bioRxiv preprint doi: https://doi.org/10.1101/082347; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



# geneXtendeR: R/Bioconductor package for functional annotation of histone modification ChIP-seq data in a 3D genome world

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

**1** Center for Therapeutic Innovation and Department of Psychiatry and Behavioral Sciences, University of Miami Miller School of Medicine, Miami, FL, USA
**2** John P. Hussman Institute for Human Genomics, University of Miami Miller School of Medicine, Miami, FL, USA

* Corresponding author: b.khomtchouk@med.miami.edu

## Abstract

### 0.1 Motivation:

Functional genomic annotation of epigenetic histone modification ChIP-seq data is a computationally challenging task. Epigenetic histone modifications 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). Varying degrees of tissue-dependent broadness of the specific epigenetic mark coupled with environmentally mediated 3D-folding of chromosomes for long-range communication make it difficult to accurately and reliably link sequencing data to biological function. Hence, it would be useful to develop a software program that can precisely tailor the computational analysis of a histone modification ChIP-seq dataset to the specific tissue-dependent, environmentally mediated characteristics of the data.

### 0.2 Results:

*geneXtendeR* is an R/Bioconductor package designed to optimally annotate a histone modification ChIP-seq peak input file with functionally important genomic features (e.g., genes associated with peaks) based on optimization calculations. *geneXtendeR* optimally extends the boundaries of every gene in a genome by some genomic distance (in DNA base pairs) for the purpose of flexibly incorporating cis and trans-regulatory elements, such as enhancers and promoters, as well as downstream elements that are important to the function of the gene relative to an epigenetic histone modification ChIP-seq dataset. *geneXtendeR* computes optimal gene extensions tailored to the broadness of the specific epigenetic mark (e.g., H3K9me1, H3K27me3), as determined by a user-supplied ChIP-seq peak input file. As such, *geneXtendeR* maximizes the signal-to-noise ratio of locating genes closest to and directly under peaks that may be linked by epigenetic regulation. By performing a computational expansion of this nature, ChIP-seq reads that would initially not map strictly to a specific gene can now be optimally mapped to the regulatory regions of the gene, thereby implicating the gene as a potential candidate, and thereby making the ChIP-seq experiment more successful. Such an approach becomes particularly important when working with epigenetic histone modifications that have inherently broad peaks. We have tested *geneXtendeR* on 547

bioRxiv preprint doi: https://doi.org/10.1101/082347; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



human transcription factor ChIP-seq ENCODE datasets and 215 human histone modification ChIP-seq ENCODE datasets, providing the analysis results as case studies.

### 0.3 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/geneXtendeR/.

### 0.4 Contact:

b.khomtchouk@med.miami.edu

## Author Summary

*geneXtendeR* is an R/Bioconductor package for histone modification ChIP-seq analysis. It is designed to optimally annotate a histone modification ChIP-seq peak input file with functionally important genomic features (e.g., genes associated with peaks) based on optimization calculations. *geneXtendeR* optimally extends the boundaries of every gene in a genome by some genomic distance (in DNA base pairs) for the purpose of flexibly incorporating cis-regulatory elements, such as enhancers and promoters, as well as downstream elements that are important to the function of the gene (relative to an epigenetic histone modification ChIP-seq dataset). *geneXtendeR* computes optimal gene extensions tailored to the broadness of the specific epigenetic mark (e.g., H3K9me1, H3K27me3), as determined by a user-supplied ChIP-seq peak input file. As such, *geneXtendeR* maximizes the signal-to-noise ratio of locating genes closest to and directly under peaks. By performing a computational expansion of this nature, ChIP-seq reads that would initially not map strictly to a specific gene can now be optimally mapped to the regulatory regions of the gene, thereby implicating the gene as a potential candidate, and thereby making the ChIP-seq experiment more successful. Such an approach becomes particularly important when working with epigenetic histone modifications that have inherently broad peaks. Vignettes detailing input/output requirements, suggested biological workflows, and underlying package infrastructure are included as part of the R/Bioconductor package. We have tested *geneXtendeR* on 547 human transcription factor ChIP-seq ENCODE datasets and 215 human histone modification ChIP-seq ENCODE datasets, providing the analysis results as case studies.

## 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 factors ranging from qualitative characteristics such as tissue-type (Rintisch et al. 2014) to temporal aspects such as developmental stage (Ha et al. 2011). The issue is further complicated by the inherently one-dimensional nature of ChIP-seq data as a static snapshot of the cellular state in the context of a highly dynamic three-dimensional genomic environment (Furey 2012, Dekker & Mirny 2016). Since ChIP-seq peak caller data originates from a single spatiotemporal snapshot of the cross-linking status of histones at the moment of lab sample preparation, the overall genomic view is naturally constrained to a single time point requiring extensive comparative follow-up analyses to establish a sense of temporal continuity (Sandmann et al. 2006, Taslim et al. 2009, Bailey et al. 2013). In total, the combined effect of all

bioRxiv preprint doi: https://doi.org/10.1101/082347; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



these factors exerts a unique influence over the genomic variability in broadness of the specific chromatin marks, which, in turn, complicates the study of their epigenetic regulation of biological function.

