Bioinformatics, 32, 2016, i629–i638
doi: 10.1093/bioinformatics/btw448
ECCB 2016

OXFORD

PeakXus: comprehensive transcription factor binding site discovery from ChIP-Nexus and ChIP-Exo experiments

Tuomo Hartonen[1,*], Biswajyoti Sahu[1], Kashyap Dave[2], Teemu Kivioja[1] and Jussi Taipale[1,2*]

[1]Genome-Scale Biology Research Program, Research Programs Unit, University of Helsinki, Helsinki, Finland and [2]Department of Biosciences and Nutrition, Karolinska Institutet, Stockholm, Sweden

*To whom correspondence should be addressed.

Abstract

Motivation: Transcription factor (TF) binding can be studied accurately *in vivo* with ChIP-exo and ChIP-Nexus experiments. Only fraction of TF binding mechanisms are yet fully understood and accurate knowledge of binding locations and patterns of TFs is key to understanding binding that is not explained by simple positional weight matrix models. ChIP-exo/Nexus experiments can also offer insight on the effect of single nucleotide polymorphism (SNP) at TF binding sites on expression of the target genes. This is an important mechanism of action for disease-causing SNPs at non-coding genomic regions.

Results: We describe a peak caller PeakXus that is specifically designed to leverage the increased resolution of ChIP-exo/Nexus and developed with the aim of making as few assumptions of the data as possible to allow discoveries of novel binding patterns. We apply PeakXus to ChIP-Nexus and ChIP-exo experiments performed both in *Homo sapiens* and in *Drosophila melanogaster* cell lines. We show that PeakXus consistently finds more peaks overlapping with a TF-specific recognition sequence than published methods. As an application example we demonstrate how PeakXus can be coupled with unique molecular identifiers (UMIs) to measure the effect of a SNP overlapping with a TF binding site on the *in vivo* binding of the TF.

Availability and Implementation: Source code of PeakXus is available at https://github.com/hartonen/PeakXus

Contact: tuomo.hartonen@helsinki.fi or jussi.taipale@ki.se

1 Introduction

Transcription factors (TFs) bind to TF-specific recognition sequences. These binding sequences can be experimentally determined *in vitro*, for example by SELEX (Jolma *et al.*, 2013) or protein-binding microarrays (PBMs) (Berger *et al.*, 2006; Mukherjee *et al.*, 2004). It is, however, also known that TFs bind to sites in the genome that do not feature a specific binding sequence and the reason for this remains largely unknown. To understand the mechanisms behind this phenomenon, TF binding positions need to be measured accurately and reliably in living cells.

Methods for studying TF binding *in vivo* have been available already over three decades (Gilmour and Lis, 1984), but genome-wide high-throughput studies have been possible only after the invention of ChIP sequencing (ChIP-seq) (Barski *et al.*, 2007). ChIP-seq reports regions bound by a specific TF but the regions can be hundreds of bps wide and arise due to different binding mechanisms. Recent upgrades to ChIP-seq, ChIP-exo (Rhee and Pugh, 2011) and ChIP-Nexus (He *et al.*, 2015), have brought the resolution of genome-wide TF binding assays to one bp regime.

The modification in ChIP-exo is the use of λ-exonuclease to digest double-stranded DNA not bound by proteins in 5′–3′ direction after enriching protein-bound DNA-regions of interest with a specific antibody. Otherwise the experiment is largely similar to ChIP-seq. The difference the λ-exonuclease makes is, however, fundamental. In ChIP-seq, the random shearing of protein-bound DNA leads to fragments where the exact location of the binding site within the fragment is unknown, whereas in ChIP-exo, the λ-exonuclease moves the 5′-end of each DNA-strand close to where the protein was bound. A schematic view of reads around a ChIP-exo/Nexus binding site is shown in Figure 1a. Digesting the 5′-end of a fragment causes loss of the adapter sequence from that end, which has to

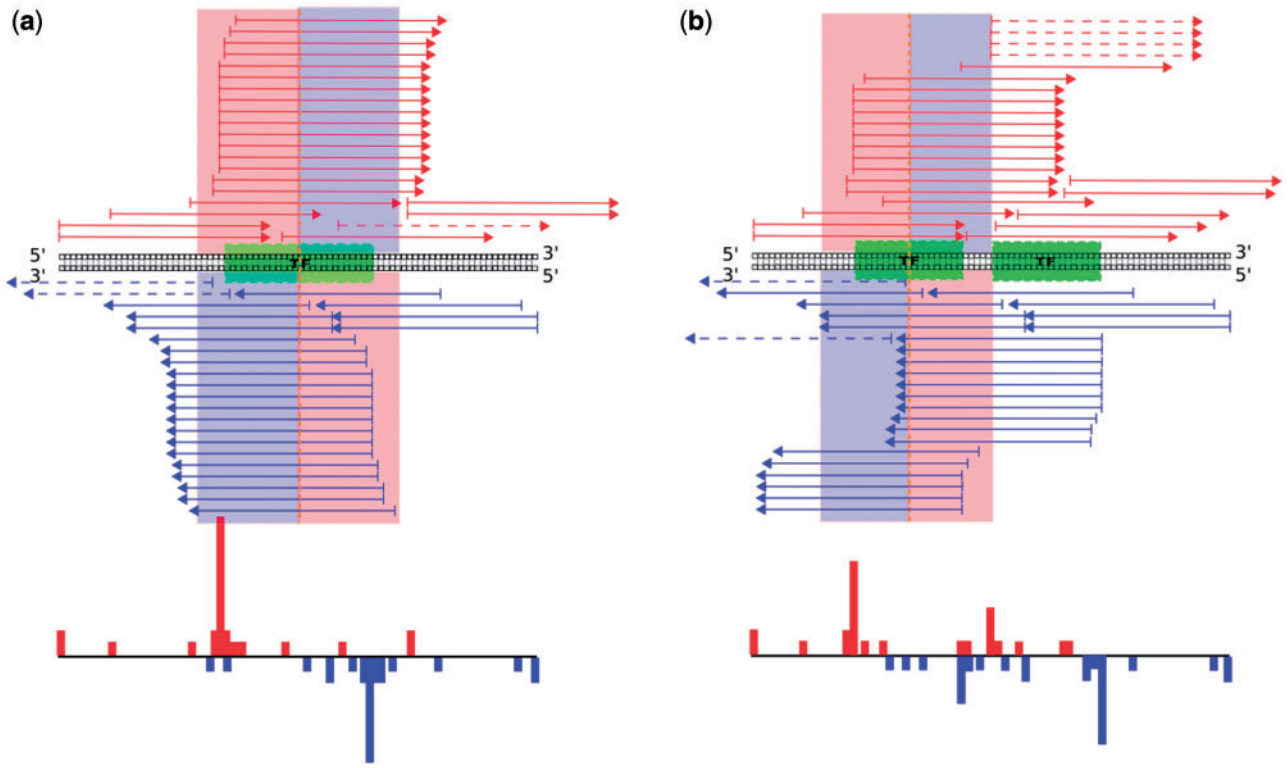

Fig. 1. Peak calling criteria. (a) Determination of a candidate peak in the presence of one true TF-DNA binding event. (b) Determination of a candidate peak in the presence of two binding events. Red arrows are reads mapped to the sense strand and blue to the antisense strand. Reads point from 5′ to 3′ direction. The red and blue bar charts below the reads correspond to counts of 5′-ends of reads (or UMIs) on the sense and antisense strands, respectively. Reads pointing toward the candidate peak center (the middle position between the borders on the sense- and antisense strands) are assumed to be true signal, while reads pointing away from the candidate peak center are assumed to be noise, as the λ-exonuclease stops at the 5′-side of a bound protein. Read 5′-end counts between the regions on red (signal) and blue (noise) background are compared against each other to separate true binding events from background noise

subsequently be re-ligated. In ChIP-exo, this is done via intermolecular ligation, whereas in ChIP-Nexus, this step is replaced with circular ligation that enhances the efficiency of the protocol (He *et al.*, 2015).

