RESEARCH ARTICLE

PAPST, a User Friendly and Powerful Java Platform for ChIP-Seq Peak Co-Localization Analysis and Beyond

Paul W. Bible[1]*, Yuka Kanno[2], Lai Wei[3], Stephen R. Brooks[4], John J. O'Shea[2], Maria I. Morasso[1], Rasiah Loganantharaj[5], Hong-Wei Sun[4]*

1 Laboratory of Skin Biology, Intramural Research Program, National Institute of Arthritis and Musculoskeletal and Skin Diseases, Bethesda, Maryland, United States of America, **2** Molecular Immunology and Inflammation Branch, Intramural Research Program, National Institute of Arthritis and Musculoskeletal and Skin Diseases, Bethesda, Maryland, United States of America, **3** State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China, **4** Biodata Mining and Discovery Section, Office of Science and Technology, Intramural Research Program, National Institute of Arthritis and Musculoskeletal and Skin Diseases, Bethesda, Maryland, United States of America, **5** Laboratory of Bioinformatics, Center for Advanced Computer Studies, University of Louisiana at Lafayette, Lafayette, Louisiana, United States of America

* paul.bible2@nih.gov (PWB); sunh1@mail.nih.gov (HWS)

OPEN ACCESS

Citation: Bible PW, Kanno Y, Wei L, Brooks SR, O'Shea JJ, Morasso MI, et al. (2015) PAPST, a User Friendly and Powerful Java Platform for ChIP-Seq Peak Co-Localization Analysis and Beyond. PLoS ONE 10(5): e0127285. doi:10.1371/journal.pone.0127285

Academic Editor: Vladimir B. Bajic, King Abdullah University of Science and Technology, SAUDI ARABIA

Received: December 10, 2014

Accepted: April 14, 2015

Published: May 13, 2015

Data Availability Statement: "Software and test data can be downloaded at https://github.com/paulbible/papst."

Funding: This work was supported by the Intramural Research Program and the Office of Science and Technology of the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of Health.