Furthermore, it is known that predictive chromatin signatures, which are characterized by strong enrichment of a specific epigenetic mark, have been used to successfully map enhancers at a genome-wide level (Heintzman et al. 2009, Hon et al. 2009), suggesting a definitive interplay between epigenetic marks and enhancer elements. In general, enhancers are known to be marked with highly cell-type-specific histone modification patterns, which, in turn, influence cell-type-dependent functional activation of gene expression programs on a global scale (Heintzman et al. 2009). From a temporal standpoint, enhancer-promoter pairs that are separated by as much as 100 kb only require time-scales on the order of minutes to be able to successfully interact in three-dimensional space via chromatin loop-extrusion (Dekker & Mirny, 2016). Consequently, enhancer-promoter interactions separated by 20 kb or less would be expected to occur with even greater frequency and duration, a feat achieved by direct physical contact via fast and dynamic chromatin looping (Drissen et al. 2004, Vakoc et al. 2005, Jing et al. 2008, Sanyal et al. 2012). Therefore, considering that enhancer elements can reside close to the promoters they control, hundreds of kilobase pairs away, or within gene bodies (Lee et al. 2015), it should come as no surprise to encounter statistically significant peaks located thousands of base pairs away from any nearby genes based on this 3D genome model.

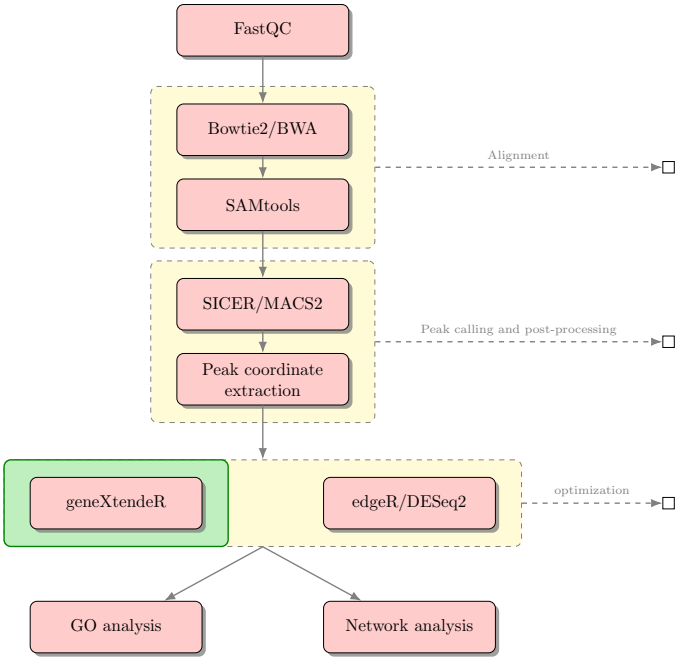


**Fig 1.** Sample biological workflow using *geneXtendeR* in combination with existing statistical software to analyze peak significance. Subsequent gene ontology or network analysis may be conducted on genes associated with statistically significant peaks. Significant peaks may be located thousands of base pairs away from their nearest genes, suggesting potential enhancer interactions in the genome. Sequences under these respective peaks may be further extracted and analyzed for the presence of known regulatory elements or repeats.

bioRxiv preprint doi: https://doi.org/10.1101/082347; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



# Materials and Methods

## Approach.

Since ChIP-seq data is an inherently one-dimensional experimental set of information, our aim is to devise a computational tool that may help inform our analyses in the context of a three-dimensional genomic environment. To this end, we propose an R/Bioconductor package designed to identify potential association of histone modification ChIP regions with functionally important genomic regions based on optimization calculations.

To achieve this, we model a gene as a composite element composed of three additive components: its gene body, *GB* (i.e., sequence between the TSS and TES), a variable upstream extension $\delta$ (in bp), and a 500 bp downstream extension $\gamma$ (in bp). To account for the three different modes of chromosomal communication (3D looping, 1D scanning, and 3D diffusion) (Dekker & Mirny, 2016), we denote this new construct the gene-sphere (*GS*).

For positive DNA strands:

$$GS = GB - \delta + \gamma \tag{1}$$

For negative DNA strands:

$$GS = GB + \delta - \gamma \tag{2}$$

*geneXtendeR* computes gene-spheres that maximize the differences in the number of genes-under-peaks between any two consecutive upstream extensions, thereby priming the peak caller data for an optimal analysis by allowing the user to choose the most optimal upstream gene extension tailored to the broadness of the specific epigenetic mark (see *geneXtendeR* package vignette for details and visualizations).

# Results & Discussion

*geneXtendeR* is designed to be used as part of a biological ChIP-seq workflow that includes read alignment, peak calling, and differential expression analysis (Fig. 1). Gene ontology and network analysis may also be integrated into the analysis pipeline.