As is the case with all next-generation sequencing applications, the vast amount of data from high-throughput ChIP experiments requires computational and statistical tools for interpretation. In ChIP-seq, the computational analysis aiming at separating the true binding events from noise is known as peak calling. There exists a plethora of peak calling software for analyzing ChIP-seq data. Maybe the most widely used being MACS (Zhang *et al.*, 2008), and other popular examples being SiSSRs (Jothi *et al.*, 2008) and QuEST (Valouev *et al.*, 2008). Even the best ChIP-seq peak callers are limited by the ramifications of the experiment. They are not optimal for accurate localization of binding sites from ChIP-exo/Nexus experiments, as the signal generated by TF binding is different from that of ChIP-seq. We test our method against two published ChIP-exo peak callers, Peakzilla (Bardet *et al.*, 2013) and MACE (Wang *et al.*, 2014). Peakzilla has already been shown to outperform MACS, SiSSRs, QuEST and several other methods which is why Peakzilla is assumed to represent the state-of-the-art ChIP-seq peak caller (Bardet *et al.*, 2013).

ChIP-seq can be used to study the effect of single nucleotide polymorphism (SNP) at TF binding sites (Bailey *et al.*, 2015; Waszak *et al.*, 2014). Many diseases are associated with SNPs at non-coding genome regions (e.g. Butter *et al.*, 2012). Often the mechanism how such a SNP causes a disease is not known. One suggested explanation is that the disease-associated SNPs may overlap important TF binding sites and a disease-associated risk allele could have a great effect on the binding of the TF and thus alter the expression of the gene controlled by the affected binding site. ChIP-exo and ChIP-Nexus are better suited for allele-specific binding studies than ChIP-seq because ChIP-exo/Nexus reads aggregate closer to the TF binding site leading to higher count of unique reads overlapping with a SNP. This is important as SNPs that directly overlap binding sites are likely to have a larger effect on TF binding than ones that are flanking binding sites.

Several recent studies have shown the importance of small mutations at non-coding regions of many cancer genomes. Katainen *et al.* (2015) show that CTCF/cohesin binding sites are often mutated in multiple cancer types and Sabarinathan *et al.* (2015) report that TF binding interferes with the nucleotide excision repair machinery resulting to higher mutation rate at TF binding sites compared with their flanks in melanoma tumors. Moreover, two recent studies have confirmed mutations at TF binding sites causing cancer. In Horn *et al.* (2013), it is shown that a mutation creating only one additional binding site for Ets-family TFs and ternary complex factors at the promoter region of TERT-gene causes up to two-fold increase in transcription. The mutation was observed in a high fraction of cell line and tumor samples from metastatic melanoma. In addition, Mansour *et al.* (2014) observed that series of small insertions create binding sites for MYB-family TFs in the promoter region of TAL1-oncogene leading to allele-specific expression of TAL1 in a subset of T-cell acute lymphoblastic leukemias.

We present here a peak caller specifically designed to leverage the increased accuracy of novel experimental methods for studying

TF binding *in vivo* in an unbiased manner. Our emphasis is on filtering out false binding events by criteria well motivated by the experimental design while avoiding any unnecessary assumptions about the outcome of the experiment. We show that using unique molecular identifiers (UMIs) (Kivioja *et al.*, 2012) to count the original number of observed molecules bypasses the need to use sophisticated modeling of duplicate reads, especially when analyzing allele-specific TF binding. We have also compared our peak calling algorithm with existing ones in the literature both in ChIP-exo and in ChIP-Nexus settings. Importantly, we demonstrate the ability of our algorithm, PeakXus in allele-specific analysis of TF binding with ChIP-Nexus data.

2 Related work

Peakzilla (Bardet *et al.*, 2013) not only is a peak caller primarily designed for ChIP-seq but is also capable of analyzing ChIP-exo experiments. The design principle of Peakzilla is to have a parameter-free peak caller that learns the shape and size of binding sites from the data. Peakzilla first estimates the peak width by calculating the average distance between sense and antisense reads on the top 200 most enriched regions and multiplying it by two. Next, expected distributions of sense and antisense strand peaks are modeled as normal distributions with a standard deviation equaling to the peak width divided by five, and locations at one-fourth and three-fourth of the peak width. This somewhat arbitrary model is later used to assess if peaks fit the 'expected' shape of the signal of a binding event.

Peakzilla identifies binding sites by scanning the genome with a half-peak-sized sliding window that computes the total read count on the sense strand downstream from the current index, and within a same sized window upstream on the antisense strand. The final set of peaks consists of such peak pairs that are local maxima at least half of a peak size apart from each other. To obtain the final peak score, the total read count is multiplied with a distribution score. Distribution score is calculated by fitting the shape of the sense-, and antisense strand peaks to the expected normal distributions.

The other tested ChIP-exo peak caller is MACE (Wang *et al.*, 2014). Its core principle is to detect so called 'borders', which are highly enriched positions along the genome and are thus expected to represent the borders of a bound TF at the positions where the λ-exonuclease stopped. Borders from opposite strands are paired using the stable matching algorithm. Scoring of the border pairs is based on the total read count, as well as the distance between the borders. Optimal border pair distance is obtained by estimating it empirically from a subset of high-confidence border pairs.

There are few key reasons for developing a new ChIP-exo peak caller. First, both Peakzilla and MACE mainly find peaks of a fixed size because they are designed to find peaks that are as wide as the average best scoring ones. This can, however, lead to missing some real events, for example in situations where two TFs bind very close to each other. It is also known that some TFs recognize several binding motifs that can be of different widths (e.g. Jolma *et al.*, 2013), which might confuse a method that is designed to find peaks of essentially fixed width. On top of that, we think it is important to devise a method that makes as few assumptions of the distribution of reads and width of the binding site as possible to allow unbiased discovery. Our only assumption is that the λ-exonuclease aggregates read 5′-ends at the borders of bound proteins, as illustrated in Figure 1. We also believe that using UMIs to remove duplicates and to conserve the information about the count of initial molecules is important for critical applications like allele-specific binding analysis, where it is essential to know accurately the number of fragments originating from each allele.

3 Algorithm for TF binding site detection

3.1 Duplicate filtering using UMIs

UMIs (Kivioja *et al.*, 2012) are a means to retain information about the original number of molecules in a sequencing library. PCR-amplification is known to suffer from amplification bias resulting to some of the initial molecules being amplified more than others (e.g. Aird *et al.*, 2011). UMIs can be used to circumvent this problem by attaching a random DNA-sequence label to each of the molecules in the initial library. If the number of available labels is sufficient, it is extremely unlikely for two different molecules with the same UMI-label to map exactly to the same genomic position. This means that as long as the complexity of the library is preserved it can be amplified and normalized freely without losing information about the original number of molecules. After sequencing at sufficient depth, all identical UMIs mapping to same positions can be treated as duplicates, and the original number of molecules can be estimated by counting each UMI-label once per position.

