Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Applications Note

OXFORD

Genome analysis

geneXtendeR: optimized functional annotation of ChIP-seq data

Bohdan B. Khomtchouk [1,2,*], Derek J. Van Booven [3] and Claes Wahlestedt [2]

[1]Department of Biology, Stanford University, Stanford, CA 94305, USA
[2]Center for Therapeutic Innovation and Department of Psychiatry and Behavioral Sciences, University of Miami Miller School of Medicine, Miami, FL 33136, USA
[3]John P. Hussman Institute for Human Genomics, University of Miami Miller School of Medicine, Miami, FL 33136, USA.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Different ChIP-seq peak callers often produce different output results from the same input. Since different peak callers are known to produce differentially enriched peaks with a large variance in peak length distribution and total peak count, accurately annotating peak lists with their nearest genes can be an arduous process. Functional genomic annotation of histone modification ChIP-seq data can be a particularly challenging task, as chromatin marks that have inherently broad peaks with a diffuse range of signal enrichment (e.g., H3K9me1, H3K27me3) differ significantly from narrow peaks that exhibit a compact and localized enrichment pattern (e.g., H3K4me3, H3K9ac). In addition, varying degrees of tissue-dependent broadness of an epigenetic mark can make it difficult to accurately and reliably link sequencing data to biological function. Thus, it would be useful to develop a software program that can precisely tailor the computational analysis of a ChIP-seq dataset to the specific peak coordinates of the data.

Results: *geneXtendeR* is an R/Bioconductor package that optimizes the functional annotation of ChIP-seq peaks using fast iterative peak-coordinate/GTF alignment algorithms focused on cis-regulatory regions and proximal-promoter regions of nearest genes. The goal of *geneXtendeR* is to robustly link differentially enriched peaks with their respective genes, thereby aiding experimental follow-up and validation in designing primers for a set of prospective gene candidates during qPCR. We have tested *geneXtendeR* on 547 human transcription factor ChIP-seq ENCODE datasets and 214 human histone modification ChIP-seq ENCODE datasets, providing the analysis results as case studies.

Availability: The *geneXtendeR* R/Bioconductor package (including detailed introductory vignettes) is available under the GPL-3 Open Source license and is freely available to download from Bioconductor at: `https://bioconductor.org/packages/devel/geneXtendeR/`.

Contact: bohdan@stanford.edu

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

1 Introduction

Epigenetic histone chromatin marks come in a variety of different shapes and sizes, ranging from the extremely broad to the extremely narrow (Squazzo et al. 2006, Pepke et al. 2009, Landt et al. 2012, Kellis et al. 2014, Heinig et al. 2015). This spectrum depends on a number of biological factors ranging from qualitative characteristics such as tissue-type (Rintisch et al. 2014) to temporal aspects such as developmental stage (Ha et al. 2011). Computational factors such as the variance observed in peak coordinate positions (peak start, peak end) depending on the peak caller used, both in terms of length distribution of peaks as well as the total number of peaks called, is an issue that persists even when samples are run at identical default parameter values (Koohy et al. 2014; Thomas et al. 2017). This variance becomes a factor when annotating peak lists