The input to *geneXtendeR* is a file consisting of a list of peak coordinates data (chromosome, peak start position, peak end position) called from a ChIP-seq peak caller (e.g., SICER (Zang et al. 2009), MACS2 (Zhang et al. 2008), etc). The final output of *geneXtendeR* is an annotated peaks file consisting of gene IDs, gene names, and proximity information (i.e., distance of peak to nearest gene) calculated via the line plot, bar chart, and differentially unique peak functions of the *geneXtendeR* package (see Bioconductor documentation for details). Peaks may further be examined through a differential expression analysis caller (e.g, edgeR (Robinson et al. 2010), DESeq2 (Love et al. 2014), etc) to check for the presence of statistically significant peaks in gene deserts (i.e., distance of peak to nearest gene >> 10 kb). Sequences under these peaks may be checked for the presence of known regulatory motifs (e.g., using TRANSFAC (Matys et al. 2006) or MEME/JASPAR (Sandelin et al. 2004, Bailey et al. 2009)), or for the presence of biological repeats (e.g., using RepeatMasker (Smit et al. 2015)). Pending a prospective GO/network analysis, functional validation may be followed up in the lab to test any potential regulatory sites or prospective enhancer elements, thereby bringing the computational analysis pipeline successfully back to the bench.

bioRxiv preprint doi: https://doi.org/10.1101/082347; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



## Case Study 1: Computational Epigenomics of Brain Tissue

*geneXtendeR* has successfully been employed in a histone modification ChIP-seq study investigating the neuroepigenetics of alcohol addiction (Barbier et al. 2016). In this study, *geneXtendeR* was used to determine an optimal upstream extension cutoff for H3K9me1 enrichment based on line plots of both significant peaks and total peaks (Fig. 2).

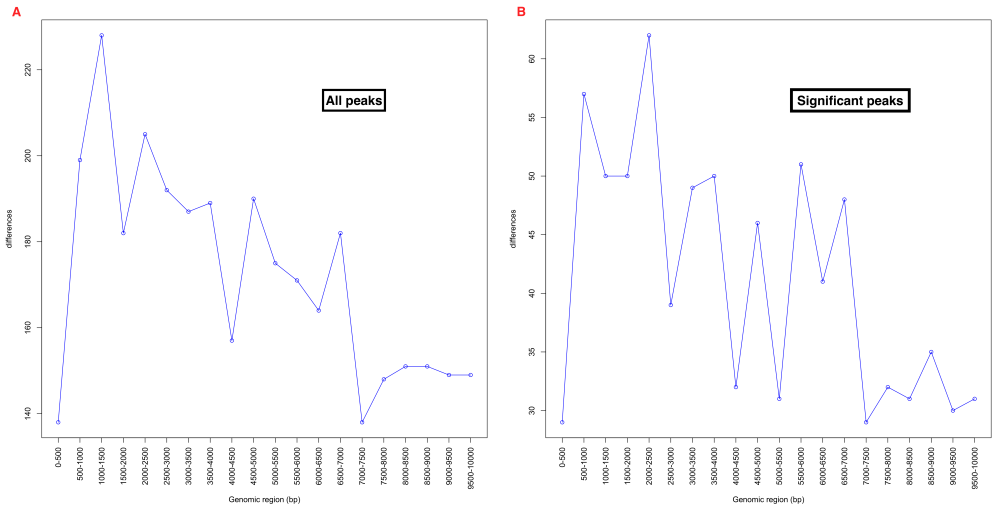


**Fig 2.** Peaks from a published study (Barbier et al. 2016) are run through *geneXtendeR* with parameters starting at 0 bp from the gene, extending out to 10000 bp away from the genes, with 500 bp increments. The number of differences in the count of genes-under-peaks is maximized at 1500 bp (panel A), but is maximized at 2500 bp in the genes-under-significant-peaks (panel B). As such, a 3000 bp extension was ultimately chosen in order to provide an extra 500 bp window that holds an additional 39 significant gene-spheres. Likewise, a 4000 bp extension could have been potentially chosen instead, due to the steady rise in number of genes-under-significant-peaks up until 4000 bp, after which a sharp drop in the number of genes-under-significant-peaks occurs in the 4000-4500 bp region. However, a 3000 bp extension was ultimately chosen in light of a prior study that utilized a 2000 bp upstream extension for H3K9me3 in the nucleus accumbens (Maze et al. 2011), a similar epigenetic mark in a similar brain region.

This decision was, in part, motivated by a previous study (Maze et al. 2011) that used a 2000 bp upstream extension of the transcription start site (TSS) such that proximal regulatory regions were also included. Since both studies investigated similar epigenetic marks (H3K9me1 vs. H3K9me3) in similar brain regions (prefrontal cortex vs. nucleus accumbens), a more conservative 3000 bp extension was ultimately chosen in favor of the alternative 4000 bp extension (see Fig. 2 for details).

