RESEARCH ARTICLE

Graph Peak Caller: Calling ChIP-seq peaks on graph-based reference genomes

Ivar Grytten[1☯]*, Knut D. Rand[2☯], Alexander J. Nederbragt[1,3], Geir O. Storvik[2], Ingrid K. Glad[2], Geir K. Sandve[1]

1 Department of informatics, University of Oslo, Oslo, Norway, **2** Department of Mathematics, University of Oslo, Oslo, Norway, **3** Department of Biosciences, University of Oslo, Oslo, Norway

☯ These authors contributed equally to this work.
* ivargry@ifi.uio.no

Abstract

Graph-based representations are considered to be the future for reference genomes, as they allow integrated representation of the steadily increasing data on individual variation. Currently available tools allow *de novo* assembly of graph-based reference genomes, alignment of new read sets to the graph representation as well as certain analyses like variant calling and haplotyping. We here present a first method for calling ChIP-Seq peaks on read data aligned to a graph-based reference genome. The method is a graph generalization of the peak caller MACS2, and is implemented in an open source tool, *Graph Peak Caller*. By using the existing tool *vg* to build a pan-genome of *Arabidopsis thaliana*, we validate our approach by showing that Graph Peak Caller with a pan-genome reference graph can trace variants within peaks that are not part of the linear reference genome, and find peaks that in general are more motif-enriched than those found by MACS2.

Author summary

The expression of genes is a tightly regulated process. A key regulatory mechanism is the modulation of transcription by a class of proteins called transcription factors that bind to DNA in the spatial proximity of regulated genes. Determining the binding locations of transcription factors for specific cell types and settings is thus a key step in understanding the dynamics of normal cells as well as disease states. Binding sites for a given transcription factor are typically obtained through an experimental technique called CHiP-seq, in which DNA binding locations are obtained by sequencing DNA fragments attached to the transcription factor and aligning these sequences to a reference genome. A computational technique known as *peak calling* is then used to separate signal from noise and predict where the protein binds. Current peak callers are based on linear reference genomes that do not contain known genetic variants from the population. They thus potentially miss cases where proteins bind to such alternative genome sequences. Recently, a new type of reference genomes based on graph representations have become popular, as they are able to also incorporate alternative genome sequences. We here present *Graph Peak Caller*, the first peak caller that is able to exploit such graph representations for the detection of

OPEN ACCESS

Citation: Grytten I, Rand KD, Nederbragt AJ, Storvik GO, Glad IK, Sandve GK (2019) Graph Peak Caller: Calling ChIP-seq peaks on graph-based reference genomes. PLoS Comput Biol 15(2): e1006731. https://doi.org/10.1371/journal.pcbi.1006731

Editor: Alice Carolyn McHardy, Helmholtz-Zentrum fur Infektionsforschung GmbH, GERMANY

Received: April 6, 2018

Accepted: December 19, 2018

Published: February 19, 2019

Data Availability Statement: All files are available from the Genomic Hyperbrowser at https://hyperbrowser.uio.no/graph-peak-caller/static/graph_peak_caller_data.html.

Funding: This work was supported by the Centre for Computational Inference in Evolutionary Life Science (CELS), but otherwise received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

transcription factor binding locations. Using a graph-based reference genome for *Arabidopsis thaliana*, we show that our peak caller can lead to better detection of transcription factor binding locations as compared to a similar existing peak caller that uses a linear reference genome representation.

This is a *PLOS Computational Biology* Methods paper.

Introduction

Transcription factors are known to play a key role in gene regulation, and detecting regions associated with transcription factor binding is an important step in understanding their function. The most common technique used to detect transcription factor binding sites is *ChIP-seq*, combining chromatin immunoprecipitation (ChIP) assays with sequencing (seq). A ChiP-seq experiment involves obtaining DNA *fragments* that bind to the transcription factor of interest and sequencing arbitrary ends of these fragments, yielding short reads. Obtaining putative binding regions from these reads is done using computational techniques known collectively as performing *peak calling*. Several *peak callers*, programs to perform peak calling, have been developed for this purpose, for example MACS2 [1] and SPP [2] [3]. Common for all current peak callers is that they take reads mapped to a *linear* reference genome, such as GRCh38, as input.

Graph-based reference genomes offer a way to include known variants within a population in the reference structure [4]. The software package *vg* supports mapping reads to a graph-based reference genome with potentially increased accuracy [5, 6] as compared to mapping reads to a standard linear reference genome using tools like BWA [7] or Bowtie [8]. Several types of genomic analyses, such as variant calling and haplotyping, can now be performed using graph-based references [5, 6]. However, no tool currently exists for performing peak calling on graph-based references.

Results

We present *Graph Peak Caller*, a first method for detecting transcription factor binding events from ChIP-seq reads mapped to a graph-based reference genome. Graph Peak Caller is based on the same principles used by MACS2 (see Fig 1 for an overview), and is able to call peaks with or without a set of control alignments. For the case of a graph that merely reflects a linear reference genome, our peak-caller produces the same results as MACS2. As input, it supports alignments in the Graph Alignment/Map format (GAM) from *vg*, as well as reads represented as genomic intervals using the Offset Based Graph Python package [9]. Graph Peak Caller can be run from the command line, and is also available through Galaxy at https://hyperbrowser.uio.no/graph-peak-caller. In the Github repository at http://github.com/uio-bmi/graph_peak_caller, we provide a simple tutorial on how to use *vg* and Graph Peak Caller to go from raw ChIP-seq reads to peaks.

