Genome analysis

SigSeeker: a peak-calling ensemble approach for constructing epigenetic signatures

Jens Lichtenberg¹*, Laura Elnitski² and David M. Bodine¹

¹Genetics and Molecular Biology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA and ²Translational and Functional Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Epigenetic data are invaluable when determining the regulatory programs governing a cell. Based on use of next-generation sequencing data for characterizing epigenetic marks and transcription factor binding, numerous peak-calling approaches have been developed to determine sites of genomic significance in these data. Such analyses can produce a large number of false positive predictions, suggesting that sites supported by multiple algorithms provide a stronger foundation for inferring and characterizing regulatory programs associated with the epigenetic data. Few methodologies integrate epigenetic based predictions of multiple approaches when combining profiles generated by different tools.

Results: The SigSeeker peak-calling ensemble uses multiple tools to identify peaks, and with user-defined thresholds for peak overlap and signal strength it retains only those peaks that are concordant across multiple tools. Peaks predicted to be co-localized by only a very small number of tools, discovered to be only marginally overlapping, or found to represent significant outliers to the approximation model are removed from the results, providing concise and high quality epigenetic datasets. SigSeeker has been validated using established benchmarks for transcription factor binding and histone modification ChIP-Seq data. These comparisons indicate that the quality of our ensemble technique exceeds that of single tool approaches, enhances existing peak-calling ensembles, and results in epigenetic profiles of higher confidence.

Availability and implementation: http://sigseeker.org

Contact: lichtenbergj@mail.nih.gov

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

To gain insight into how a genome guides the building of a complete organism, the Encyclopedia of DNA Elements (ENCODE) projects have mapped biochemical activity signatures to various genomes, linked them to epigenetic features, and explored these features in terms of gene regulation (Dunham et al., 2012; Yue et al., 2014). Epigenetics refers to the study of heritable, reversible modifications that are critical to the organization of chromatin and regulation of gene expression across the genome without changing the underlying primary nucleotide sequence (Bird, 2007; Richards, 2006). The most well characterized epigenetic alterations are DNA methylation and histone modifications, which influence transcription factor binding with direct implications on gene regulation. A high-throughput sequencing adaptation of chromatin immunoprecipitation (ChIP-Seq) is often used to map epigenetic marks as well as transcription factor binding on a genome-wide scale. While the general experimental approach used to characterize transcription factor binding, methylation and histone modification is conceptually very similar, the analyzed data are considerably different. Transcription factor binding sites are expected to be relatively narrow, typically...
ranging from 5-25 bp, whereas regions subjected to histone modifications are at least 145 base pairs in size, which corresponds to the size of a nucleosome.

The vast amounts of epigenetic data collected to characterize genomic regulation in human, mouse, and many other organisms has led to a surge in computational techniques to analyze those data. Over the last several years, more than 50 approaches to determine regions enriched for epigenetic modifications (peak calling) have been published (Supplementary Information S1). They have been applied to a variety of cell types and epigenetic features, often with discordant results. Several surveys have been conducted to characterize and compare the techniques available (Barski and Zhao, 2009; Chen et al., 2012; Laajala et al., 2009; Leleu et al., 2010; Malone et al., 2011; Micsinai et al., 2012; Park, 2009; Pepke et al., 2009; Rye et al., 2011; Wilbanks and Facciotti, 2010). They determined that the existing approaches differ strongly in terms of the underlying algorithms as well as the predicted results and consequently their result quality.

To determine which approaches yield the highest value data, it is necessary to compare them against common benchmarks (Micsinai et al., 2012; Rye et al., 2011). For example, the performance of high-throughput sequencing methods for determining the presence of epigenetic or transcription factor binding can be tested using experimental validation via qPCR, proximity to computationally predicted binding sites or manual inspection of the profiles. Since unique PCR primer pairs must be designed to characterize each site, qPCR can only validate a small percentage of the predictions and may introduce a bias towards certain types of sites. Since benchmarks using computationally predicted binding sites for validation may introduce a large number of false positives and false negative rates, recent benchmarks have focused on the visual inspection of several sites to label regions with presence and absence of binding (Hocking et al., 2017; Micsinai et al., 2012; Rye et al., 2011).