Fig. 2 shows that most peaks (228 peaks) are located between 1000 and 1500 bp upstream of the first exon of their respective genes, but only 50 of these peaks are significant according to the statistical criterion chosen by the user (in this case, $p < 0.05$). However, the largest number of significant peaks (62 peaks) occur in regions located between 2000 and 2500 bp upstream of the first exon of their respective genes. To determine the identity of the unique peaks between any two user-specified upstream cutoffs (e.g., 2000 vs. 2500), the user can do $distinct(organism, 2000, 2500)$ (where *organism* is rat, see R package vignette for details), thereby returning a table listing of annotated peaks found between 2000 and 2500 bp upstream of their respective genes

bioRxiv preprint doi: https://doi.org/10.1101/082347; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



(Fig. 3).

```
> head(distinct(rat, 2000, 2500))
[1] "1\t20328600\t20329799\t1\t20329476\t20412881\tENSRNOG00000047783\tTmem200a\t0"
[2] "1\t78830000\t78830999\t1\t78830657\t78838123\tENSRNOG00000016756\tPtgir\t0"
[3] "1\t80224200\t80226199\t1\t80213994\t80224210\tENSRNOG00000046667\tFosb\t0"
[4] "1\t102896200\t102897799\t1\t102897786\t102910207\tENSRNOG00000013009\tLdha\t0"
[5] "1\t134693400\t134696999\t1\t134696799\t134743014\tENSRNOG00000012874\tRgma\t0"
[6] "1\t227004600\t227006999\t1\t226995737\t227004639\tENSRNOG00000020925\tCcdc86\t0"
```

**Fig 3.** A truncated table listing of the first six annotated peaks found between 2000 and 2500 bp upstream of their respective genes. `\t` denotes the tab-delimited separator, where the first column is the peak chromosome number, the second column is the peak start position, the third column is the peak end position, the fourth column is the gene chromosome number, the fifth column is the gene start position, the sixth column is the gene end position, the seventh column is the Ensembl gene ID, the eigth column is the gene symbol, and the ninth column is the gene-to-peak genomic distance (in bp), where distance is calculated between 5-prime end of the gene and 3-prime end of the peak. There are 62 peaks returned with this command, which represent the specific genes-under-significant-peaks that are found in the 2000-2500 bp interval. Investigating the gene ontologies and/or network analyses of these genes may prove useful in resolving consistent biological patterns in the peak data.

One interesting genome metric to consider is:

$$\psi = \frac{\text{Number of significant peaks at a specific interval}}{\text{Total number of peaks at this specific interval}} \quad (3)$$

where $\psi$ represents the relative abundance of significant peaks at each genomic interval normalized by the total number of peaks found at that interval. Since the user is responsible for uploading the 3-column peaks list (comprised of chromosome number, start position of peak, and end position of peak; see R package vignette for details), it is easy to upload either all the peaks generated by a peak caller or just the significant peaks generated by a peak caller (where significance is set according to some statistical criterion chosen by the user). Computing $\psi$ on Fig. 2 reveals that the locations of significant peaks are more or less randomly distributed across the genome (Fig. 4), with about 25% of the total peak pool returning significant at a statistical cutoff of $p < 0.05$.

bioRxiv preprint doi: https://doi.org/10.1101/082347; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



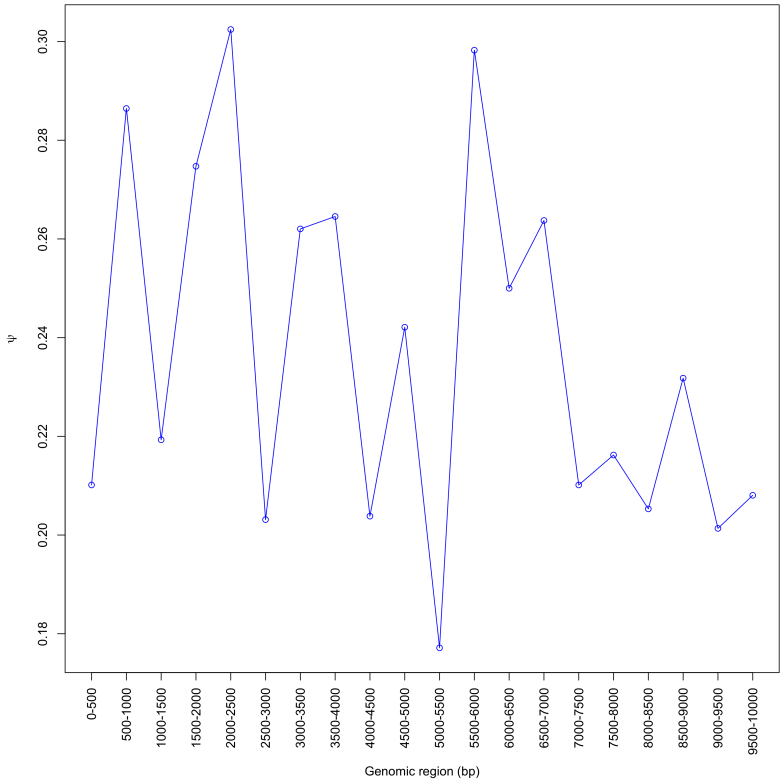


**Fig 4.** Levels of $\psi$ (Eq. 3) showing random fluctuations characteristic of stochastic noise demonstrate that, on average (across the whole genome), there is no specific distance from a gene that can be definitively associated with an overabundance of significant peaks (relative to the total number of peaks).