The output from Graph Peak Caller consists of graph intervals, but the tool is also able to transform these into approximate positions on a linear reference genome (by projecting them to the nearest position on the linear reference genome), making it possible to analyse detected

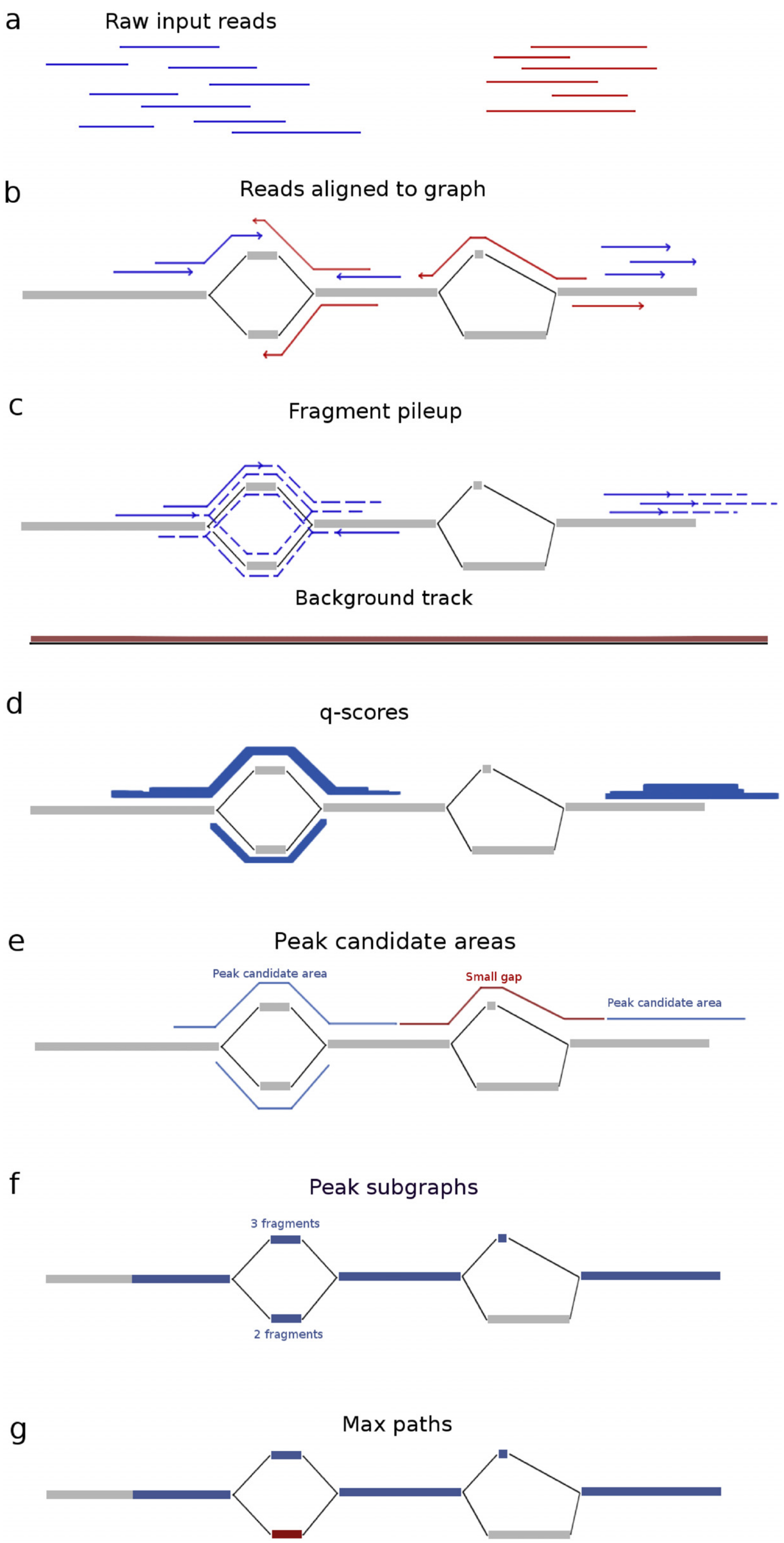

Fig 1. Overview of how Graph Peak Caller works. Example of peak calling on an example graph (nodes in gray, edges in black). After raw reads (a) in the form of input reads (blue) and control reads (red) have been mapped to the graph-based reference genome and filtered based on mapping quality (b), the fragment pileup is created (c) by extending the forward input alignments and reverse input alignments (extensions shown as dotted lines) along all possible paths in their corresponding direction. A background track is created by projecting the alignments resulting from the control

reads onto a linear path and calculating a local average of read counts. Then the linear track is projected back to the graph again. The fragment pileup is treated as *counts* and the background track as *rates* in a Poisson-distribution, and p-values are computed for each position for the observed count, given the corresponding rate. Adjusted q-values are computed to control the false discovery rate (d) (figure shows q-scores, which are -$log_{10}(qvalue)$). The q-values are thresholded on a user-defined threshold (default 0.05), resulting in a set of peak candidate areas with gaps between them (e). Small gaps are filled, resulting in a set of peak subgraphs (connected subgraphs)(f). Graph Peak Caller finds a single "maximum path" (g, maximum path in blue) through each peak subgraph by selecting the path that has the highest number of input reads mapped to it.

https://doi.org/10.1371/journal.pcbi.1006731.g001

peaks further using existing "linear" approaches. Graph Peak Caller is also able to output peak candidates for differentially expressed peaks.

To showcase and test Graph Peak Caller, we chose to perform peak calling on *Arabidopsis thaliana*, as the 1001 Genomes Project for *A. thaliana* makes it possible to build a high-quality reference graph with a high density of variants (on average one SNP or indel for every 9 base pairs, compared to one SNP or indel for every 27 base pairs in the human 1000 Genomes Project). This graph is pruned by vg in order to make it possible to create indices for read alignment within reasonable time, reducing the number of variants to on average one variant for every 18 base pairs.

We called peaks on this graph-based reference genome for *A. thaliana* and compared the results to peaks called on the *Tair10* [10] linear reference genome by MACS2 (Methods). Table 1 and Fig 2 shows an overview of peaks found by Graph Peak Caller and MACS2. Most of the peaks found by one peak caller are also found by the other. Among these, the peaks found by Graph Peak Caller are slightly more enriched for DNA-binding motifs than the peaks found by MACS2 for all transcription factors, except SEP3, where the numbers are the same. Fig 3 shows an example of a peak detected on chromosome 1, illustrating how both peak callers find a peak, but only the peak found by Graph Peak Caller has a match against the motif.