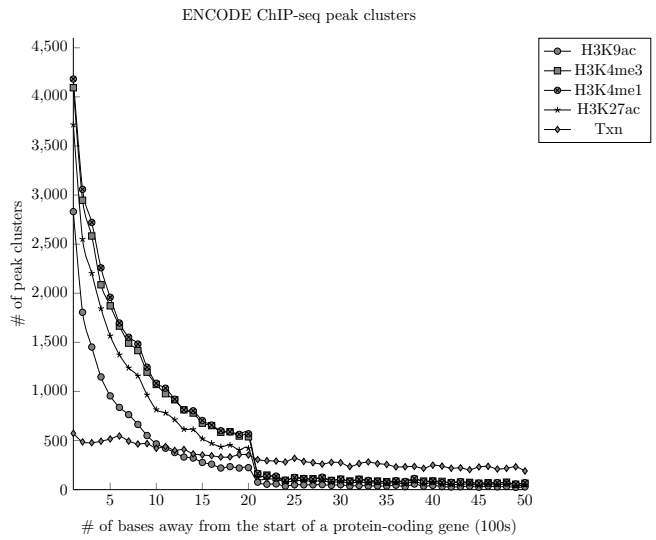

Fig. 1. Large-scale computational geneXtendeR analysis of 45 H3K27ac datasets, 48 H3K4me1 datasets, 94 H3K4me3 datasets, 27 H3K9ac datasets, and 547 transcription factor ChIP-seq datasets from ENCODE. The y-axis represents a raw count of peak clusters, where a peak cluster is defined as a genomic locus harboring at least 5 overlapping peaks. The x-axis represents a genomic distance (in bp) of the closest protein-coding gene to each respective peak cluster. A sharp drop in peak cluster count is detected at around 2000 bp for all chromatin marks, i.e., most peak clusters congregate proximally within 0-2000 bp of their respective protein-coding genes, yet a number of peak clusters reside further upstream of their nearest gene. In contrast to histone modification ChIP-seq studies, transcription factor ChIP-seq datasets do not exhibit a sharp drop in peak clusters at any particular genomic distance, although a steady decline is certainly evident.

genome-wide with their nearest genes as, depending on the peak caller employed, peaks can be either shifted in genomic position (towards 5' or 3' end) or be of different lengths. In total, the combined effect of all these factors exerts a unique influence over the functional annotation and understanding of genomic variability found in the broadness of chromatin mark peak data, which complicates the study of epigenetic regulation of biological function. Despite the existence of Bioconductor software and command line tools available for peak annotation to nearby features (e.g., ChIPpeakAnno (Zhu et al. 2010), HOMER (Heinz et al. 2010), BEDTools (Quinlan and Hall, 2010)), the aforementioned issues have not yet been addressed. As such, most studies still arbitrarily assign gene body definitions when mapping peaks to genomic features (Maze et al. 2011). To this end, we propose geneXtendeR, an R/Bioconductor package designed to assess the variability of peak overlap with cis-regulatory elements and proximal-promoter regions of nearest genes. geneXtendeR therefore represents a first step towards tailoring the functional annotation of a ChIP-seq peak dataset according to the details of the peak coordinates (chromosome number, peak start position, peak end position).

2 ENCODE analysis

We tested geneXtendeR on all publicly available transcription factor and histone modification ChIP-seq datasets in ENCODE (see Supplementary Information). Our large-scale analysis (Fig. 1) revealed that ChIP-seq peaks concentrate within the first 2000 bp upstream of their nearest protein-coding genes, with an immediate sharp drop observed in peak count for narrow chromatin marks (H3K9ac, H3K4me3, H3K4me1, and H3K27ac) but not transcription factors. Although 2000 bp is a good general guideline cutoff for capturing proximal histone modifications, this is not the case for transcription factor studies. In addition, there are still hundreds of peak clusters that reside in proximal-promoter regions that are 2000-3000 bp away from their nearest protein-coding genes and in distal regions beyond 3 kbp. When applying geneXtendeR to both proximal and distal transcription factor (TF) binding peaks for all cell types, we observed some cell type-dependent and TF-dependent peak aggregation dynamics in intervals ranging from 0 to 10 kbp (Fig. S1). Likewise, examining distal peaks in representative plots of different chromatin marks in different cell types indicates that peaks indeed aggregate in a cell type and chromatin mark-dependent manner (Fig. S2).

3 Functions

This complexity motivated us to design functions that can calculate ratios of statistically significant peaks to total peaks in various genomic intervals (see hotspotPlot() documentation in geneXtendeR vignette). Similarly, users can transform peaks into merged peaks (see peaksMerge()). geneXtendeR also allows users to explore gene ontology differences at various extensions (see diffGO()) as interactive network graphics (see makeNetwork()) or word clouds (see makeWordCloud()). Furthermore, users can investigate mean (average) peak lengths within any genomic interval (see meanPeakLengthPlot()), showing how average peak broadness can change at different upstream extensions, or examine the variance of peak lengths within a specific genomic interval (see peakLengthBoxplot()). It is also possible to examine unique genes and their associated ChIP-seq peaks between any two upstream extension levels (see distinct()). For example, Fig. S3 displays all unique genes (and their respective gene ontologies) that are associated with peaks located between 2-3 kbp across the genome. geneXtendeR also allows users to examine the distribution of peak lengths across the entire peak set (see allPeakLengths()), a function that is useful for visualizing the length distribution of all peaks from a peak caller. These functions and more are all explored in detail within the package vignette. After a user has explored the peak coordinates data using these functions to determine the optimal alignment of peaks to a GTF file, the peaks file can be functionally annotated with the annotate() function. We have successfully applied geneXtendeR during the analysis of a histone modification ChIP-seq study investigating the neuroepigenetics of alcohol addiction (Barbier et al. 2016), where geneXtendeR was used to determine an optimal upstream extension cutoff for H3K9me1 enrichment (a commonly studied broad peak) in rat brain tissue based on line plots of both significant peaks and total peaks. This analysis helped us to identify, functionally annotate, and experimentally validate synaptotagmin 1 (Syt1) as a key mediator in alcohol addiction and dependence (Barbier et al. 2016). All in all, geneXtendeR's functions are designed to be used as an integral part of a broader biological workflow (Fig. S4).