Hence, even though the 2000-2500 bp and 5500-6000 bp regions maximize levels of $\psi$, the random and noisy fluctuations of the line plot suggest that there is no clearly defined region (i.e., genomic interval upstream of a gene) harboring a cluster of significant peaks (relative to total peaks). However, we anticipate that datasets that may exhibit such a cluster could reveal hotspots of statistically significant chromatin mark activity in either proximal (e.g., promoter) or distal (e.g., enhancer) regulatory regions. For instance, transcription factor (TF) ChIP-seq datasets with an overabundance of peaks in proximal-promoter regions (Fig. 5) are expected to harbor more statistically significant peaks in these genomic regions relative to distal regions.

## Case Study 2: ENCODE TF ChIP-Seq Analysis

In this regard, not only is *geneXtendeR* useful for histone modification ChIP-seq analysis, it is also instructive when applied to transcription factor ChIP-seq data analysis. We applied *geneXtendeR* to 547 human ENCODE datasets for both proximal and distal transcription factor (TF) binding peaks for all cell types. Both distal and proximal lists were combined to form a single comprehensive peak list, which was then annotated with *geneXtendeR* using protein-coding genes from the human Gencode v19 GTF file. As such, all peaks were run through *geneXtendeR* with parameters starting at 0 bp from the gene, extending out to 10000 bp away from the genes, with 500 bp increments. As expected, the further the distance from the gene, the less peaks appear, due in part to the sharp and narrow nature of peaks residing in promoter-proximal

bioRxiv preprint doi: https://doi.org/10.1101/082347; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



elements that are characteristic of TF ChIP-seq datasets. Most of the results undergo a significant drop in the number of genes-under-peaks after the proximal region, and a steady dropoff is seen in the distal region (Fig. 5). This is expected, as the majority of TF peaks tend to reside in the promoter-proximal regions of their respective genes, a pattern that is evident across multiple TF identities and cell types (Comprehensive TF ChIP-seq ENCODE analysis). A full list of all ENCODE TF datasets analyzed by *geneXtendeR*, including bioinformatic shell scripts, is provided as Supplementary Material (Comprehensive TF ChIP-seq ENCODE analysis).