For all transcription factors, except ERF115, the peaks uniquely found by Graph Peak Caller (not overlapping peaks found by MACS2) are more enriched for DNA-binding motifs than the peaks uniquely found by MACS2. In aggregate, the ratios of motif-matches of the uniquely found peaks are 6.33% for Graph Peak Caller and 5.74% for MACS2, yielding a z-value for the difference of 2.08 and a p-value of 1.9% using a one-sided z-test for difference in population proportions [11]. For the peaks found by both peak callers, Graph Peak Caller has a few more peaks with with motif match, but the difference is not statistically significant ($p = 0.42$). Fig 4

Table 1. Overview of peaks reported by Graph Peak Caller and MACS2 on *A. thaliana* for 5 transcription factors (TFs). *Total* is the total number of peaks reported by the peak caller, *shared* is the number of peaks that overlap with a peak from the other peak caller, and *unique* are peaks reported by one peak caller and not the other. In the categories *shared* and *unique*, both the number of peaks with motif match (the number before the /) and the number of peaks found are shown (percent of peaks with motif match are shown in parentheses). All peaks have been trimmed to 120 base pairs around the peak summit (position in peak with lowest q-value), to make the comparison clearer.

	Graph Peak Caller					**MACS2**				
	Total	**Shared**		**Unique**		**Total**	**Shared**		**Unique**	
ERF115	24976	3121/21364	14.61%	284/3612	7.86%	23167	3108/21364	14.55%	214/1803	11.87%
SEP3	14976	989/11982	8.25%	125/2994	4.18%	15517	978/11982	8.16%	148/3535	4.19%
AP1	16797	764/13405	5.70%	102/3392	3.01%	17030	754/13405	5.62%	94/3625	2.59%
SOC1	15502	1676/14297	11.72%	116/1205	9.63%	16407	1681/14297	11.76%	142/2110	6.73%
PI	17518	1790/14084	12.71%	300/3434	8.74%	16084	1793/14084	12.73%	152/2000	7.60%
SUM	89769	8340/75132	11.10%	927/14637	6.33%	88205	8314/75132	11.07%	750/13073	5.74%

https://doi.org/10.1371/journal.pcbi.1006731.t001

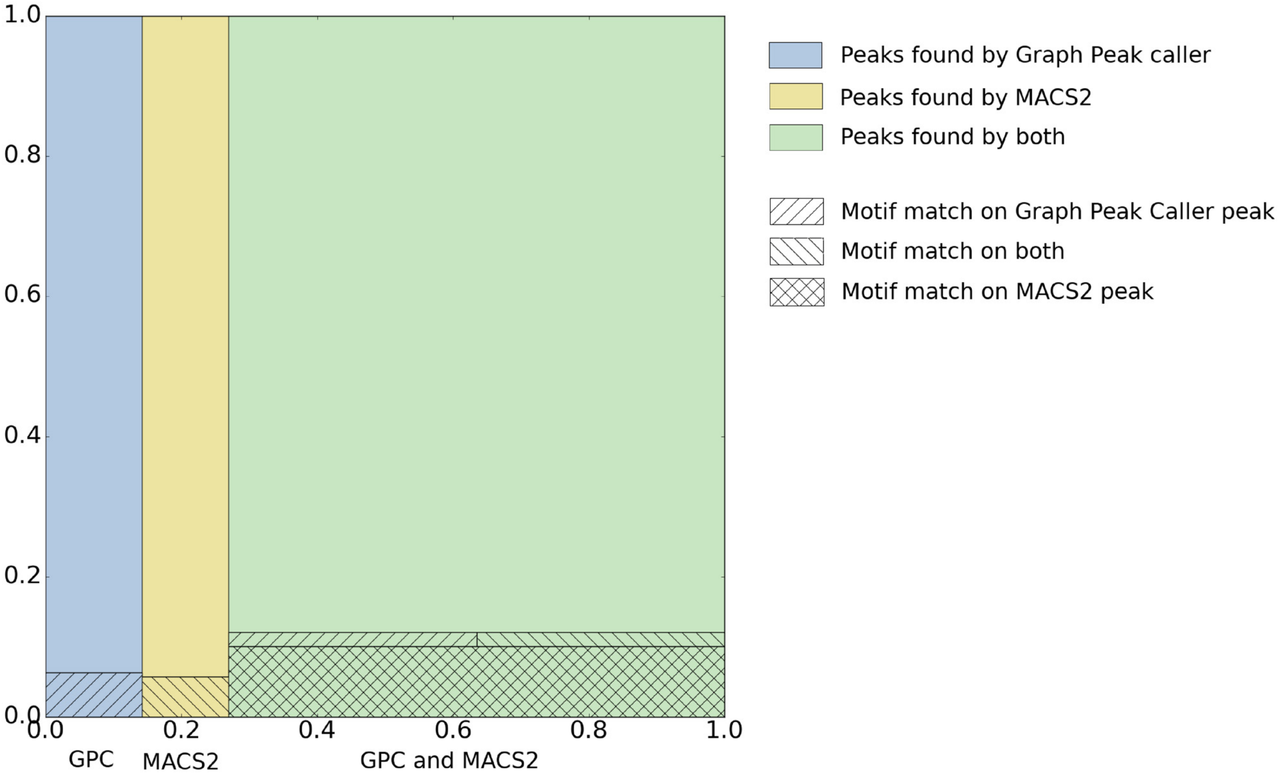

Fig 2. Visual overview of all peaks found by Graph Peak Caller and MACS2 on *A. thaliana*. Showing the proportions of peaks found by both peak callers and peaks found uniquely by each peak caller. All rectangles are scaled to the proportion between the number of peaks they represent and the number of peaks in the union of the peaks found by both peak callers. The colors denote which peak caller found the peaks, while the hatching denotes which peaks had a motif match. Note that the sequence of a shared peak can differ between peak callers and thus the motif count among the shared peaks differs between the peak callers.

https://doi.org/10.1371/journal.pcbi.1006731.g002

shows one of the cases on chromosome 1 where Graph Peak Caller detects a peak that is not detected by MACS2, due to that the input reads are aligned against an indel that is not part of the linear reference genome.