4 Conclusion

Motivated to optimally annotate ChIP-seq peak data based on the cis-regulatory and proximal promoter regions of genes, we propose an R/Bioconductor package to be used as an integral part of modern ChIP-seq workflows. geneXtendeR optimally annotates a ChIP-seq peak input file with functionally important genomic features (e.g., genes associated with peaks) based on optimization calculations in cis-regulatory and proximal promoter regions. As such, the user can effectively customize a ChIP-seq analysis to the tissue-specific, peak caller-specific, and environment-specific details that inherently affect the broadness, location, and total number of peaks in their dataset.

5 Supporting Information

Comprehensive TF ChIP-seq ENCODE analysis:

***geneXtendeR* analysis on 547 human TF ChIP-seq ENCODE datasets.** Files available here: `https://github.com/Bohdan-Khomtchouk/ENCODE_TF_geneXtendeR_analysis`

Comprehensive histone modification ChIP-seq ENCODE analysis:

***geneXtendeR* analysis on 214 human histone modification ChIP-seq ENCODE datasets.** Files available here: `https://github.com/Bohdan-Khomtchouk/ENCODE_histone_geneXtendeR_analysis`

6 Author's contributions

BBK conceived the study, designed the algorithms, implemented the R code and C code, engineered the R/C interface, implemented the Bioconductor package, and wrote the paper. BBK and DV analyzed the data. CW supervised the study and participated in the management of the source code and its coordination. All authors read and approved the final manuscript.

7 Acknowledgements

BBK dedicates this work to the memory of his uncle, Taras Khomchuk. BBK wishes to acknowledge the financial support of the United States Department of Defense (DoD) through the National Defense Science and Engineering Graduate Fellowship (NDSEG) Program: this research was conducted with Government support under and awarded by DoD, Army Research Office (ARO), National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a. BBK, DV, and CW thank Martin Morgan, Hervé Pagès, and Mohammed K. Sayed for useful technical support during the R/Bioconductor peer-review process.

8 Funding

This work has been supported by the Army Research Office (ARO), National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a.

References

[1]Barbier E, Johnstone AL, Khomtchouk BB, Tapocik JD, Pitcairn C, Rehman F, Augier E, Borich A, Schank JR, Rienas CA, Van Booven DJ, Sun H, Nätt D, Wahlestedt C, Heilig M: *Dependence-induced increase of alcohol self-administration and compulsive drinking mediated by the histone methyltransferase PRDM2*. Molecular Psychiatry. 2016, Nature Publishing Group. doi: 10.1038/mp.2016.131.

[2]Ha M, Ng DW, Li WH, Chen ZJ. *Coordinated histone modifications are associated with gene expression variation within and between species.* Genome Research. 2011, 21 (4): 590–598.

[3]Heinig M, Colomé-Tatché M, Taudt A, Rintisch C, Schafer S, Pravenec M, Hubner N, Vingron M, Johannes F. *histoneHMM: Differential analysis of histone modifications with broad genomic footprints.* BMC Bioinformatics. 2015, 16:60.

[4]Heinz S, Benner C, Spann N, Bertolino E et al.: *Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities.* Mol Cell 2010, 38(4): 576–589.

[5]Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, Dunham I, Elnitski LL, Farnham PJ, Feingold EA, Gerstein M, Giddings MC, Gilbert DM, Gingeras TR, Green ED, Guigo R, Hubbard T, Kent J, Lieb JD, Myers RM, Pazin MJ, Ren B, Stamatoyannopoulos JA, Weng Z, White KP, Hardison RC: *Defining functional DNA elements in the human genome.* Proceedings of the National Academy of Sciences. 2014, 111 (17): 6131–6138.