In ChIP-seq, peaks caused by TF binding are wide and broad, and the sensitivity of peak calling does usually not suffer from removing duplicated reads by simply deleting all reads that map to identical location and strand. This strategy is no longer viable in ChIP-exo/Nexus, because theoretically true ChIP-exo/Nexus binding events look very much like artifacts caused by PCR-duplicates, as seen from Figure 1! UMIs allow discarding duplicated reads without removing borders caused by true binding events.

3.2 Discovering candidate peaks

Read count on sense strand is marked with $c_+(i)$ and on antisense strand $c_-(i)$ at genomic coordinate (base pair) i. Each read is saved corresponding to the location of the 5′-end of the read as the λ-exonuclease stops at the 5′-end of a bound protein. Two schematic examples of how the read count profile is built can be seen in Figure 1. When UMIs are used, $c_+(i)$ and $c_-(i)$ represent counts of *unique* molecules.

A TF-DNA binding event in a ChIP-exo/Nexus experiment should be located between 'borders' on the sense and antisense strands because the λ-exonuclease should always stop at the edge of a bound protein. Theoretically, we expect to see signatures as wide as the bound protein or protein complex completely devoid of reads in the middle, with a large amount of reads mapping to the sense strand on the left side, and to the antisense strand, on the right side of the binding site, as illustrated in Figure 1a.

Thus the first phase of binding site recognition is to search for these transition points. The algorithm goes through the genome looking for positions where the total read 5′-end count $c_+(i) - c_-(i) < 0$ and $c_+(i-1) - c_-(i-1) \geq 0$. For each such position, index of the left border of the candidate peak is $k = arg\,max\{c_+(k) - c_-(k)\}$ such that $i - w < k < i$ and $c_+(k) - c_-(k) > 0$. Similarly for each such position, index of the right border of the candidate peak is $j = arg\,min\{c_+(j) - c_-(j)\}$ such that $i \leq j < i + w$ and $c_+(j) - c_-(j) < 0$. A candidate peak is accepted only if both c_+(k)−c_−(k) > 0 and c_+(j)−c_−(j) < 0. At this point, we store all the candidate peaks even if they overlap. The true peak among overlapping ones is decided after a peak score has been calculated for all candidate peaks.

3.3 Significance testing and removing overlapping peaks

The read 5′-end counts from ChIP-exo/Nexus binding events exhibit two distinct features. First, true events should have a high total read 5′-end count around the binding site, and as earlier discussed, there should be the signature borders on opposite strands flanking the

site. To account for both these features, we propose a scoring scheme illustrated in Figure 1.

The underlying idea of this scheme is that due to the λ-exonuclease digesting DNA from 5′ to 3′ direction, only reads pointing towards the center of a candidate peak can be regarded as resulting from true binding events. This means that ideally there should not be any 5′-ends of reads mapping to areas with blue background, but instead the areas with red background should have a plenty of reads (Fig. 1a). Moreover, if distance from a candidate peak summit is calculated for each 5′-end of a read, the red areas should give rise to a high number of distances approximately equal to half of the length of the candidate peak, while the blue areas should lack any pattern since the few reads observed there should be randomly positioned. To clarify the division into signal and background, background reads are drawn with a dashed line. To separate signal from background, we thus create histograms of distances between each read 5′-end and the candidate peak summit separately for both the red and blue regions, and compare these distributions against each other to determine if the signal-region (red) is different from the background-region (blue).

In reality, the binding signatures can be more complex than in the simple example in Figure 1a. In many cases, two (or sometimes more) proteins bind close to each other, producing two sets of borders in ChIP-exo/Nexus experiments, as the random shearing of DNA creates both fragments that contain both, and fragments that contain only one of the bound proteins. Even though being geared towards finding single TF binding events to avoid making assumptions of the shape or size of the binding signature, PeakXus will not overlook more complex events, as illustrated in Figure 1b. The read density distributions on the red and the blue background are still significantly different from each other, since only other half of the signal region gets convoluted with the reads that belong to another binding event.

3.3.1 Distinguishing true events from background

Let us denote the non-normalized frequency distribution of distances between background read 5′-end and candidate peak summit for a given candidate kj (k marking the left edge coordinate and j the right, respectively) with $\mathbf{B^{kj}}$, and with $\mathbf{S^{kj}}$ for distances between signal read 5′-ends and candidate peak summit. Then, the G-test statistic is

$$G^{kj} = 2\sum_{i=0}^{d_{max}} S_i^{kj} \ln\left(\frac{S_i^{kj}}{B_i^{kj}}\right), \quad (1)$$

where index i enumerates all possible distances excluding the ones where $B_i^{kj} = 0$, i.e. the positions where the read 5′-end count at position i is zero. The maximum allowed distance is determined by the edges of the candidate peak kj, denoted with k_{kj} being the left edge and j_{kj} being the right edge, considering also the d immediate flanking positions. In other words, the maximum distance between a 5′-end of a read and the peak candidate summit for candidate kj is $d_{max} = \frac{j_{kj}-k_{kj}}{2} + d$. We used $d = 5$ as a default to allow some uncertainty for the stop base of the λ-exonuclease.

The G-test is our method of choice because its test statistic approximates the χ^2-distribution better than Pearson's χ^2-test (Harremoës and Tusnády, 2012). This test addresses both the distinctive features described above as it gives a larger test statistic value (smaller P value) both when S_i^{kj}'s and S_i^{kj}/B_i^{kj}'s are large.

3.3.2 Pseudocount

By definition, G-test does not account for distances that are not found from the distribution of background distances. This can be problematic, since the assumption is that especially at the immediate flanks of a bound TF, there is a lot of true signal and minimally background noise. G-test overlooks these differences if $B_i^{kj} = 0$ even though these are especially the kind of differences we want to capture. To solve this problem, we applied the pseudocount-approach (Durbin *et al.*, 1998). After measuring the distances between all reads and the summit of the candidate peak, a constant p is added to each B_i^{kj} and S_i^{kj}. This leads to the final test statistic

$$G_{pseudo}^{kj} = 2\sum_{i=0}^{d_{max}} (S_i^{kj} + p) \ln\left(\frac{S_i^{kj} + p}{B_i^{kj} + p}\right). \quad (2)$$

Given G_{pseudo}^{kj}, a P value is then calculated from the χ^2 cumulative distribution function. The null hypothesis that the background and signal distributions are identical is rejected if P value < 0.05, the same significance threshold is used also in Peakzilla and MACE. A pseudocount accounts for the fact that there exists a non-zero probability of observing reads at any position, but due to the large size of the genome and a limited sample size, only a fraction of the positions are covered by reads.

3.3.3 Peak score

The final ranking of peaks is based on a peak score. With the peak score, we want to emphasize that given two peaks where the difference between the background and signal regions is small, the peak with a higher total read count is likely more important. Thus each peak is assigned a score

$$SC^{kj} = G^{kj}\left(\sum_{i=k_{kj}-d}^{\lfloor m \rfloor} c_+(i) - c_-(i) + \sum_{i=\lceil m \rceil+1}^{j_{jk}+d} c_-(i) - c_+(i)\right), \quad (3)$$

where the middle position of the candidate peak is $m = (k_{kj} + j_{kj})/2$. Coming back to the issue of two proteins binding close to each other, using the testing and scoring scheme presented above can lead to a somewhat smaller score for a peak that is flanked by another peak if the signal reads of the flanking peak overlap with the background-region of the first peak, as is illustrated in Figure 1b. However, strong binding events will not be missed even if flanked by another binding protein.