The peaks found uniquely by Graph Peak Caller have more than twice the number of basepairs not part of the linear reference genome, compared to the peaks found by Graph Peak Caller that also have been found by MACS2 (S2 Table). Fig 5 shows the proportion of peaks enriched for DNA-binding motifs for the peaks that are uniquely found by each peak caller (for a similar plot including all peaks, see S1 Fig). As seen in the figure, Graph Peak Caller has a better correspondence between high peak scores and motif enrichment for all transcription factors, except ERF115.

To validate that our peak caller works on other species, we repeated the experiment on similar datasets for *Drosophila melanogaster* and human (see S2 Fig and S1 Table). In both cases Graph Peak Caller got a higher ratio of motif-matches on uniquely found peaks, but only statistically significant for *D. melanogaster* ($p_{\mathrm{human}} = 24.9\%$, $p_{\mathrm{D.melanogaster}} = 0.42\%$).

We also checked how MACS2 performs when using reads aligned by vg that have been projected to the linear reference genome using *vg surject*, in order to see how MACS2 performs when taking advantage of the improved mapping to a graph-based reference genome. Using these alignments, MACS2 performs better than when using using alignments from BWA, but is still slightly outperformed by Graph Peak Caller on motif matches (S3 Table).

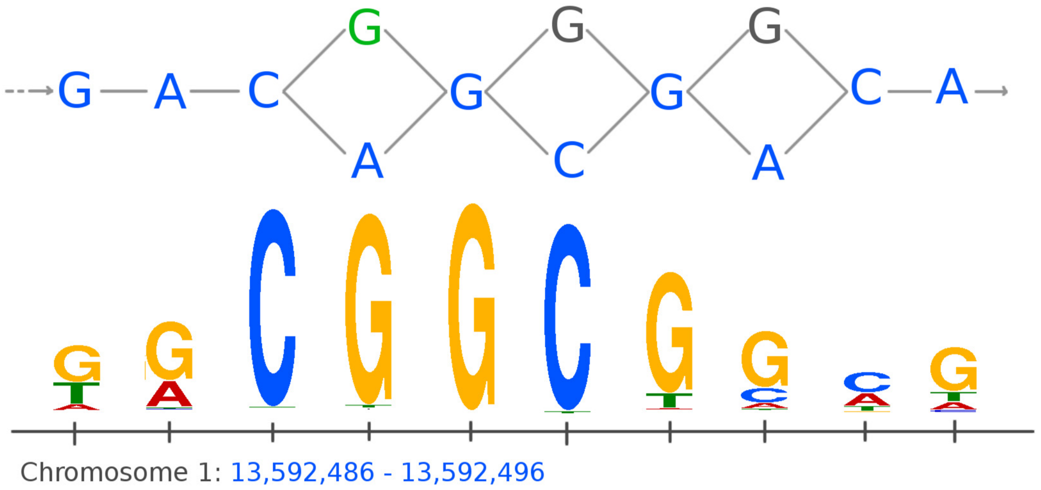

Fig 3. Example of a motif match for a peak following a variant not part of the linear reference genome. Showing part of the graph-based reference genome for *A. thaliana* on chromosome 1 (top) with the linear reference genome represented by blue nodes (bases). The figure shows a peak found for the ERF115 transcription factor that matches a DNA-binding motif. The peak detected by Graph Peak Caller follows the linear reference genome (blue nodes) except for the first SNP shown (green node), where the peak follows the green G instead of the blue A, making this a significant match against the DNA-binding motif (shown as a sequence logo at the bottom). The peak detected by MACS2 does not have a significant motif match. All alignments intersecting with this binding site are present in a common haplotype among the haplotypes used to build the reference graph (S2 Appendix).

https://doi.org/10.1371/journal.pcbi.1006731.g003

Alignment bias evaluation

A problem with using motif enrichment to compare the peaks detected by MACS2 and Graph Peak Caller is that the two peak callers use different sets of alignments. Thus, we checked whether vg tends to align reads to the graph reference genome that more often end up in alignments with motif match than what BWA does. We found no such tendency. In aggregate for

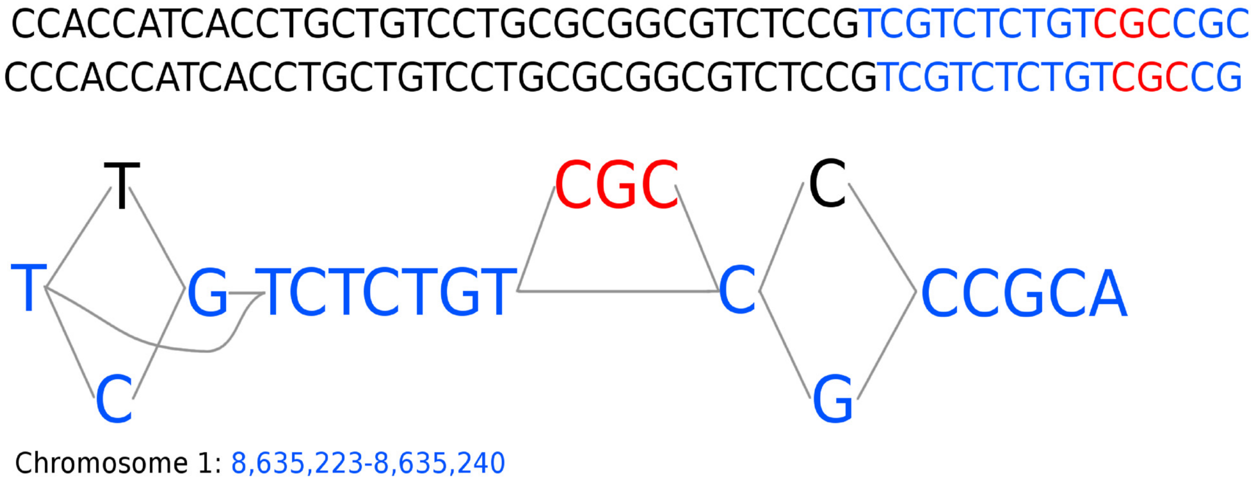
CCACCATCACCTGCTGTCCTGCGCGGCGTCTCCGTCGTCTCTGTCGCCGC
CCCACCATCACCTGCTGTCCTGCGCGGCGTCTCCGTCGTCTCTGTCGCCG
T
T
C
G
TCTCTGT
CGC
C
C
G
CCGCA
Chromosome 1: 8,635,223-8,635,240