[6]Koohy H, Down TA, Spivakov M, Hubbard T: *A Comparison of Peak Callers Used for DNase-Seq Data*. PLoS One. 2014, 9(8): e105136.

[7]Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, Chen Y, DeSalvo G, Epstein C, Fisher-Aylor KI, Euskirchen G, Gerstein M, Gertz J, Hartemink AJ, Hoffman MM, Iyer VR, Jung YL, Karmakar S, Kellis M, Kharchenko PV, Li Q, Liu T, Liu XS, Ma L, Milosavljevic A, Myers RM, Park PJ, Pazin MJ, Perry MD, Raha D, Reddy TE, Rozowsky J, Shoresh N, Sidow A, Slattery M, Stamatoyannopoulos JA, Tolstorukov MY, White KP, Xi S, Farnham PJ, Lieb JD, Wold BJ, Snyder M: *ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.* Genome Research. 2012, 22 (9): 1813–1831.

[8]Pepke S, Wold B, Mortazavi A: *Computation for ChIP-seq and RNA-seq studies.* Nature Methods. 2009, 6 (11 Suppl): S22–S32.

[9]Quinlan AR, Hall IM: *BEDTools: a flexible suite of utilities for comparing genomic features*. Bioinformatics. 2010, 26(6): 841–842.

[10]Rintisch C, Heinig M, Bauerfeind A, Schafer S, Mieth C, Patone G, Hummel O, Chen W, Cook S, Cuppen E, ColomÃ©-TatchÃ© M, Johannes F, Jansen RC, Neil H, Werner M, Pravenec M, Vingron M, Hubner N: *Natural variation of histone modification and its impact on gene expression in the rat genome.* Genome Research. 2014, 24 (6): 942–953.

[11]Squazzo SL, O'Geen H, Komashko VM, Krig SR, Jin VX, Jang S, Margueron R, Reinberg D, Green R, Farnham PJ: *Suz12 binds to silenced regions of the genome in a cell-type-specific manner.* 2006. Genome Research 16: 890–900.

[12]Thomas R, Thomas S, Holloway AK, Pollard KS: *Features that define the best ChIP-seq peak calling algorithms*. Briefings in Bioinformatics. 2017, 18(3): 441–450.

[13]Zhu L, Gazin C, Lawson N, Pages H, Lin S, Lapointe D, Green M: *ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data*. BMC Bioinformatics. 2010, 11(1), pp. 237.