Rye et al. compared five peak-calling approaches and revealed that the inclusion of background control data can limit the number of false positive predictions, however none of the tools returned peak lists with satisfactory benchmarking metrics, due to false positive results, which remained high (Rye et al., 2011). Using histone modification data, Micsinai et al. performed a validation discriminant analysis to optimally select validation sites for an additional benchmark set (Micsinai et al., 2012). Hocking et al. used narrow histone marks (H3K4me3) and broad histone domains (H3K36me3) to characterize seven existing unsupervised ChIP-Seq peak callers and compared it against a supervised approach (Hocking et al., 2017).

Both Rye et al. and Micsinai et al. noted that the benchmarking quality of peak-calling algorithms had improved considerably in more recent approaches and that the techniques were also improving in terms of histone modification data. They also recommend combining results from different peak-calling approaches to achieve more reliable peak definitions (Micsinai et al., 2012; Rye et al., 2011). Hocking et al. further determined that the visual inspection of only a few regions by a user and the subsequent use of the labelled regions in a supervised peak calling algorithm can lead to higher accuracy for the overall profiling (Hocking et al., 2017).

Despite the importance of these observations, only one peak-calling ensemble has been made available to date. This ensemble, the Peak Finder Metaserver (PFMS), aggregates regions reported by several peak-calling approaches that overlap by at least one base pair (bp) and combines them into a single output with normalized scores (Kruczyk et al., 2013). While it is important to consider peak similarity from a positional perspective, peak-prediction tools often generate co-localized peaks with intensity scores that differ greatly, despite normalization (Supplementary Information S2). While calculated differently across various peak-calling tools, intensity relates to the number of high-throughput sequencing reads that map to a genomic region. We propose that positional co-localization should be complemented by an assessment of the correlations among the intensity profiles of co-localized peaks, to remove false positive predictions, in a manner similar to the irreproducible discovery rate (IDR) (Landt et al., 2012). IDR (https://sites.google.com/site/anshulkun/dajo/projects/idr) measures the reproducibility of peak-calling replicates in transcription factor profiling experiments with narrow predicted sites, to suggest parameter guidelines for the analysis, using a copula mixture model that fits a curve by comparing the scores assigned to ranked pairs of peaks between replicates (Li et al., 2011).

The SigSeeker approach presented here addresses the necessity of epigenetic peak-calling ensembles established by Rye et al. and Micsinai et al. and combines it with the concept of intensity correlations. SigSeeker enhances existing peak comparisons by including peak overlaps of varying minimal width (a positional filter) and peak-intensity correlation models (a quantitative filter) to refine peak predictions in an unsupervised approach. Instead of generating optimized peak-calling results for a single tool or a set of all overlapping peaks from multiple tools, SigSeeker combines positional and quantitative filtering as part of a peak-calling ensemble. The resulting predictions contain those peaks identified by a user-defined number of tools at the same or overlapping position in the genome. The peaks furthermore retain a comparable level of intensity across different prediction algorithms, and consequently represent a consolidated and high quality epigenetic dataset. Using accepted benchmarks for epigenetic peak calling, we show that a positional filter greatly improves results, while using different suggested minimal overlap parameters for histone modification and transcription factor binding data. We also demonstrate that the application of a quantitative filter results in further improvement of peak-calling results through enforcing intensity correlations between different peak-calling approaches.

2 Materials and methods
2.1 Methodological pipeline

The input for the SigSeeker peak-calling ensemble is one or more sets of sequencing reads that have been previously aligned against a reference genome. Although approximately 95% of peak-calling approaches support the analysis of samples against input or background control data, a few approaches analyze sample sequencing reads only. These approaches allow analysis of epigenetic experiments with very limited biological materials or time-series comparisons that address changes in epigenetic profiles. Consequently, the SigSeeker peak-calling ensemble requires at least one set of sample sequencing reads and allows control read sets as an option.