In a nutshell, the following steps are performed for all candidate peaks: (1) calculate the P value using G-test. If P value is larger than 0.05, then discard the candidate. (2) Calculate the peak score for all remaining peaks. (3) Find all sets of overlapping peaks and discard all others but the peak with the highest peak score from each set. (4) Calculate false discovery rates using the Benjamini–Hochberg procedure for dependent test statistics (Benjamini and Yekutieli, 2001) using the initial number of candidate peaks as the total number of tested null hypotheses.

4 Data and resources

The in-house CTCF ChIP-Nexus experiment was conducted on human LoVo-cell line (adenocarcinoma of the colon). DNA-fragments were sequenced with Illumina HiSeq2000-sequencer. The experiment uses UMI-labels of length 5 that include all possible combinations of A, C, G and T. Reads were aligned against the hg19-reference genome using the Burrows–Wheeler alignment tool (Li and Durbin, 2009) and specifically the bwa aln-algorithm, with default parameters. The SAMtools and BEDtools toolkits (Li *et al.*, 2009; Quinlan and Hall, 2010) were used to perform e.g. filtering (MAPQ < 20 used for all experiments) and various genome arithmetics tasks.

The MAX and TWIST ChIP-Nexus experiments used in this work are published in He *et al.* (2015). These experiments were performed in *Drosophila melanogaster* cells and were aligned to dm3 reference genome with bwa aln-algorithm with default parameters. These experiments utilize similar UMI-design as described above for CTCF ChIP-Nexus.

The CTCF ChIP-exo experiment by Katainen *et al.* (2015) was aligned to hg19 similarly as described above. The CTCF ChIP-exo experiment by Rhee and Pugh (2011) was downloaded readily aligned. These experiments do not utilize UMIs.

The whole genome sequencing (WGS) experiment was conducted on the same LoVo-cell line used in the in-house ChIP-exo/Nexus experiments. Sequencing was performed as paired-end sequencing with Illumina HiSeq2000. Reads with identical start and end positions for the both paired fragments were discarded as duplicates.

The *in vitro* binding specificities of TFs can be presented as positional weight matrices (PWMs) that give the affinity of a given TF towards any DNA-sequence of given length. Using PWMs, it is possible to scan the genome for hits of the binding motifs and rank the resulting regions according to their affinity towards the given TF. This requires both the PWMs and a software for scanning the genome. These PWM hits to the reference genome are called high-affinity recognition sequences (HARSs) of the TF. It is expected that the true occupied binding sites are enriched at HARS sites relative to other positions, but not all the HARS sites bind the corresponding TF *in vivo* due to for example the local chromatin context. The CTCF-PWM used in this work is from Jolma *et al.* (2013) and the *Drosophila melanogaster* PWMs are from the JASPAR database (Mathelier *et al.*, 2013) (MAX: MA0058.3 and TWIST: MA0249.1). The lists of locations of binding motifs on the reference genomes were produced with the MOODS software (Korhonen *et al.*, 2009).

Variant calls used in this work were obtained from the 1000 Genomes database (1000 Genomes Project Consortium, 2010). The initial phase 3 release of the database includes more than 79 million variant sites. We used the initial phase 3 SNPs to assess the allele specificity of CTCF binding.

5 Results

To make the comparison among PeakXus, MACE and Peakzilla as neutral as possible, all methods were run with their default parameters with one exception: Peakzilla reported considerably fewer peaks than the other methods when using the default parameter values. Therefore, we used input flags '-c 0 -s 0', where c is the fold enrichment cut-off and s the peak score cut-off. The default parameters for PeakXus are the following: maximum binding site width $w = 60$ bps, number of flanking bases considered when calculating peak score $d = 5$, pseudocount $p = 1$. UMIs were used by PeakXus when available. As neither MACE nor Peakzilla offers a method for controlling the false discovery rate (FDR) for ChIP-exo/Nexus experiments, PeakXus was also run without controlling for FDR. GeneTrack (Albert *et al.*, 2008) is a method essentially listing the most enriched ChIP-exo/Nexus genomic sites without making any more sophisticated analysis. Therefore, the results obtained with GeneTrack are shown as a baseline in Figures 2 and 4.

TF binding occurs both at sites with a TF-specific high-affinity recognition sequence (HARS) and at sites devoid of such sequences. Thus, there is no absolute way to assess if a peak reported by a peak caller corresponds to a true binding event. It is nevertheless reasonable to assume that a large fraction of the strongest peaks overlap with an HARS, as these are the sites that are known to bind the corresponding TF strongly. Not all the binding sites are accessible to binding. It is known that local chromatin context is a major determinant of where TFs bind (e.g. Li *et al.*, 2011). DNA methylation has also been shown to reduce TF binding (Wang *et al.*, 2012). Thus, the expectation is to observe both peaks that lack the underlying HARS as well as locations with an HARS that do not bind the associated TF in the ChIP-exo/Nexus experiments.

Figure 2 shows how many of the peaks found by a given peak caller overlap with HARS sites for three different CTCF experiments (one ChIP-Nexus and two ChIP-exo) on human and two ChIP-Nexus experiments (MAX and TWIST) on *Drosophila melanogaster* cells. Slopes of the curves indicate the efficiency of finding peaks overlapping with a TF-specific HARS. It is seen that for the very top

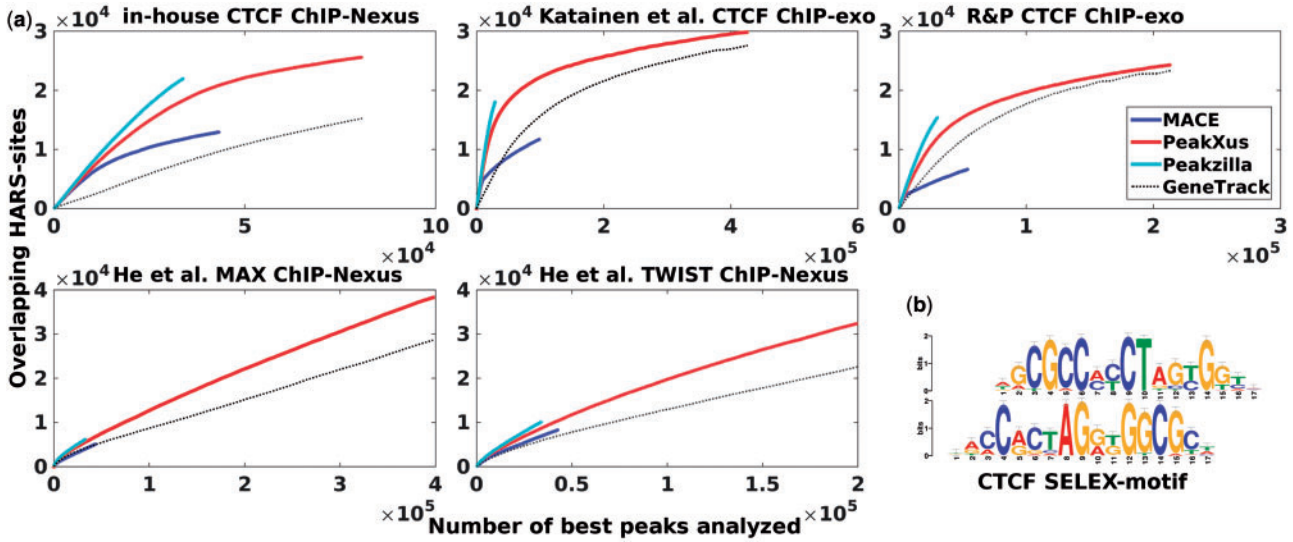

Fig. 2. (a) Total number of high affinity recognition sequences (HARS) found from ChIP-exo/Nexus experiments as a function of top scoring peaks analyzed. The *y*-axis corresponds to the number of HARSs that were found to overlap with *x* (value shown on the *x*-axis) top peaks. Peaks were sorted by the score given by each peak caller. A peak was considered to match with an HARS if distance between the center of the HARS and the center of the peak was less than or equal to 20 bps. A total of 300 000 highest affinity HARS sites were considered both for human CTCF experiments and for fly MAX and TWIST experiments. For GeneTrack, the same amount of top peaks is shown as is reported by PeakXus. (b) Canonical CTCF binding sequence as measured in Jolma *et al.* (2013)