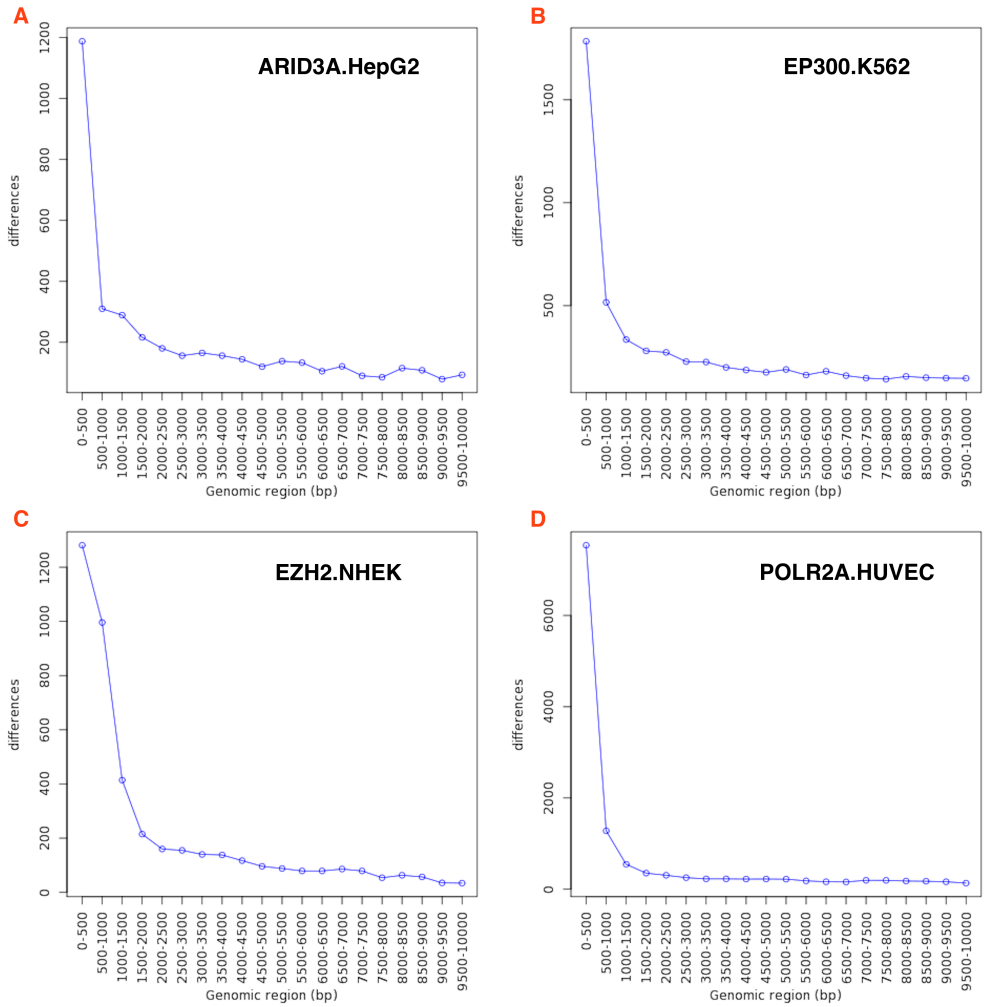


**Fig 5.** Running *geneXtendeR* on 547 human transcription factor ChIP-seq datasets obtained from ENCODE shows that most peaks reside within 500 bp upstream of their respective protein-coding genes. Depending on the identity of the transcription factor (e.g., EP300) and the specific cell type (e.g., K562), there may be more or less peaks located further upstream. Therefore, choosing an optimal gene extension is a simple exercise in $\psi$ calculation (Eq. 3) for various upstream extension levels at a given user-specified statistical criterion (e.g., p-value and/or FDR cutoffs), thereby giving the user complete freedom and control over what statistical apparatus and stringency to use during this procedure.

bioRxiv preprint doi: https://doi.org/10.1101/082347; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



## Case Study 3: ENCODE Histone Modification ChIP-Seq Analysis

We applied *geneXtendeR* to 215 human ENCODE datasets for both proximal and distal histone modification peaks for all cell types with the following chromatin marks: H3K27ac, H3K4me1, H3K4me3, and H3K9ac. Distal, proximal, and combined (distal + proximal) lists were annotated with *geneXtendeR* using protein-coding genes from the human Gencode v19 GTF file. Each set of peaks was run through *geneXtendeR* at two different sets of parameters (0-2000 bp at 100 bp increments, and 0-10000 at 500 bp increments). It was found that most distal peaks congregate within 5000 bp upstream of their respective protein-coding genes and that optimal upstream extensions are dependent on the specific chromatin mark and cell type (Fig. 6).

bioRxiv preprint doi: https://doi.org/10.1101/082347; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



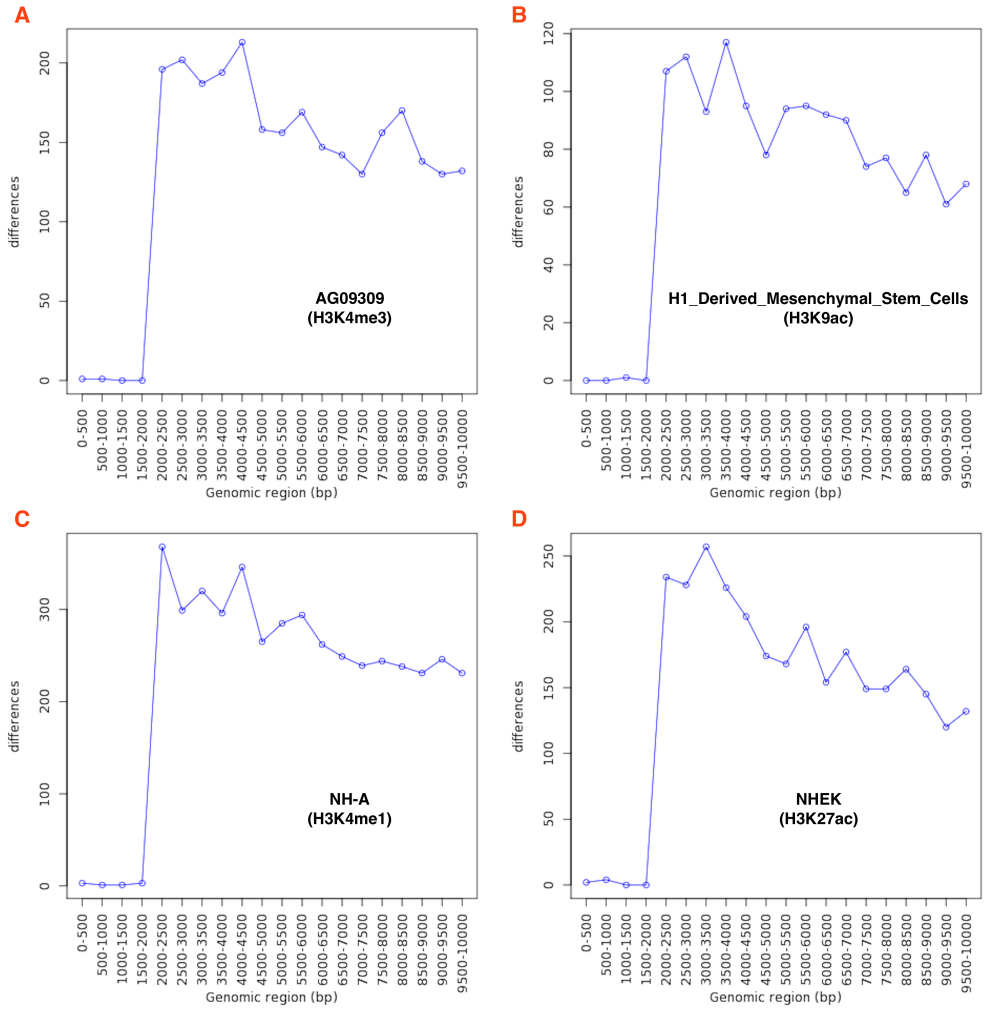


**Fig 6.** Running *geneXtendeR* on 215 human histone modification ChIP-seq distal peak datasets obtained from ENCODE reveals that most distal peaks congregate within 5000 bp upstream of their respective protein-coding genes. Additional comprehensive analyses (Comprehensive histone ChIP-seq ENCODE analysis) were also run for proximal peaks as well as the complete set of peaks (proximal + distal) from all 215 histone modification ChIP-seq datasets. Each representative figure panel demostrates 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 would be preferable to a 2500 bp extension for capturing the optimal number of genes-under-distal-peaks. On the contrary, datasets like H1_Derived_Mesenchymal_Stem_Cells experience a single spike at 3500-4000 bp, followed by a gradual decline.