The SigSeeker ensemble is open source and can be used with any peak-calling approach whose input (1) consists of aligned sequencing reads in the BED file format or (2) is in a form that can be converted to BED (e.g. SAM, BAM, ELAND) and that generates peak lists in BED or compatible formats. The webserver at http://sigseeker.org currently supports commonly used and widely reviewed approaches for sharp peaks (e.g. transcription factors) and broad peaks (e.g. epigenetic domains). cisGenome (Ji et al. 2008), ERANGE (Mortazavi et al., 2008), MACS (Zhang et al., 2008), MACS2 (Liu, 2015), SWEMBLE (Wilder, 2010) are designed for the detection of sharp peaks, while hidden domains (Starmer and
2.2 Peak co-localization approach

To establish the set of peaks supported by an ensemble of peak-calling techniques, it is necessary to compare the co-localization of peaks predicted by different tools. Positional similarity between peaks is traditionally based on minimum 1 bp overlap (Hogart et al., 2012; Kruczyk et al., 2013), i.e. if peaks from different tools overlap even 1 bp, the regions will be intersected. In the context of SigSeeker peak integration, we allow user-defined overlap thresholds greater than or equal to 1 bp. An intersection of multiple peaks is thus retained if it covers a user-defined minimum number of base pairs and if the peaks that make up the intersection are drawn from a (user-defined) minimum number of analysis tools (Supplementary Information S3). Each of these refined peaks can be assigned a new intensity score, e.g. the number of tools supporting the peak, the average intensity reported by each tool for the peak or a novel score based on an underlying intensity model described below. In addition to intersections with the choice of adjustable overlap lengths, we implemented an option to allow for peak unions, where independent of a specific overlap the results of various peak callers are merged into a single comprehensive peak list.

While it is necessary to consider peak similarity from a positional perspective, peak-prediction tools often generate co-localized peaks whose original intensities differ greatly. Using the intensity profiles of overlapping peaks reported by all tools, we use the total least squares approximation to fit a line to all reported scores (Markovsky and Huffel, 2007). While linear regression represents residuals as a vertical line from an error-free observation to a fitted curve, total least squares approximation to fit a line to all reported scores (Markovsky and Vanlloon, 1980). The parameters of the model enable us to normalize and characterize the correlation between results from different tools and allow us to highlight peaks that differ significantly from the model as outliers. This allows us to prune the set of predicted peaks by removing peaks with discordant scoring. Such quantitative filtering is of particular importance for peaks that have less-than-complete overlap, as they may represent neighboring, distinct epigenetic marks (Supplementary Information S2). Outliers are characterized as intensity measures outside upper and lower quantiles, which are determined by an estimated underlying distribution of the peak intensities, similar to the Cho et al. approach for mass spectrometry (Cho et al., 2008). The quantiles are applied to filter the lists of intersecting peaks of multiple tools into a single result (Fig. 1).

This approach ensures retention of not only peaks with strong scores across all tools but also those detected by multiple approaches with lower, but still relevant, intensities. Scores of intersecting peaks with similar intensity profiles are assigned in accordance with the regression model line.

2.3 Compiling data outputs

Since it cannot be known ahead of time which combination of tools in an ensemble provides the best results, outputs of various tool combinations are combined into a final result. For this purpose we are joining peak sets predicted by all k-combinations of the set of n available peak calling tools, thus each resulting peak list is a merged list of C(n, k) peak lists. For example, a 2-combination of peak lists using MACS, MACS2 and ERANGE would build the intersections of MACS and MACS2, MACS and ERANGE, as well as MACS2 and ERANGE. These three intersections would then be merged into a single peak list.