Supplementary Figures

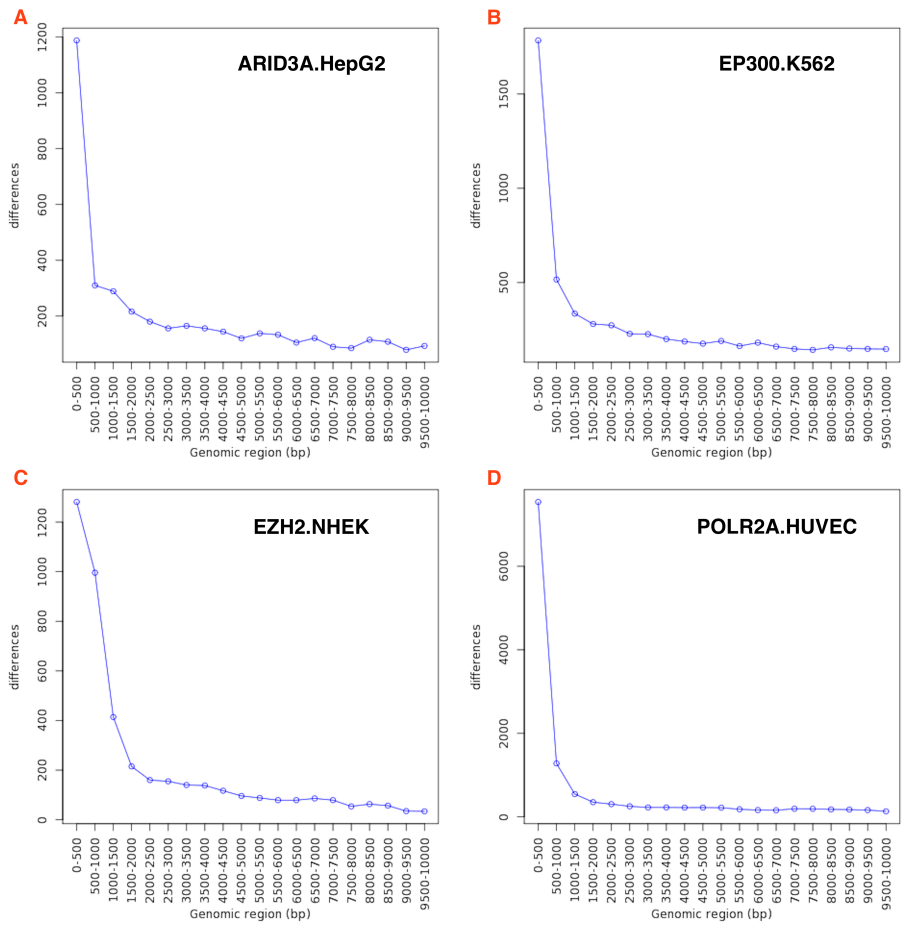

Fig. S1. Running geneXtendeR on 547 human transcription factor ChIP-seq datasets obtained from ENCODE shows that most peaks reside within 500 bp upstream of their respective protein-coding genes. Depending on the identity of the transcription factor (e.g., EP300) and the specific cell type (e.g., K562), there may be more or less peaks located further upstream. Therefore, choosing an optimal gene extension is a simple exercise in ψ calculation (see hotspotPlot() in vignette) for various upstream extension levels at a given user-specified statistical criterion (e.g., p-value and/or FDR cutoffs).