Fig 4. Example of part of a peak detected by Graph Peak Caller and not by MACS2. Showing part of the graph-based reference genome for *A. thaliana* on chromosome 1, containing two SNPs (black/blue) and an indel (red) that are not part of the linear reference genome (blue). From the raw ChIP-seq data for the ERF115 transcription factor (NCBI SRA SRR931836) two reads (shown at the top) align perfectly to the graph-based reference genome and Graph Peak Caller is able to detect a peak in this area. Mapping to the linear reference genome does not give sufficiently high mapping score, and so the peak is missed by MACS2.

https://doi.org/10.1371/journal.pcbi.1006731.g004

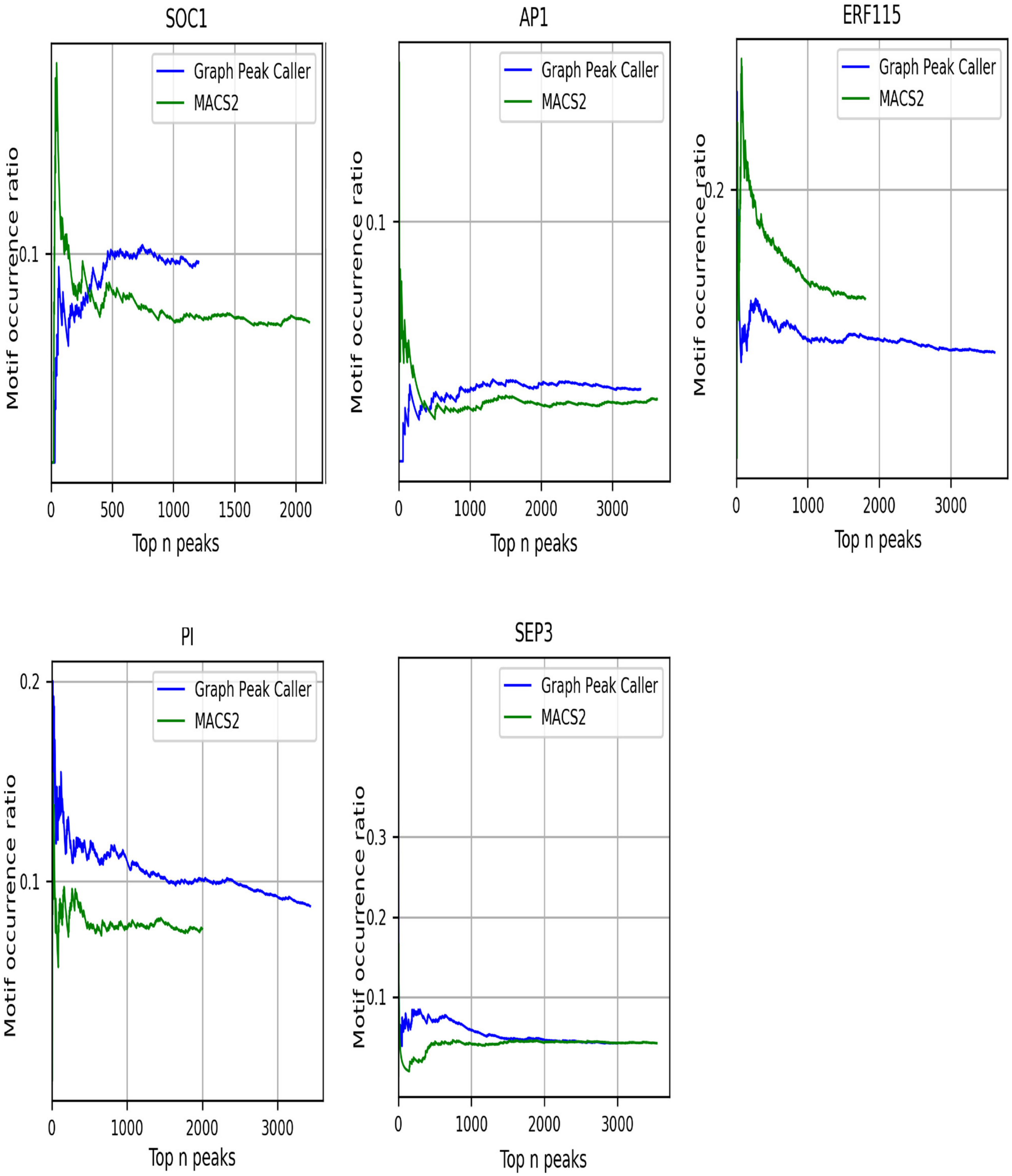

Fig 5. DNA-binding motif enrichment plots. The proportion of peaks enriched for a DNA-binding motif (Y-axis) when iteratively including more peaks from the set of uniquely found peaks for each peak caller, sorted descending on score (X-axis).

https://doi.org/10.1371/journal.pcbi.1006731.g005

the peaks detected on *A. thaliana*, we found on average 0.0208 motif matches per graph alignment, and and on average 0.0214 motif matches per linear reference genome alignment.