bioRxiv preprint doi: https://doi.org/10.1101/082347; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



## Case Study 4: A Tale of Three Peak Callers

Since *geneXtendeR* asks the user to supply a peaks list, a user can run multiple peak callers and then supply *geneXtendeR* with a list of mutual peaks called in common by the various peak callers. Such an approach was employed in Barbier et al. 2016, where SICER (Zang et al. 2009), MACS2 (Zhang et al. 2008), and CisGenome (Hongkai et al. 2008) were run at default settings and the peak results were intersected to examine mutually inclusive peaks called in common by the three peak callers. This input was then passed to *geneXtendeR*, which annotated the 3-column peaks list (chromosome number, start position of peak, and end position of peak) with gene information via the $annotate(organism, extension)$ command (see package documentation for details). Specifically, it was of interest to reveal which genes lay directly under peaks, even at a 0 bp upstream extension. One such gene, Syt1, became a prime candidate for experimental validation because it exhibited biologically relevant gene ontology and network analysis follow-up results, which is a suggested component of the biological workflow (Fig. 1). A follow-up quantitative real-time PCR (qPCR) confirmed the presence of differential H3K9me1 enrichment at the Syt1 gene in control versus postdependent rats (Fig. 7), therefore successfully validating this gene candidate in the study, and thereby validating the bioinformatics pipeline used in the ensuing computational analysis, where SICER was utilized for peak calling.

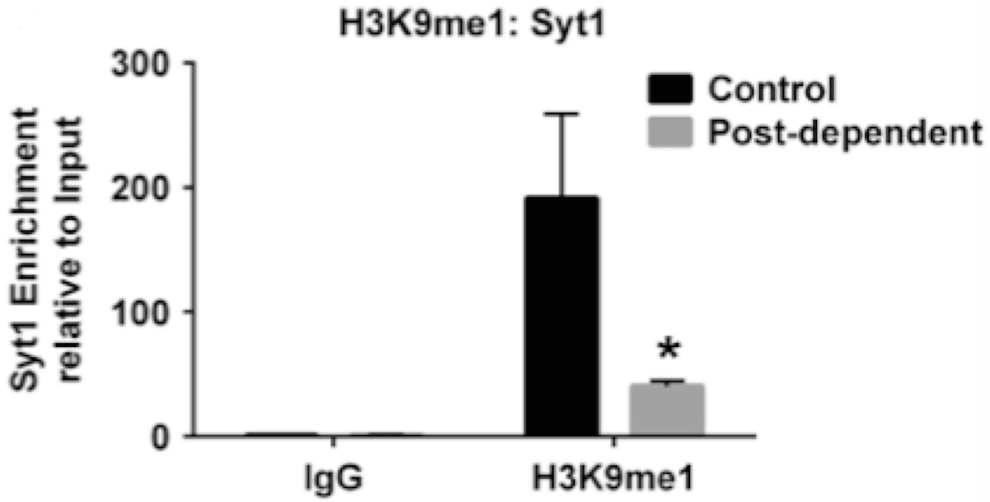


**Fig 7.** Reproduced with permission from Molecular Psychiatry advance online publication 23 August 2016. doi:10.1038/mp.2016.131. * denotes $P < 0.05$, where quantitative real-time PCR (qPCR) was used to verify differential H3K9me1 enrichment at the Syt1 gene in control versus postdependent rats. In this study, intersecting the output of multiple peak callers and annotating the resultant shortlist of peaks using *geneXtendeR* proved instrumental in pinpointing Syt1 amongst thousands of other potential gene candidates. This, in turn, validated the bioinformatic pipeline employed in the rest of the study.

The lesson learned from this case study was: intersecting the output of multiple peak callers and annotating the resultant shortlist of peaks using *geneXtendeR* in the context of a biological workflow proved instrumental in pinpointing Syt1 amongst thousands of other gene candidates (i.e., other annotated peaks).

bioRxiv preprint doi: https://doi.org/10.1101/082347; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



## Conclusion

Motivated by mounting evidence for a 3D genome chromatin looping model, we propose an R/Bioconductor package to be used as an integral part of modern histone modification ChIP-seq workflows. *geneXtendeR* optimally annotates a histone modification ChIP-seq peak input file with functionally important genomic features (e.g., genes associated with peaks) based on optimization calculations. As such, the user can effectively customize their ChIP-seq analysis to the tissue-specific and environment-specific details that inherently affect the broadness and location of epigenetic histone marks in their dataset.