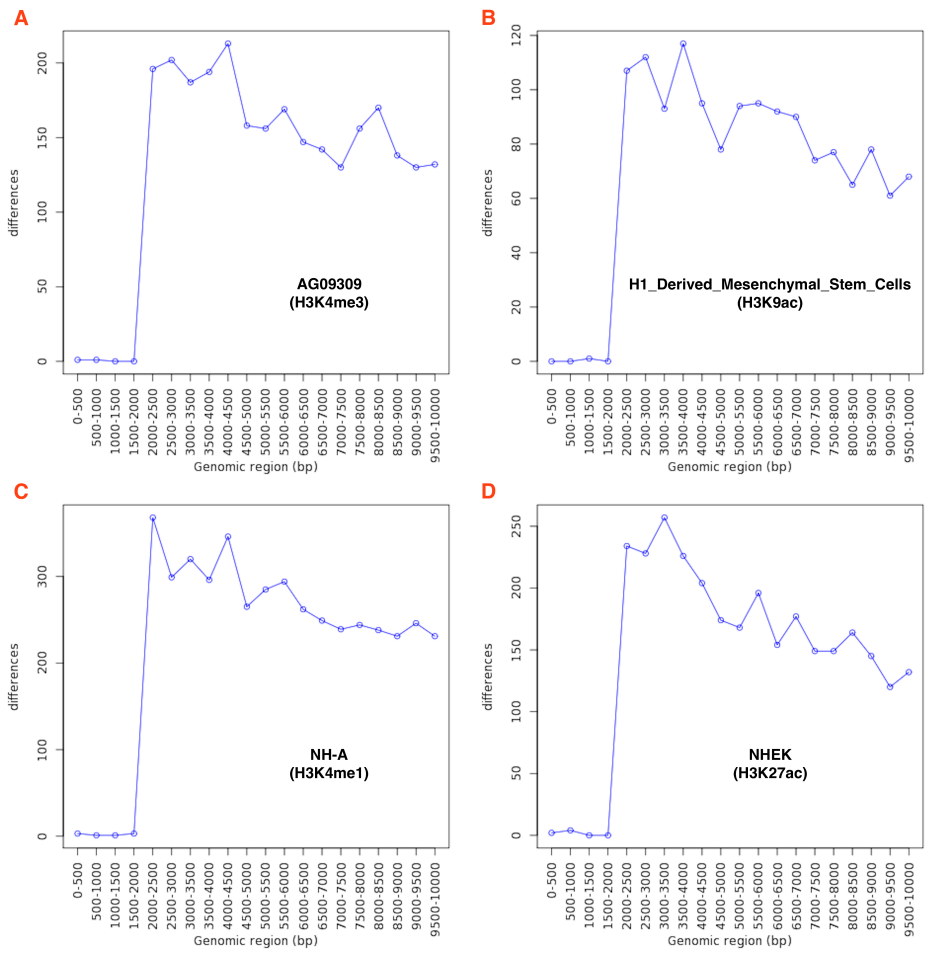

Fig. S2. Running geneXtendeR on 214 human histone modification ChIP-seq distal peak datasets obtained from ENCODE reveals that most distal peaks congregate within 5000 bp upstream of their respective protein-coding genes. Additional comprehensive analyses (see Supplementary Information) were also run for proximal peaks as well as the complete set of peaks (proximal + distal) from all 214 histone modification ChIP-seq datasets. Each representative figure panel demonstrates that a spike in the number of distal peaks in a particular upstream interval differs from one dataset to another. For example, AG09309 experiences a spike in the 4000-4500 bp region, whereas NHEK experiences a spike in the 3000-3500 bp region. This demonstrates the simple observation that arbitrarily extending genes by some generalized upstream cutoff is unlikely to capture the optimal number of genes-under-distal-peaks for any one specific dataset. For instance, a totally different set of dynamics is seen with NH-A, where the highest spike occurs immediately at 2000-2500 bp, but then another spike of almost identical magnitude occurs at 4000-4500 bp, suggesting that a 4500 bp upstream global extension of each gene might be preferable to a 2500 bp extension for capturing the optimal number of genes-under-distal-peaks (may be verified with hotspotPlot() function). On the contrary, datasets like H1_Derived_Mesenchymal_Stem_Cells experience a single spike at 3500-4000 bp, followed by a gradual decline.