We also checked whether the higher motif match among the peaks detected by Graph Peak Caller might be a result of peaks coming from artificially clustered alignments, something which can be the case if there are complex regions in the graph reference containing the motif. Such regions could allow for many different reads to align to the same place, resulting in high-scoring peaks with motif match. We found that 99.6% of the motif matches among the peaks detected by Graph Peak Caller are consistent with a single haplotype. For the unique peaks detected by Graph Peak Caller, 98.6% of the motif matches were consistent with a single haplotype. For each detected peak, we also checked what percentage of the reads among those aligned to the peak were aligned to one or two haplotypes (allowing for diploidy). In aggregate, we found that 55.6% of the unique peaks detected by Graph Peak Caller had all reads compatible with maximum two haplotypes. We found that the peaks having alignments from more than two haplotypes were more often enriched for motif, which might indicate that peaks coming from mismapped reads to high complexity areas inflate the motif match percentage. However, when we removed those peaks from the analysis, the ratio of motif peaks matching motif among the peaks detected by Graph Peak Caller is still significantly higher than the ratio of peaks matching motif among the peaks detected by MACS2. In addition, the peaks found by MACS2 are also more motif enriched when the vg alignments overlapping those peaks are consistent with more than two haplotypes, indicating that the correspondence between multiple haplotypes and motif enrichment is not solely attributable to mismapping of reads containing motifs to high-complexity regions. The full analysis results of motif enriched alignments and haplotypes can be found in S3 Appendix.

Discussion

We have presented a first peak caller for ChIP-seq data mapped to a graph-based reference genome. We have tested our method by calling peaks on a graph-based reference genome for *A. thaliana* and comparing the detected peaks to those found by MACS2 using a linear reference genome of the same species. In the instances where the peak callers find peaks that overlap, Graph Peak Caller is able to find peaks that are more enriched for motifs. This is likely because Graph Peak Caller can trace variants within the peaks that are not part of the linear reference genome. Even if such peaks are also found by MACS2, the specific sequence comprising the DNA-binding motif may not be part of the linear reference.

Furthermore, Graph Peak Caller has a significantly higher proportion of motif-enriched peaks among its uniquely found peaks, showing that Graph Peak Caller is finding peaks enriched for motifs that MACS2 is not finding. These peaks cover more variations from the linear reference genome compared to the peaks found by both peak callers, and thus seem to be in areas where the advantages of graph-based reference genomes are more pronounced.

Also, when running MACS2 with vg alignments projected to the linear reference genome, Graph Peak Caller still has a slightly higher ratio of motif matches. This is probably because even though MACS2 in this case finds many of the same peaks as Graph Peak Caller, these peaks are represented on the linear reference genome and will in some cases miss motif match occurrences that cover variants.

We checked whether the higher motif enrichment among the peaks detected by Graph Peak Caller could be a result of a bias in alignments more often containing motif or alignments clustered to motif enriched areas of the graph, but we found no evidence for such biases.

Interestingly, Graph Peak Caller performs worse than MACS2 on the transcription factor ERF115. The reason for this is unclear. The ERF115 experiment is different from the rest of the experiments in a number of ways. The motif for ERF115 is GC-rich, in contrast to the motifs for all the other transcription factors as well as the *A. haliana* genome in general, which are all AT-rich. Also, the ERF155 reads come from a *TChAP* experiment, while the reads from the other transcription factors come from a normal ChIP-seq experiment. The ERF115 motif is not present in the JASPAR database and was gathered from a different source than the other transcription factors. However it is unclear how these factors might influence the results for ERF115, and since we have not seen similar behaviour in any other experiment it is difficult to deduce which of the factors, if any, are responsible for the lower motif enrichment.

We chose to develop Graph Peak Caller by tightly following the principles of MACS2, so we easily could validate our graph-based approach and accurately measure the benefits of doing peak calling on a graph-based reference genome rather than on a linear referencegenome. Having this as a working first approach to graph-based peak calling, it will now be natural and interesting to extend our work by drawing ideas from other peak callers or develop new peak calling principles to further improve graph-based peak calling. Also, Graph Peak Caller currently assumes no known information about the specific paths of the diploid genome of the individual that ChIP-seq data has been collected from. It would be interesting to develop a ChIP-seq pipeline where the path(s) through the reference graph are known (or estimated based on the ChIP-seq data), and compare that approach to Graph Peak Caller.

There are a few challenges with a graph-based ChIP-seq approach. Mapping to graphs is still in its infancy, and has not yet reached its full potential. Also, both mapping to graph-based reference genomes as well as many of the operations required for doing peak calling, such as expanding input reads, are, in our experience, still a lot slower than existing solutions on linear reference genomes.

We believe that our peak caller represents an important step towards creating a more comprehensive toolset for functional genomics on graph-based reference genomes, extending the possible applications of graph-based reference genomes and bringing the genomics community an important step closer to widespread adoption of these reference structures.

Conclusion

We have developed Graph Peak Caller, a tool for performing peak calling from ChIP-seq reads mapped to a graph-based reference genome. Graph Peak Caller is based on the same principles as MACS2. We have validated our approach by using both Graph Peak Caller and MACS2 to call peaks using ChIP-seq datasets on *A. thaliana*, showing that the peaks found by Graph Peak Caller in general are more enriched for DNA-binding motifs than those found by MACS2 on a linear reference genome. Graph Peak Caller is also able to provide candidates for differentially expressed peaks, and together with *vg* it provides a first method for doing peak calling on graph-based reference genomes.

Methods

Peak calling

Our approach to graph-based peak calling is implemented in an open source Python 3 package, *Graph Peak Caller*. Graph Peak Caller was developed by extending the methodologies and concepts from MACS2 to directed acyclic graphs (DAGs). The MACS2 algorithm can be divided into five steps: estimating the fragment length, creating a fragment pileup by extending input reads to match the estimated fragment length, calculating a background track based on local and global average number of reads, calculation of p/q scores based on the fragment pileup and background track, and finding peaks based on thresholded scores. We have adopted each of these steps to work on DAGs. Fig 1 illustrates the method on a graph-based reference genome, and the following describes the details of each step.