2.4 Program implementation and availability

SigSeeker’s peak location and intensity integration schemes are available at http://sigseeker.org, and the source code is available through GitHub at https://github.com/lichtenj/sigseeker (Supplementary Information S4). Each peak-calling technique employed in the ensemble is installed in accordance with the instructions provided in its manual. Peak overlap for positional filtering is realized using the multi-inter function of the pracma package. The k-combinations join function, is implemented using the Math::Combinatorics package in Perl and the merge program in BedTools2. Visualization and characterization are carried out in R using the ggplot2 and mcr packages. The intersection and filtering functionality implemented as part of SigSeeker operates with a complexity of O(N^k) (Supplementary Information S7).

A web interface for SigSeeker is implemented via Apache web server (Version 2.2.22) and uses Perl CGI (Version 5.14.2 for Perl and Version 3.52 for the Perl CGI Module) and JavaScript (Version 1.5 and above). The user characterizes the project and the desired

Magnuson, 2016) and SICER (Zang et al., 2009) operate best when applied to broader epigenetic peaks. CCAT (Xu et al., 2010) has explicit settings for the detection of narrow and broad peaks.

Each peak-calling tool appropriate for the experimental design is run with the parameter space supplied by the user, and genomic locations and intensities are collected from the individual tools and the output for each tool is compared to the output of all other tools. Peak lists resulting from these comparisons can then be further analyzed based on relative location to genomic landmarks (e.g. PinkThing (Nielsen et al., 2013)), associated expression profiles (e.g. ChIPseeker (Yu et al., 2015)), or embedded regulatory binding sites (e.g. GREAT (McLean et al., 2010)). All data visualization is performed in R (version 3.1.2).
2.5 Performance bench-marking

The performance of individual peak-calling approaches has been evaluated using benchmark datasets for transcription factor binding (Rye et al., 2011) and histone modification (Hocking et al., 2017; Micsinai et al., 2012) datasets. Rye et al. conducted a ChIP-Seq analysis for the transcription factors in human K562 (NRSF and MAX) and GM12878 (SRF) cells. The resulting profiles were manually inspected and peaks with strong correspondence to the transcription factor binding regions were chosen as positive sites, while regions clearly absent of binding characteristics marked negative sites.

Mikkelsen et al. (2007) published a survey of the narrow H3K4me3 histone mark and the broad H3K27me3 histone mark in embryonic mouse stem cells. Micsinai et al. visually inspected these data and evaluated a relevant subset for true presence or absence of the histone mark. Hocking et al. evaluated H3K4me3 and H3K36me3 data for a variety of human cell types that is available through the McGill Epigenomes Portal (http://epigenomes.portal.ca). We added another benchmark by visually inspected highly significant profiles of H3K9me3 in mouse embryonic stem cells using a NimbleGen mouse tiling array (ChIP-chip Data: GSM360357, (Schnetz et al., 2009)) and labeled absent and present binding sites to evaluate corresponding ChIP-Seq peak calling profiles (Read Data: GSM307621, (Mikkelsen et al., 2007)).

SigSeeker and each of the existing single tool approaches run with the default parameters was evaluated using the transcription factor benchmarks and histone modification benchmarks. Presence of a feature at a site (‘validated positive’) or an absence of a feature (‘validated negative’) can be used to characterize each predicted peak by a given approach as correct (TP: true positive, TN: true negative) or incorrect (FP: false positive, FN: false negative). Using these characterizations, it is possible to assign sensitivity and specificity measures to each peak calling profile. For the purpose of analyzing the performance of the SigSeeker ensemble approach, we use the Informedness (Powers, 2011) and the Jaccard Index Precision (Steinhauser et al., 2016).

\[
\text{Informedness} = \text{Sensitivity} + \text{Specificity} - 1
\]

\[
\text{Jaccard Index Precision} = \left( \frac{1}{N} \sum_{i=0}^{N} \frac{|A_i \cap B_{i0}|}{|A_i \cup B_{i0}|} \right)
\]

The norms used as part of the Jaccard Index Precision metric represent the genomic length of the intersection and union of a peak with an overlapping validated binding site respectively. It penalizes imprecise peak calling, i.e. those peaks that span regions outside the validated site (over-calling) and those that map represent only a small percentage of the validated site (under-calling).