## 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 ChIP-seq ENCODE analysis

***geneXtendeR* analysis on 215 human histone modification ChIP-seq ENCODE datasets.** Files available here:
https://github.com/Bohdan-Khomtchouk/ENCODE_histone_geneXtendeR_analysis

## Acknowledgments

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.

## Author's contributions

BBK conceived the study, designed the algorithms, wrote 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.

## References

1. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS: *MEME SUITE: tools for motif discovery and searching.* Nucleic Acids Research. 2009, 37 (2): W202–W208.

bioRxiv preprint doi: https://doi.org/10.1101/082347; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.




2. Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, Madrigal P, Taslim C, Zhang J: *Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data.* PLoS Computational Biology. 2013, 9(11): e1003326.

3. 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.

4. Calo E, Wysocka J: *Modification of Enhancer Chromatin: What, How, and Why?* Molecular Cell. 2013, 49 (5): 825–837.

5. Drissen R, Palstra RJ, Gillemans N, Splinter E, Grosveld F, Philipsen S, de Laat W: *The active spatial organization of the beta-globin locus requires the transcription factor EKLF.* Genes & Development. 2004, 18 (20): 2485–2490.

6. Furey TS: *ChIP-seq and Beyond: new and improved methodologies to detect and characterize protein-DNA interactions.* Nature Reviews Genetics. 2012, 13 (12): 840–852.

7. 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.

8. 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.

9. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, Ching KA, Antosiewicz-Bourget JE, Liu H, Zhang X, Green RD, Lobanenkov VV, Stewart R, Thomson JA, Crawford GE, Kellis M, Ren B: *Histone modifications at human enhancers reflect global cell-type-specific gene expression.* Nature. 2009, 459: 108–112.

10. Hon GC, Hawkins RD, Ren B: *Predictive chromatin signatures in the mammalian genome.* Human Molecular Genetics. 2009, 18 (2): R195–R201.

11. Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH: *An integrated software system for analyzing ChIP-chip and ChIP-seq data.* Nature Biotechnology. 2008, 26: 1293–1300.

12. Jing H, Vakoc CR, Ying L, Mandat S, Wang H, Zheng X, Blobel GA: *Exchange of GATA factors mediates transitions in looped chromatin organization at a developmentally regulated gene locus.* Molecular Cell. 2008, 29 (2):232–242.

13. 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.

bioRxiv preprint doi: https://doi.org/10.1101/082347; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.




14. Krig SR, Jin VX, Bieda MC, O'Geen H, Yaswen P, Green R, Farnham PJ: *Identification of genes directly regulated by the oncogene ZNF217 using chromatin immunoprecipitation (ChIP)-chip assays.* Journal of Biological Chemistry. 2007. 282: 9703–9712.

15. 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.

16. Lee K, Hsiung CC, Huang P, Raj A, Blobel GA: *Dynamic enhancer-gene body contacts during transcription elongation.* Genes & Development. 2015, 29 (19): 1992–1997.

17. Love MI, Huber W, Anders S: *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biology. 2014, 15: 550.

18. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E: *TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.* 2006. Nucleic Acids Research. 34 (Database issue): D108–110.

19. Maze I, Feng J, Wilkinson MB, Sun H, Shen L, Nestler EJ: *Cocaine dynamically regulates heterochromatin and repetitive element unsilencing in nucleus accumbens.* 2011. Proceedings of the National Academy of Sciences. 108 (7): 3035–3040.

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

21. Rintisch C, Heinig M, Bauerfeind A, Schafer S, Mieth C, Patone G, Hummel O, Chen W, Cook S, Cuppen E, Colomé-Tatché 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.

22. Robinson MD, McCarthy DJ, Smyth GK: *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics. 2010, 26: 139–140.

23. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B: *JASPAR: an open-access database for eukaryotic transcription factor binding profiles.* Nucleic Acids Research. 2004, 32 (Database issue): D91–D94.

24. Sandmann T, Jensen LJ, Jakobsen JS, Karzynski MM, Eichenlaub MP, Bork P, Furlong EEM: *A temporal map of transcription factor activity: mef2 directly regulates target genes at all stages of muscle development.* Developmental Cell. 2006, 10: 797–807.

25. Sanyal A, Lajoie BR, Jain G, Dekker J: *The long-range interaction landscape of gene promoters.* Nature. 2012, 489 (7414): 109–113.

bioRxiv preprint doi: https://doi.org/10.1101/082347; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



26. Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0. 2013-2015 <http://www.repeatmasker.org>.

27. 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.

28. Taslim C, Wu J, Yan P, Singer G, Parvin J, Huang T, Lin S, Huang K: *Comparative study on ChIP-seq data: normalization and binding pattern characterization.* Bioinformatics. 2009, 25: 2334–2340.

29. Vakoc CR, Letting DL, Gheldof N, Sawado T, Bender MA, Groudine M, Weiss MJ, Dekker J, Blobel GA: *Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1.* Molecular Cell. 2005, 17 (3):453–462.

30. Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W: *A clustering approach for identification of enriched domains from histone modification ChIP-Seq data.* Bioinformatics. 2009, 25(15): 1952–1958.

31. 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, 9(9): R137.