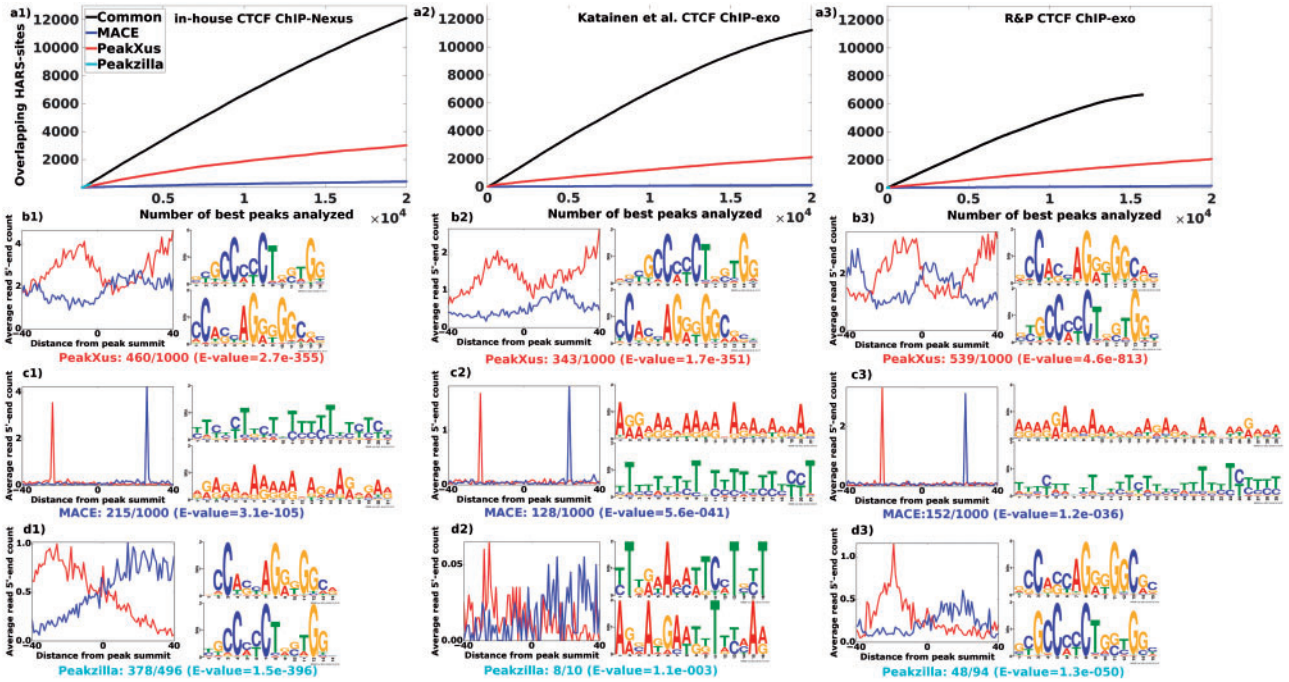

Fig. 3. Analysis of peaks specific to each of the peak callers in CTCF-experiments. Methods are color coded as the following: Peakzilla = cyan, MACE = blue, PeakXus = red and the peaks reported by all methods are plotted in black. Leftmost column (from a1 to d1) corresponds to results from our in-house ChIP-Nexus experiment, middle column (from a2 to d2) to ChIP-exo experiment from Katainen *et al.* (2015) and rightmost column (from a3 to d3) to ChIP-exo experiment from Rhee and Pugh (2011). (a1–a3) The *y*-axis corresponds to the number of high-affinity binding sites (HARSs) that were found to overlap with *x* (value shown on the *x*-axis) top peaks. Only 20 000 highest scoring peaks are shown for clarity. A peak was considered to match with an HARS if distance between the HARS center and the peak center was less than or equal to 20 bps. For each method, only the peaks that did not overlap with any of the peaks found by the other two methods were included. Peakzilla-specific peaks are so few (in-house ChIP-Nexus: 496, Katainen *et al.* ChIP-exo: 11, R&P ChIP-exo: 94) that the curves are hardly visible. (b1–b3) More detailed analysis of the top 1000 peaks specific to PeakXus. Each column contains on left the average read 5′-end count around the peak center for sense- (red) and antisense (blue) strands. On the right are shown the both orientations of the best motifs reported by MEME. The number of hits to the best scoring MEME-motif among the top peaks is shown below each of the motif pairs along with the corresponding *E*-value reported by MEME. (c1–c3) More detailed analysis of the top 1000 (or less, if 1000 were not found) peaks specific to Peakzilla and d1–d3) MACE

peaks, there is negligible difference in the efficiency of finding peaks with an HARS. Efficiency drops significantly faster for MACE than for the other methods. Highest points of the curves reveal that PeakXus consistently finds more peaks with an underlying HARS than the other tested methods.

One way to compare peak callers is to look at how different are the peaks reported by them that the other methods do not find. To investigate this, we selected for each method only the peaks that the other two methods did not report and repeated the analysis shown in Figure 2. Comparison for the three CTCF experiments is shown in Figure 3(a1–a3), now cutting the *x*-axis after 20 000 peaks to focus on the highest scoring peaks. As expected, the peaks reported by all three methods contain a notably higher fraction of peaks with a match to an HARS than the peaks specific to any given method. Peakzilla finds less than 100 peaks the other methods do not from two of the three experiments, and thus these curves are not clearly visible. Importantly, PeakXus reports in total considerably more peaks that are not found by the other two methods but still co-occur with a CTCF HARS.

Surprisingly, we observe that only a few percent of the peaks found by MACE alone overlap with a CTCF HARS. A possible explanation for this could be that there is another sequence CTCF binds to that yields such a different signal that PeakXus or Peakzilla miss these peaks. To further investigate this, we analyzed the top 1000 method-specific peaks with MEME-software (Bailey *et al.*, 2006). MEME is a tool that finds significantly enriched subsequences occurring in a set of input sequences. As an input, we gave the 100 bp-wide regions centered around the top 1000 peaks. MEME was evoked with parameters '-dna -mod anr -nmotifs 5 -maxsites 1000 -minw 4 -maxw 50 -revcomp'. The top motifs found by MEME are shown in Figure 3(b–d) at the corresponding columns for each experiment and at the corresponding rows for each peak caller. On the left hand side of each motif is shown the corresponding average read 5′-end counts around the same top 1000 method-specific peaks. Comparing the found sequence motifs to the *in vitro* motif in Figure 2b, it is seen that the expected CTCF motif is found by PeakXus from all the experiments as the best-scoring motif. MEME finds the CTCF *in vitro* motif from 2/3 sets of Peakzilla-specific peaks and from 0/3 sets of MACE-specific peaks. Furthermore, the read densities of the top method-specific peaks show that the peaks reported by PeakXus have a signature where most of the sense strand reads are on the left side of the peak summit and antisense strand reads on the right side of the peak summit, relatively close to the peak center. Read densities specific to Peakzilla and MACE peaks are radically different. The read counts around Peakzilla-specific peaks are low, meaning the corresponding binding events are weaker. MACE-specific peaks, in turn, do have the borders on opposite strands around the summit, but they completely lack the CTCF-motif. It is also interesting to note that the read density profiles of peaks specific to PeakXus exhibit multiple maximas around the peak middle position. This suggests that these signatures correspond to binding events where multiple TFs are bound close to each other producing multiple sets of borders, as described in the Introduction. All in all, PeakXus reports significantly more peaks coinciding with a CTCF HARS than MACE or Peakzilla.