3 Results

We used the ChIP-Seq benchmarking datasets for narrow peak calling (Micsinai et al., 2012; Rye et al., 2011) and broad epigenetic domains (Hocking et al., 2017; Mikkelsen et al., 2007; Schnetz et al., 2009) to evaluate the peak-calling performance of SigSeeker. We also assessed the improvements gained by positional peak integration and quantitative ensemble techniques. Positional integration is determined by peak co-localization using user-defined peak overlap lengths, and quantitative integration is implemented via the total least square approximation. Our peak calling pipeline included the tools cisGenome, ERANGE, MACS (Version 1.4 and Version 2.1), SWEMBLE, and CCAT for narrow peak profiles and SICER, hid-domains, and CCAT for broad epigenetic domains. The tools were chosen due to their wide usage in their respective areas of peak profiling.

Our analysis found that the union (i.e. join) of output from peak calling tools produced an increased number of overall retained peaks, however, significantly fewer peaks were retained as the number of tools required to support the predictions increased (Fig. 2).
The number of retained peaks was further reduced by (1) increasing the minimal overlap threshold between peaks predicted by different tools and (2) requiring conformation to the intensity scores based total least square approximation for the peaks. Under these conditions, joining the filtered peak lists for all combinations of a specific number of tools in a union increased the number of retained peaks to a lesser extent. For tool sets of high cardinality, as well as for large overlaps, the number of retained peaks quickly approached zero.

In confirmation of the observation that unsupervised peak calling approaches yielded increased predictions in only some experiments (Hocking et al., 2017; Micsinai et al., 2012; Rye et al., 2011), we found that none of the individual peak-calling approaches integrated into SigSeeker emerged as clearly superior across all data sets (Table 1 and Supplementary Information 6 (Tables S4–S9)). ERANGE was associated with the best Informedness for detecting MAX binding sites in K562 cells and MACS2 profiles had the best Jaccard Index Precision (Table 1). MACS2 showed the best Informedness and Jaccard Index Precision for NRSF binding in K562 cells, while SWEMBLE had the worst Jaccard Index Precision (Supplementary Information 6 (Table S4)). For SRF binding in GM12878 ERANGE reported the best Informedness, while MACS2 was associated once again with the best Jaccard Index Precision and SWEMBLE with the worst (Supplementary Information 6 (Table S5)). In general, the Rye et al. benchmarks for transcription factor peak calling indicate that either MACS2 or ERANGE provide the best Informedness, while MACS2 provides the optimum width peak predictions in all cases.

For both the narrow and broad histone benchmarks, hiddendomains, SICER and CCAT provided very high specificity measures (Supplementary Information 6 (Tables S6–S9)). Of the broad peak callers, SICER reported the best Informedness for the narrow H3K4me3 mark with 0.32, while SWEMBL generated profiles with a slightly higher informedness score of 0.40 (Supplementary Information 6 (Table S6)). The best Jaccard Index Precision for H3K4me3 was achieved by cisGenome (0.61), while CCAT had the most precise mapping among the broad domain approaches (0.55). For the H3K27me3 histone marks, which have both narrow and domain-like properties, hiddendomains reported the best Informedness, and only cisGenome had a higher Jaccard Index Precision modeled for just a single correctly identified true positive peak (Supplementary Information 6 (Table S7)). SICER once again had the best informedness for the broad H3K9me3 and H3K36me3 marks, while profiles generated via hiddendomains had the best informedness for the H3K27me3 benchmark. The best Jaccard Index Precision was reported for CCAT in H3K4me3 as well as H3K9me3, SICER for H3K36me3 and hiddendomains for H3K27me3.

The general performance of individual approaches falls in line with the observation made by Hocking et al., that narrow unsupervised peak callers were better for identifying transcription factor binding sites and somewhat better for the detection of narrow histone marks, while broad unsupervised peak callers operated best on wide histone marks.