Graph Peak Caller uses the linear estimation algorithm from MACS2 to estimate the fragment length f by using the linear path through the graph with the highest number of aligned reads as reference. Graph Peak Caller generates *the fragment pileup* by extending each read to the estimated fragment length f, and counting the number of extended reads that cover each base pair in the graph. For a single read with length r, the extension is done by including all possible paths of length $f - r$ in the graph that start at the read's end position, using a breadth first search. The background track is an estimate of the expected number of reads mapping to each position in the reference. This is, for a given position in the reference, estimated by measuring the amount of reads mapping in the "neighbourhood" of that position. The reads can either be the input reads or a set of control reads. On a linear reference genome, the background track is simply estimated by taking the average pileup count in a local window around each base pair. This is less trivial to do on a graph-based reference genome, since the concept of neighborhood is not as well defined. We solve this problem by projecting the graph onto a single linear path where parallel paths are projected to the same position on the linear path. This allows us to perform background track estimation much the same way as MACS2 does, using a linear reference, and then projecting the resulting track back to the graph again. If control reads are used to generate the background track, the background track is scaled with the ratio of control reads to input reads.

The fragment pileup and background track are then treated as counts and rates in Poisson distributions, and p-values are computed for each position for the observed *count*, given the corresponding *rate*. Since one test is performed for each position in the graph, we compute q-values (adjusted p-values) to control the false discovery rate. The q-values are thresholded at a user specified threshold, yielding a binary track of potential binding regions.

Graph Peak Caller then removes small gaps (similarly to MACS2) between these potential binding regions. On a graph, this is done by joining regions that are connected by a path shorter than the read length. If a gap consists of several paths, all paths of length shorter than the read length are included in the joined region. Then, the resulting regions are grouped into connected subgraphs, representing areas of potential binding events. The final peaks are selected by finding the path through each subgraph that has the highest number of input reads mapped to it. Similarly to MACS2, peaks that are shorter than the estimated fragment length are removed.

For each subgraph, Graph Peak Caller can also report an "alternative" peak in addition to the main peak. This is done by using Fimo [12] to estimate the exact location within the peak subgraph that matches the binding motif, and looking for an alternative path through this area which is covered by at least one input read. Such alternative peaks can be used to infer differential binding.

Validation and testing

To test our peak caller, we used *vg* [6] to create a whole genome *Arabidopsis thaliana* reference graph by using variants from the 1001 Genomes Project [13]. We selected all transcription factors listed in the transcription factor database of *Expresso* [14] that also had a motif in the *Jaspar* database of transcription factor binding profiles [15], resulting in a set of 5 transcription factors: ERF115, SEP3, AP1, SOC1, and PI. (Two transcription factors, SVP and ATAF1, were omitted due to invalid fastq files. AP2 and AP3 were omitted based on their close relatedness to AP1. Also, PIF3 was omitted since neither the detected binding events by Graph Peak Caller nor MACS2 had any association with the motif we found in the Jaspar database). Raw ChiP-seq reads were downloaded from the NCBI Sequence Read Archive (SRA) (SRA accession numbers in S1 Appendix) and trimmed using *Trim Galore!* v0.4.4 [16] (default parameters). Reads were mapped both to our graph-based reference genome using *vg* and to the *Tair10* [10] reference genome using *BWA* v0.7.12 (`bwa aln` followed by `bwa samsse`, default parameters). In the linear case we filtered out low-quality alignments using *SAMtools* v0.1.19 [17] with the command `samtools view -F 1804 -q 37`, and for the graph alignments we used `vg filter -q 37`. *MACS2* v2.1.0 was used to call peaks on the linear reference genome, using default parameters. We created DNA-binding motif enrichment plots (Fig 5) for each set of detected peaks (URLs to the motif models that were used are in S1 Appendix). We have created a Docker repository with the *A. thaliana* graph-based reference genome, Graph Peak Caller, *vg* and all other software and scripts used to generate the results in this article. A simple guide on how to re-run the experiments can be found in the wiki in the Github repository for Graph Peak Caller.

To investigate how MACS2 performed when using vg alignments projected to the linear reference genome, we projected each vg alignment by finding a corresponding start and end position for the alignment on the linear reference genome. Given a start or end position on the graph, the new corresponding position on the linear reference genome was found by finding the shortest distance in the graph going from the original position backwards through the graph to a node shared with the linear reference genome, and then going this distance forward through the graph by following only nodes shared by the linear reference genome.

The analysis of motif enrichment alignments was performed by extracting the sequences from the alignments to the linear and graph-based reference genomes. We extracted the sequences from the linear alignments by first using Bedtools version 2.26.0 to convert the alignments from BAM to BED format by running *bamtools bamtobed* (default parameters), and then running *bedtools getfasta* (default parameters) on the BED files. We extracted sequences from the graph alignments by running *graph_peak_caller vg_json_alignments_to_fasta* using graphs created by *graph_peak_caller create_ob_graph*. We then ran Fimo with default parameters using these sequences as input.

The haplotype analysis was performed by finding all alignments overlapping a peak, and finding which variants each alignment contained, then finding the haplotypes from the VCF file that contained each variant. An alignment was decided to be compatible with a haplotype if all the variants included in the alignment were present in the haplotype. A set of alignments was decided to be compatible with two haplotypes if we could find two haplotypes in the vcf such that all the alignments in the set were compatible with at least one of them.