Figure 4 shows the localization accuracy of the top 1000 peaks reported by each method with respect to HARS sites in CTCF ChIP-exo/Nexus experiments. Around 20% of the top 1000 peaks localize within 5 bps of the HARS center regardless of the peak caller. Respective superiority of the peak callers in precisely localizing the summits of the top peaks seems to strongly depend on the experiment, as none of the methods consistently outperforms the others.

6 Allele-specific binding analysis

Determining allele specificity of TF binding in ChIP experiments is in principle simple. For each polymorphism, one needs to retrieve all reads mapping to the location and just count the fraction of reads mapping to each allele. Straightforward comparison of read counts mapping to different alleles has, however, serious caveats. PCR-amplification can introduce library amplification bias which can result into false-positive allele-specific binding (ASB) calls. There have been attempts to control this in ChIP-seq ASB analysis in the past (Waszak *et al.*, 2014). However, the problem can be made trivial by using UMIs in the experiment, as each UMI represents a distinct molecule in the initial library and if there are enough UMI-labels in use, the number of initial molecules in the library can always be accurately recovered. Even if PCR amplifies all reads evenly, measuring ASB from difference between read counts mapping to different alleles is problematic when sequencing deeply. With high enough read counts any small difference stemming from any artifact will eventually result into a statistically significant result. Using UMIs, this does not happen as there will never be more UMIs than initial molecules in the library.

Second, a naïve expectation is to observe 50% of the reads to map to each of the alleles. However, biases caused by copy number variation, clonal heterogeneity or the tendency of reads to map better to the reference than the alternate allele can render this null hypothesis unrealistic. These biases can be controlled by calculating the genomic allelic ratio (gAR) (Bailey *et al.*, 2015; Degner *et al.*, 2009) from the WGS reads of the genome used in the ChIP experiment and using this ratio as the null expectation. The gAR is simply fraction of reads that overlap with a given SNP mapping to the reference allele.

Outline of the ASB analysis is simple: (1) select the SNPs that overlap with a peak. (2) Filter out duplicated reads by counting each UMI-label once per position. (3) Obtain a list of nucleotides overlapping with each SNP from the filtered reads. (4) Test each SNP-position for significant ASB. Most of the steps are straightforward except the significance testing, which is discussed in detail below. To avoid including homozygous SNPs to our analysis, all SNP-locations with less than three WGS reads with unique positions for both ends of paired-end reads per allele were discarded.

After step three, our goal is to compute if the fraction of reads mapping to reference allele in ChIP experiment significantly differs from the fraction of reads mapping to reference allele in the WGS experiment for each SNP. The null hypothesis is that there is no difference. To test if the null hypothesis holds, we assume that we draw reads (in the case of WGS) or UMIs (in the case of ChIP) overlapping a SNP from a large pool of reads with replacement. This means that observing one UMI does not alter the probabilities of observing any of the UMIs. This is a reasonable assumption as we are actually sampling from the pool of PCR-duplicates of the original molecules, which is for practical purposes infinite even though the number of *observed* binding events might be small for one SNP. Drawing a read/UMI that overlaps with the SNP of interest is a rare event. With these assumptions, one could calculate the local gARs from the WGS experiment and use these as the parameters of binomial test to determine if the allelic ratios differ between the experiments (Bailey *et al.*, 2015; Degner *et al.*, 2009; McDaniell *et al.*, 2010; Rozowsky *et al.*, 2011). This, however, assumes that there is no error in the gARs. To account for the uncertainty of the gARs caused by the varying coverage of the WGS reads, we use the general probability distribution governing occurrence of the same rare event in duplicate experiments (Audic and Claverie, 1997). The two experiments are viewed as replicates with k reads that overlap with the SNP i mapping to reference allele in the ChIP experiment while n reads are observed to map to reference allele in the WGS experiment. Thus the probability of observing k given n by chance is

$$P_i(k|n) = \binom{N_2}{N_1}^n \frac{(k+n)!}{k!n!(1+N_2/N_1)^{k+n+1}}, \quad (4)$$

where N_1 is the total number of reads overlapping with SNP i in ChIP and N_2 in WGS experiment. $P_i(k|n)$ is calculated for each SNP and treated as the test statistic.

6.1 Allele-specific binding analysis in CTCF ChIP-Nexus

We performed whole genome sequencing for the LoVo-cell line and calculated the gARs from the WGS reads. These gARs were used as the control for allele-specific binding (ASB) of CTCF in a ChIP-Nexus experiment. Table 1 shows statistics of the ASB analysis. In the following, we study how SNPs that change the binding sequence of CTCF affect binding and thus focus on SNPs that overlap with a high-affinity recognition sequence (HARS) and a peak.

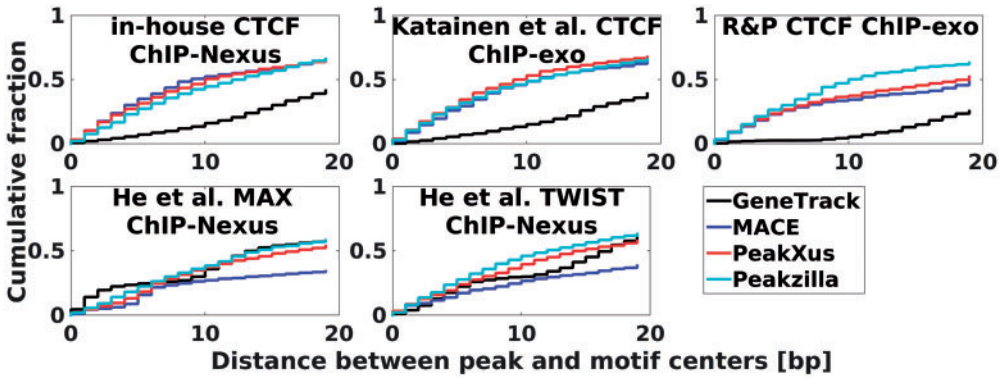

Fig. 4. Localization accuracy of peaks in three ChIP-Nexus and two ChIP-exo experiments. The *y*-axes show the fraction of top 1000 peaks that are within *x* bps (shown on the *x*-axis) of a high-affinity recognition sequence (HARS). Distance is measured between the center of an HARS and center of a peak

Table 1. Statistics of allele-specific binding analysis on the in-house CTCF ChIP-Nexus experiment

Total number of peaks	80 806
Peaks overlapping with a SNP	4226
SNPs with a peak and an HARS	314
– Significant with UMIs	$73 \rightarrow F^{peak}_{UMI} = 73/314 = 0.232$
– Significant without UMIs	$142 \rightarrow F^{peak}_{read} = 142/314 = 0.452$
SNPs flanking HARS sites	1494
– Significant with UMIs	$170 \rightarrow F^{flank}_{UMI} = 170/1494 = 0.114$
– Significant without UMIs	$407 \rightarrow F^{flank}_{read} = 407/1494 = 0.272$
SNPs 50 bps from an HARS	1278
– Significant with UMIs	$84 \rightarrow F^{50bp}_{UMI} = 84/1278 = 0.066$
– Significant without UMIs	$205 \rightarrow F^{50bp}_{read} = 205/1278 = 0.160$