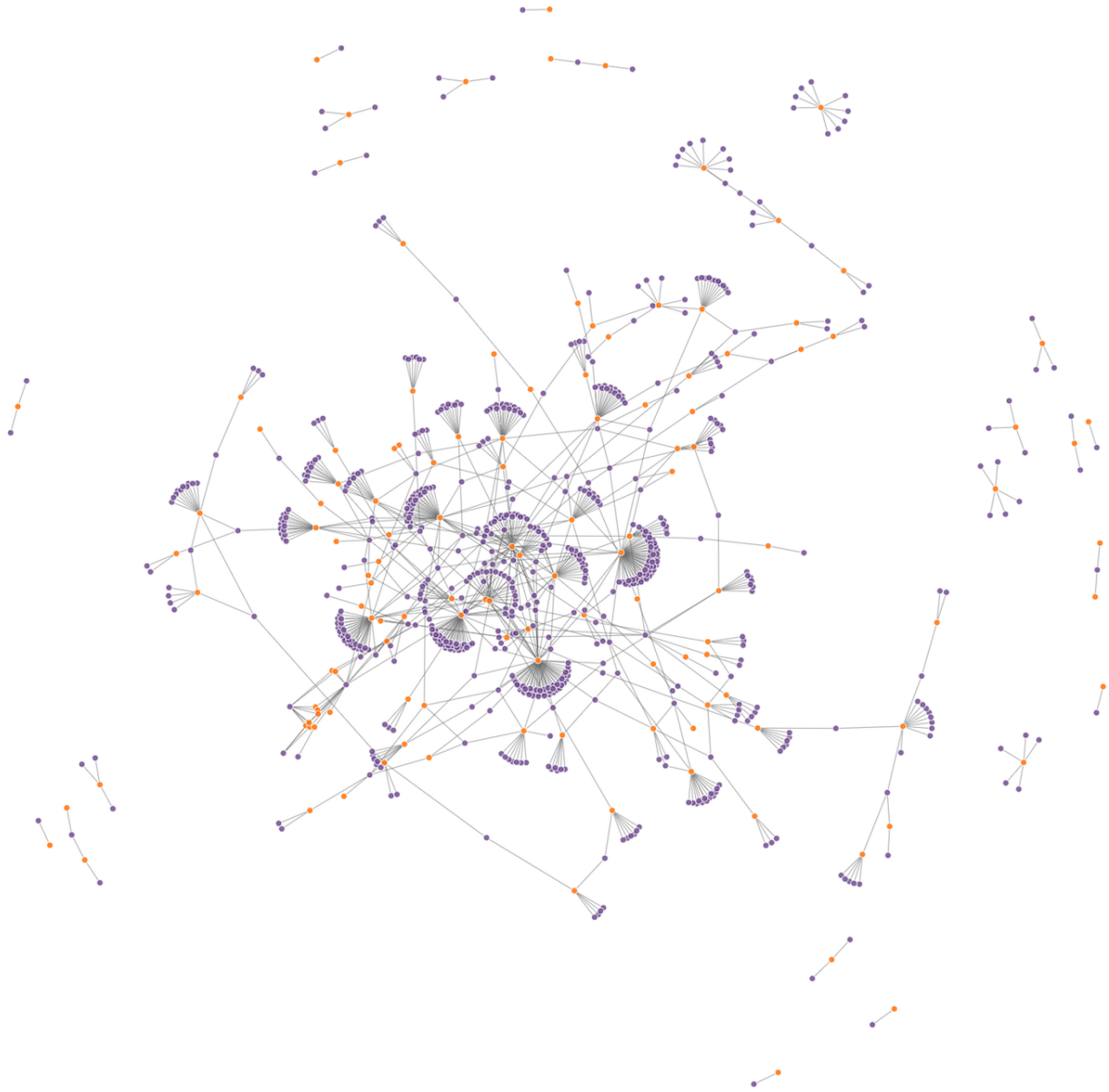

Fig. S3. All unique genes (and their respective gene ontologies) that are associated with peaks located between 2-3 kbp across the genome. Put another way, these are all gene-GO pairs associated with peaks that are distinct between 2000 and 3000 bp upstream extensions. Orange color denotes gene names, purple color denotes GO terms. A user can hover the mouse cursor over any given node to display its respective label directly within R Studio. Likewise, users can dynamically drag and reorganize the spatial orientation of nodes, as well as zoom in and out of them for visual effect.

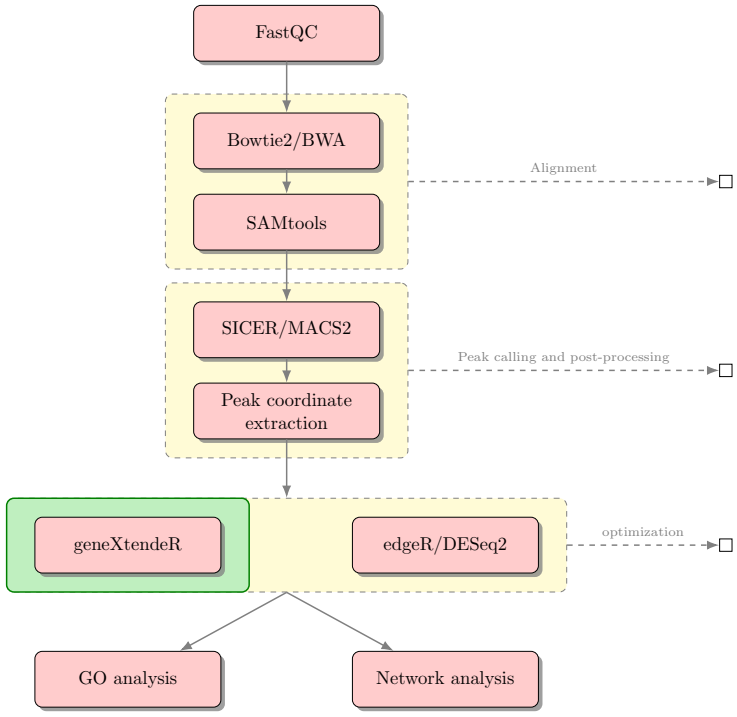

Fig. S4. Sample biological workflow using geneXtendeR in combination with existing statistical software to analyze peak significance. Note that geneXtendeR's built-in hotspotPlot() function takes as input both the total peaks and the statistically significant peaks as returned by the peak caller. However, a user may wish to also run an additional differential expression (DE) analysis (e.g., using edgeR/DESeq2) to count reads from the loci belonging to the peak coordinates as a way to manually cross-check the statistical results returned from the original peak caller. Likewise, geneXtendeR has built-in gene ontology and network analysis functions, but these may also be cross-checked from the results of a user's chosen DE caller. As such, subsequent gene ontology or network analysis may be conducted on genes associated with statistically significant peaks returned from such DE callers, and an overlap with geneXtendeR's results may be conducted to assess stringency. Significant peaks may be located thousands of base pairs away from their nearest genes, suggesting that sequences under these respective peaks may be further extracted and analyzed for the presence of known regulatory elements or repeats (e.g., using TRANSFAC, MEME/JASPAR, or RepeatMasker).