Supporting information

S1 Table. Overview of peaks detected on human and *D. melanogaster*.
(PDF)

S2 Table. Average number of base pairs not part of linear reference genome.
(PDF)

S3 Table. Overview of peaks found on *A. thaliana* when MACS2 is using vg alignments projected to the linear reference genome.
(PDF)

S1 Fig. DNA-binding motif enrichment plots (as Fig 5) for all peaks detected on *A. thaliana*. Contrary to Fig 5, these plots include all peaks found by both peak callers.
(TIF)

S2 Fig. DNA-binding motif enrichment plots (as Fig 5) for all peaks detected on *D. melanogaster* and human. Left plots are proportion of peaks matching motif when all peaks are included. Right plots are proportion of peaks matching motif when only unique peaks found by each peak caller are included.
(TIF)

S3 Fig. Diagram showing peaks found on *D. melanogaster* and human.
(TIF)

S1 Appendix. URLs to motifs and accession numbers to data used in experiments.
(PDF)

S2 Appendix. Reads and alignments behind Fig 3. List of fastq IDs and resulting alignment sequences intersecting with the binding site shown in Fig 3, showing that all of these appear in the same haplotype.
(PDF)

S3 Appendix. Details of haplotype analysis.
(PDF)

Acknowledgments

The sequence data used for creating the *Arabidopsis thaliana* reference graph were produced by the Weigel laboratory at the Max Planck Institute for Developmental Biology.

Author Contributions

Conceptualization: Ivar Grytten, Knut D. Rand, Alexander J. Nederbragt, Geir O. Storvik, Ingrid K. Glad, Geir K. Sandve.

Formal analysis: Geir O. Storvik, Ingrid K. Glad.

Investigation: Ivar Grytten, Knut D. Rand.

Methodology: Ivar Grytten, Knut D. Rand, Alexander J. Nederbragt, Geir K. Sandve.

Project administration: Ivar Grytten, Knut D. Rand.

Software: Ivar Grytten, Knut D. Rand.

Supervision: Alexander J. Nederbragt, Geir O. Storvik, Ingrid K. Glad, Geir K. Sandve.

Validation: Ivar Grytten, Knut D. Rand.

Visualization: Ivar Grytten, Knut D. Rand.

Writing – original draft: Ivar Grytten, Knut D. Rand.

Writing – review & editing: Ivar Grytten, Knut D. Rand, Alexander J. Nederbragt, Geir O. Storvik, Ingrid K. Glad, Geir K. Sandve.

References

1. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. Model-based Analysis of ChIP-Seq (MACS). Genome Biology. 2008 Nov; 9(9):R137. https://doi.org/10.1186/gb-2008-9-9-r137 PMID: 18798982
2. Kharchenko PV, Tolstorukov MY, Park PJ. Design and Analysis of ChIP-seq Experiments for DNA-binding Proteins. Nature Biotechnology. 2008 Dec; 26(12):1351. https://doi.org/10.1038/nbt.1508 PMID: 19029915

3. Wilbanks EG, Facciotti MT. Evaluation of Algorithm Performance in ChIP-seq Peak Detection. PloS One. 2010 Jul 8; 5(7):e11471. https://doi.org/10.1371/journal.pone.0011471 PMID: 20628599
4. Paten B, Novak AM, Eizenga JM, Garrison E. Genome Graphs and the Evolution of Genome Inference. Genome Research. 2017 May 1; 27(5):665–76. https://doi.org/10.1101/gr.214155.116 PMID: 28360232
5. Novak AM, Hickey G, Garrison E, Blum S, Connelly A, Dilthey A, Eizenga J, Elmohamed MS, Guthrie S, Kahles A, Keenan S. Genome Graphs. bioRxiv. 2017 Jan 1:101378.
6. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, Jones W, Lin MF, Paten B, Durbin R. Sequence Variation aware Genome References and Read Mapping with the Variation Graph Toolkit. bioRxiv. 2017 Jan 1:234856.
7. Li H. Aligning Sequence Reads, clone Sequences and Assembly Contigs with BWA-MEM. arXiv preprint arXiv:1303.3997. 2013 Mar 16.
8. Langmead B, Salzberg SL. Fast Gapped-read Alignment with Bowtie 2. Nature Methods. 2012 Apr; 9(4):357. https://doi.org/10.1038/nmeth.1923 PMID: 22388286
9. Rand KD, Grytten I, Nederbragt AJ, Storvik GO, Glad IK, Sandve GK. Coordinates and Intervals in Graph-based Reference Genomes. BMC Bioinformatics. 2017 Dec; 18(1):263. https://doi.org/10.1186/s12859-017-1678-9 PMID: 28521770
10. Arabidopsis Genome Initiative. Analysis of the Genome Sequence of the flowering Plant Arabidopsis thaliana. Nature. 2000 Dec; 408(6814):796. https://doi.org/10.1038/35048692 PMID: 11130711
11. Devore JL, Berk KN. Modern Mathematical Statistics with Applications. 2nd ed. New York: Springer-Verlag; 2012. p. 519–521
12. Grant CE, Bailey TL, Noble WS. FIMO: Scanning for Occurrences of a given Motif. Bioinformatics. 2011 Feb 16; 27(7):1017–8. https://doi.org/10.1093/bioinformatics/btr064 PMID: 21330290
13. 1000 Genomes Project Consortium. A global Reference for Human genetic Variation. Nature. 2015 Oct; 526(7571):68. https://doi.org/10.1038/nature15393 PMID: 26432245
14. Aghamirzaie D, Velmurugan KR, Wu S, Altarawy D, Heath LS, Grene R. Expresso: A Database and Web Server for exploring the Interaction of Transcription Factors and their Target Genes in Arabidopsis thaliana using ChIP-Seq peak Data. F1000Research. 2017; 6. https://doi.org/10.12688/f1000research.10041.1 PMID: 28529706
15. Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic acids research. 2004 Jan 1; 32:D91–4. https://doi.org/10.1093/nar/gkh012 PMID: 14681366
16. Krueger F. Trim galore. A Wrapper Tool around Cutadapt and FastQC to consistently apply Quality and Adapter trimming to FastQ files. 2015.
17. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map Format and SAMtools. Bioinformatics. 2009 Aug 15; 25(16):2078–9. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943