It is reasonable to assume that the SNPs directly overlapping with a CTCF HARS have a larger effect on CTCF binding than the SNPs that lay further away. If using UMIs to discard duplicated reads helps in detecting ASB events, we would expect UMIs to separate the SNPs overlapping with an HARS from other SNPs better than if the analysis is conducted with raw reads. This means we expect the UMI-analysis to report higher fraction of SNPs to be significant under the HARS sites relative to other positions. Using UMIs, we observe 2.04 times higher fraction of SNPs to be significant under an HARS relative to the 16 motif-flanking bps ($F^{peak}_{UMI}/F^{flank}_{UMI}$). Without UMIs, this fraction is 1.66 ($F^{peak}_{read}/F^{flank}_{read}$). When comparing the SNPs under an HARS to region 50 bps–58 bps away from a CTCF HARS, we observe 3.52 times higher fraction of SNPs to be significant under an HARS with UMIs ($F^{peak}_{UMI}/F^{50bp}_{UMI}$), and 2.83 without UMIs ($F^{peak}_{read}/F^{50bp}_{read}$). This suggests that UMIs help to distinguish the SNPs underlying an HARS from other SNPs.

Reference allele ratio ($N_{ref}/(N_{ref}+N_{alt})$, where N_{ref} is the number of hits to reference allele and N_{alt} to alternate), should depend more or less linearly on the affinity change a SNP causes to a TF binding site. This is because the more a SNP changes affinity of the binding sequence, the larger should be the difference in binding between the alleles. Figure 5 shows scatter plots of reference allele ratio versus affinity change of the CTCF recognition sequence due to a SNP for the in-house CTCF ChIP-Nexus experiment. The correlation is stronger when using UMIs indicating that UMIs help to discard false-positive ASB calls.

Methylation of CG sites in the binding motif disrupts CTCF-binding (Wang *et al.*, 2012), which could explain some outliers in Figure 5. Let us assume that we have a reference CTCF-binding site AGCAGACCTAGTGGTA (sequence 1), and a SNP A→G at position four, changes it to AGCGGACCTAGTGGTA (sequence 2). Comparing with the *in vitro* CTCF recognition sequence in Figure 2b, sequence 2 should have a higher affinity towards the CTCF protein (reference allele ratio < 0.5). However, the mutation introduces an additional CG to positions 3–4, which can potentially lead to methylation of the site and stronger observed binding to sequence 1 even though it has lower affinity towards the recognition sequence. However, we observe no obvious outliers that could be caused by methylation based on the ASB analysis conducted using UMIs.

7 Conclusions and discussion

We have shown that the developed ChIP-exo/Nexus peak caller is a valuable addition to the computational tools available for studying TF binding locations and patterns *in vivo*. The main advantages of PeakXus compared with the previously published methods are at least 2-fold. First, PeakXus constantly reports more peaks that overlap with known strong binding sites than MACE and Peakzilla. This is currently the most reliable way to asses if the reported peaks correspond to true binding events. Second, PeakXus does not achieve this by fitting all peaks to a profile based on a fraction of sites with high read counts, an approach taken by the other two peak callers. Fitting to high coverage sites is a safe approach as the high coverage sites correspond to true binding events with high likelihood. However, this can also be dangerous, since it can lead to missing some weaker binding patterns that result in a differently shaped signal.

Current peak callers cannot localize peak summits with the advertised one-base pair resolution of ChIP-exo/Nexus for others than the very top peaks, as shown in Figure 4. Peak localization accuracies of all the tested methods are very similar for the top 20% of the most accurately localized peaks, after which there is some deviation depending on the experiment. There are at least two possible reasons as for why only 20% of the best scoring peaks are within 5 bps of the high-affinity recognition sequence (HARS) center. The most obvious explanation is biological—there are likely several cases among the top peaks where some other protein binding next to

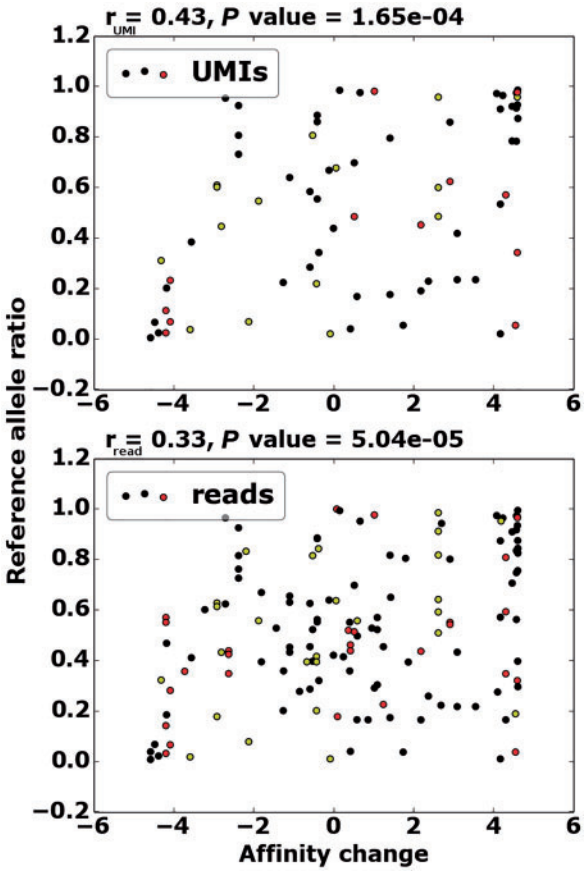

Fig. 5. Reference allele ratio as a function of binding sequence affinity change in the in-house CTCF ChIP-Nexus experiment. The *y*-axis shows reference allele ratio while the *x*-axis is the affinity change (reference minus alternate sequence affinity). SNPs with *P* value > 0.01 are filtered out. Red dots represent SNPs where alternate allele creates a CG site to the sequence but reference does not. Yellow dots represent SNPs where sequence with reference allele has an extra CG. Other SNPs are black. Values of correlation coefficients (*r*) along with the corresponding *P* values are shown above the subfigures. *TOP*: *P* values are calculated using UMIs (73 significant SNPs). *BOTTOM*: *P* values are calculated using raw read counts (142 significant SNPs)

the measured TF is present in a large fraction of the DNA-fragments possibly shifting the border created by the λ-exonuclease. PeakXus may be more sensitive to this than the other methods, as it does not fit the peak width to most enriched sites. This might be a strength of PeakXus in the future, if the experiment is developed more sensitive by, for example, shortening the fragment size. This would mean that it becomes more unlikely for two TFs to fit into a same fragment, allowing PeakXus to better separate TFs bound close to each other. Another possible explanation is that there is some noise associated with the stop base of the λ-exonuclease and it would not always stop at the same exact position. However, this effect is probably not as large as the biological one.