Using the application of the k-combination joining of all ensembles containing a specific number of tools resulted in an overall increased number of retained peaks compared to filtering alone (Fig. 2). The associated sensitivity only drops below the median of single tool approaches for 4 tool ensembles (Supplementary Information 6 (Fig. S5)). The specificity of joined ensemble results starts to exceed the median specificity of single tool approaches when 3 or more tools are included in the ensemble.

In terms of transcription factor related benchmarks, an increase in Informedness is significant for ensembles of exactly 4 tools or 2-3 tools when applying a 200bp overlap filter (Supplementary Information 6 (Fig. S6)). The Informedness associated with joined ensembles consisting of 2 tools with a 200bp overlap filter and no quantitative filter exceeds the measure for every single tool approach (Fig. 3a). A good mapping on the actual observed peak profiles as measured by the Jaccard Index Precision was also maintained for the ensemble directly, the precision of the retained mapping is best when considering the union of all 3 peak profiles and a minimum overlap of 700bp across the peaks and a quantitative filter of ±5% (Fig. 3b).

### Table 1. Benchmarking metrics for MAX binding in K562 cells in accordance with 289 genomic sites queried for MAX binding using qPCR (Rye et al., 2011)

<table>
<thead>
<tr>
<th>K562 (MAX)</th>
<th>MACS</th>
<th>MACS2</th>
<th>ERANGE</th>
<th>SWEMBLE</th>
<th>cisGenome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>22361</td>
<td>10264</td>
<td>2888</td>
<td>24815</td>
<td>22606</td>
</tr>
<tr>
<td>TP</td>
<td>186</td>
<td>176</td>
<td>171</td>
<td>224</td>
<td>184</td>
</tr>
<tr>
<td>FP</td>
<td>31</td>
<td>49</td>
<td>54</td>
<td>11</td>
<td>39</td>
</tr>
<tr>
<td>TN</td>
<td>32</td>
<td>14</td>
<td>9</td>
<td>52</td>
<td>24</td>
</tr>
<tr>
<td>FN</td>
<td>40</td>
<td>50</td>
<td>55</td>
<td>2</td>
<td>42</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.82</td>
<td>0.78</td>
<td>0.76</td>
<td>0.99</td>
<td>0.81</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.49</td>
<td>0.78</td>
<td>0.88</td>
<td>0.18</td>
<td>0.62</td>
</tr>
</tbody>
</table>

**Note:** Best scores in a category are shaded in green, while worst scores are highlighted in orange.

4 Conclusions

We show that using an ensemble, operating on the intersections of individual peak-calling results, without additional filters increases specificity and decreases sensitivity of peak-calling results but has no significant impact on the informedness of the result. The combination of a positional filter and the subsequent joining of the results for all ensemble combinations of a specific number of tools, implemented in the SigSeeker toolkit, generated peak results with significantly improved informedness for transcription factor peak calling over single peak calling approaches, while the additional application of a quantitative filter using a total least squares approximation resulted in an improved informedness for the detection of very broad histone marks (H3K36me3).

In general, when analyzing transcription factor binding profiles, our results support the use of an ensemble, composed of pairs of suitable tools requiring a minimum overlap between intersecting peaks of approximately 200bp for transcription factor binding site discovery and an overlap of 2900bp as well as a quantitative scoring model that removes peaks outside the 90th percentile broad H3K36me3 histone marks.
Fig. 3. Measure of best possible Informedness and Jaccard Index Precision for single tool approaches and unions of ensembles combining the peak lists of these tools. K562-MAX, K562-NRSF and Gm12878-SRF were chosen to generate the ‘Transcription Factor’ boxplot (a), while H3K9me3 (b) and H3K36me3 (e) ensemble results were used for broad histone marks. Different minimal overlaps provided the optimal results in all three benchmarks. ‘Transcription Factors’ did not require the addition of a quantitative filter, while a quantitative filter of 90% was used for each of the broad histone marks.

Acknowledgements

The authors would like to thank Valer Gotea, Gennady Margolin and Elisabeth Heuston for the many discussions relating to this work.

Funding

Funding was provided by the NHGRI Intramural Research Program.

Conflict of Interest: none declared.

References