We compared the three peak callers by analyzing the peaks uniquely found by each of the methods. Figure 3a shows that a very high fraction of peaks commonly reported by the three methods contain the CTCF-specific HARS. These are the 'easiest' peaks to find as it is expected that the top peaks are found from sites with the highest affinity for the corresponding TF. PeakXus finds thousands of true binding events the other methods do not detect. The result that only 1–2% of the peaks found by MACE alone overlap with an HARS is surprising. We mined the sequences under the top 1000 peaks reported by PeakXus, Peakzilla and MACE for common sequence motifs to see if the existence of some other binding sequence could explain why so few of the peaks reported only by MACE overlap with a CTCF HARS. We found no strong common binding motifs from the set of peaks unique for MACE. Furthermore when reads with identical strand and 5′-end were removed and the analysis re-run, MACE reproduced only 2005 of the 99 564 peaks from ChIP-exo experiment by Katainen *et al.*, and 10 549 of the 54 510 peaks from CTCF ChIP-exo experiment by Rhee and Pugh. After removing duplicates, MACE was unable to find any peaks from in-house ChIP-Nexus data. This highlights the difficulty of separating true ChIP-exo/Nexus binding events from artifacts created by PCR-duplicates if UMIs (Kivioja *et al.*, 2012) are not used in the experiment and supported by the peak caller. Importantly, comparing the read 5′-end count distributions of the method-specific peaks, it is seen that peaks found by PeakXus but not by the other methods have the expected appearance of a true binding event whereas Peakzilla and MACE-specific peaks seem to be either very small in terms of total read count or resulting from duplicated reads (Figures 3(b–d)).

We also demonstrated how PeakXus can be used to measure the level of allele-specific binding (ASB). We applied, to our knowledge first time in ASB analysis, UMIs to remove duplicate reads to avoid library amplification bias. In previous ChIP-seq ASB analyses, accounting for read duplication bias has either required complicated computational approaches (Waszak *et al.*, 2014) or has been neglected (Bailey *et al.*, 2015). We introduced a way to account for the uncertainty of the allelic ratios calculated from the whole genome sequencing reads that are used as a control for the ChIP-Nexus ASB calls. To our knowledge, this uncertainty has previously not been considered. In Figure 5, we show a strong correlation between reference allele ratio and affinity change induced by a SNP to the binding sequence in a CTCF ChIP-Nexus experiment indicating that our approach can be used to reliably study the allele specificity of TF binding. Moreover, the correlation is stronger when using UMIs compared with analysis with read counts suggesting that utilizing UMIs in ASB analysis filters out false-positives. We also observed that conducting the ASB analysis with UMIs reports relatively more significant ASB events under HARS sites than at the immediate proximity of them compared with the analysis using read counts. This further highlights usefulness of UMIs in ASB analysis as the SNPs overlapping an HARS are expected to have the largest effect on ASB.

Our ASB analysis was conducted using SNPs from the 1000 Genomes database. Analyzing variants from the same cells used in the ChIP experiments will give more power to ASB analysis because there will be more variants available. We also discussed how DNA-methylation distorting TF binding could manifest in ChIP-exo/Nexus ASB analysis. To properly asses the effect of methylation on TF binding, the ChIP experiments should be coupled with bisulfite sequencing, which is ongoing.

Large-scale screenings of TF binding locations and patterns have thus far been conducted using ChIP-seq (e.g. ENCODE Project Consortium, 2012). To perform such high-throughput analysis with ChIP-exo or ChIP-Nexus requires analysis tools specifically designed for these experiments. Here we have presented a peak caller capable of capturing protein–DNA binding events in an unbiased manner without making *a priori* assumptions about the nature of the event. Together with a high-throughput ChIP-Nexus laboratory protocol PeakXus will allow studying the binding patterns of different TFs leaving room for the discovery of novel binding modes.

Acknowledgements

The authors would like to thank Dr. E. Kaasinen for help with the WGS data and T. Leung for technical assistance.

Funding

This work has been supported financially by the Academy of Finland (Finnish Center of Excellence Program 2012–2017, 250345, Personal Grant 274555 to B.S.) and the Integrative Life Sciences (ILS) Doctoral Program.

Conflict of Interest: none declared.

References

1000 Genomes Project Consortium. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

Aird, D. *et al.* (2011) Analyzing and minimizing pcr amplification bias in illumina sequencing libraries. *Genome Biol.*, **12**, R18.

Albert, I. *et al.* (2008) Genetrack – a genomic data processing and visualization framework. *Bioinformatics*, **24**, 1305–1306.

Audic, S. and Claverie, J.M. (1997) The significance of digital gene expression profiles. *Genome Res.*, **7**, 986–995.

Bailey, S.D. *et al.* (2015) ABC: a tool to identify SNVs causing allele-specific transcription factor binding from ChIP-seq experiments. *Bioinformatics*, **31**, 3057–9.

Bailey, T.L. *et al.* (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, **34**, W369–W373.

Bardet, A.F. *et al.* (2013) Identification of transcription factor binding sites from ChIP-seq data at high resolution. *Bioinformatics*, **29**, 2705–2713.

Barski, A. *et al.* (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.

Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, 1165–1188.

Berger, M.F. *et al.* (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, **24**, 1429–1435.

Butter, F. *et al.* (2012) Proteome-wide analysis of disease-associated SNPs that show allele-specific transcription factor binding. *PLoS Genet.*, **8**, e1002982.

Degner, J.F. *et al.* (2009) Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, **25**, 3207–3212.

Durbin, R. *et al.* (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Gilmour, D.S. and Lis, J.T. (1984) Detecting protein-DNA interactions in vivo: distribution of RNA polymerase on specific bacterial genes. *Proc. Natl. Acad. Sci.*, **81**, 4275–4279.

Harremoës, P. and Tusnády, G. (2012) Information divergence is more chi squared distributed than the Chi squared statistics. *ArXiv e-Prints*.

He, Q. *et al.* (2015) Chip-nexus enables improved detection of in vivo transcription factor binding footprints. *Nat. Biotechnol.* **33**, 395–401.

Horn, S. *et al.* (2013) TERT promoter mutations in familial and sporadic melanoma. *Science*, **339**, 959–961.

Jolma, A. *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–339.

Jothi, R. *et al.* (2008) Genome-wide identification of in vivo protein–DNA binding sites from ChIP-Seq data. *Nucleic Acids Res.*, **36**, 5221–5231.

Katainen, R. *et al.* (2015) CTCF/cohesin-binding sites are frequently mutated in cancer. *Nature Genetics*, **47**, 818–821.

Kivioja, T. *et al.* (2012) Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods*, **9**, 72–74.

Korhonen, J. *et al.* (2009) MOODS: fast search for position weight matrix matches in DNA sequences. *Bioinformatics*, **25**, 3181–3182.

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

Li, H. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li, X.Y. *et al.* (2011) The role of chromatin accessibility in directing the widespread, overlapping patterns of drosophila transcription factor binding. *Genome Biol.*, **12**, R34.

Mansour, M.R. *et al.* (2014) An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science*, **346**, 1373–1377.

Mathelier, A. *et al.* (2013) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **42**, D142–147.

McDaniell, R. *et al.* (2010) Heritable individual-specific and allele-specific chromatin signatures in humans. *Science*, **328**, 235–239.

Mukherjee, S. *et al.* (2004) Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat. Genet.*, **36**, 1331–1339.

Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Rhee, H.S. and Pugh, B.F. (2011) Comprehensive genome-wide protein–DNA interactions detected at single-nucleotide resolution. *Cell*, **147**, 1408–1419.

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

Sabarinathan, R. *et al.* (2015) Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **532**, 264–267

Valouev, A. *et al.* (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods* **5**, 829–834.

Wang, H. *et al.* (2012) Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res.*, **22**, 1680–1688.

Wang, L. *et al.* (2014) MACE: model based analysis of ChIP-exo. *Nucleic Acids Res.*, **42**, e156–e156.

Waszak, S.M. *et al.* (2014) Identification and removal of low-complexity sites in allele-specific analysis of chip-seq data. *Bioinformatics*, **30**, 165–171.

Zhang, Y. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.