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

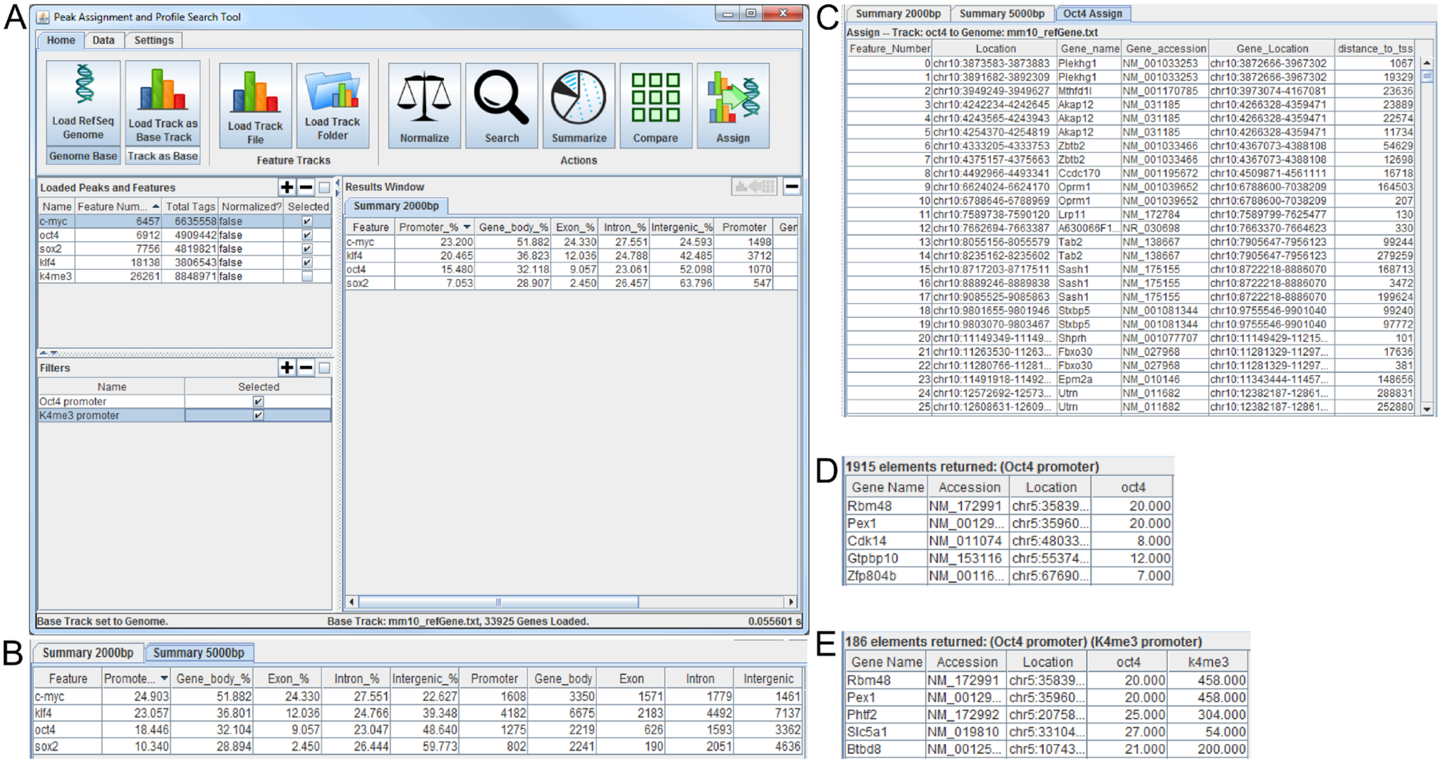
Abstract

Comparative co-localization analysis of transcription factors (TFs) and epigenetic marks (EMs) in specific biological contexts is one of the most critical areas of ChIP-Seq data analysis beyond peak calling. Yet there is a significant lack of user-friendly and powerful tools geared towards co-localization analysis based exploratory research. Most tools currently used for co-localization analysis are command line only and require extensive installation procedures and Linux expertise. Online tools partially address the usability issues of command line tools, but slow response times and few customization features make them unsuitable for rapid data-driven interactive exploratory research. We have developed PAPST: Peak Assignment and Profile Search Tool, a user-friendly yet powerful platform with a unique design, which integrates both gene-centric and peak-centric co-localization analysis into a single package. Most of PAPST's functions can be completed in less than five seconds, allowing quick cycles of data-driven hypothesis generation and testing. With PAPST, a researcher with or without computational expertise can perform sophisticated co-localization pattern analysis of multiple TFs and EMs, either against all known genes or a set of genomic regions obtained from public repositories or prior analysis. PAPST is a versatile, efficient, and customizable tool for genome-wide data-driven exploratory research. Creatively used, PAPST can be quickly applied to any genomic data analysis that involves a comparison of two or more sets of genomic coordinate intervals, making it a powerful tool for a wide range of exploratory genomic research. We first present PAPST's general purpose features then apply it to several public ChIP-Seq data sets to demonstrate its rapid execution and potential for cutting-edge research with a case study in enhancer analysis. To our knowledge, PAPST is the first software of its kind to provide efficient and sophisticated post peak-calling ChIP-Seq data analysis as an easy-to-use interactive application. PAPST is available at https://github.com/paulbible/papst and is a public domain work.

Introduction

ChIP-Seq, one of the most powerful research applications of Next Generation Sequencing (NGS) technology, is fundamental to the investigation of TF-to-DNA interactions and histone modifications. A general ChIP-Seq data analysis pipeline includes a) mapping sequence reads to a genome, b) calling statistically significant peaks, and c) broad down-stream analysis of called peaks including co-localization analysis and assignment of peaks to genes and other annotated genomic regions. There are well-established mapping programs [1–3] and a large number of peak-calling programs [4–8] including specialized peak-callers with robust normalization features [9]. However, availability of easy-to-use yet powerful platforms for post peak-calling data analysis has been limited. Addressing this need is particularly important for bench scientists who, in most cases, are responsible for making biological sense of the data. Chojnowski et al. [10] has recently highlighted the lack of packages for post peak-calling analysis and the need for more user-friendly software to empower a greater number of researchers in the community. Their program, jChIP, is the only downstream ChIP analysis package with a cross platform graphical interface. That program focuses on visualization of raw data as histograms and viewing quality check information. Two other recently developed tools, with some co-localization analysis features, include PAVIS [11] and ChIPseek [12]. Both are web-based. The former is peak-centric only, and the latter brings online some features of popular Linux command line tools such as HOMER [13] and BEDTools [14]. Online tools such as these have the advantage of few system requirements and support more devices. However, this advantage comes at the cost of limits placed on the amount of data processed, longer response times from the software, and fewer customization features. In comparison, a stand-alone desktop tool gives the user more customization and control over the data being analyzed, as successfully demonstrated by the popular Java platform genomics viewer IGV [15].

We have developed PAPST (peak assignment and profile search tool), with bench scientists in mind, as an easy-to-use yet powerful stand-alone platform for post peak-calling data analysis, focusing on co-localization analysis of transcription factors and histone modifications, a critical component of integrated ChIP-Seq data analysis (Fig 1A). PAPST has been especially designed for efficient data-driven exploratory research. Among its most notable distinct features are its lightning-fast speed, its capability to perform customizable co-localization analysis on an unlimited number of factors, and its unique design that integrates both gene-centric and peak-centric (see the Methods section for Special term definitions) ChIP-Seq data analysis into a single package. These powerful and unique features are highly desirable in data-driven exploratory research, where data analysis parameters can be quickly changed and the output results can easily be used as inputs for the next round of analysis on a fine-tuned hypothesis. We present first examples of the main features of PAPST then apply it to select mouse embryonic stem cell (ESC) ChIP-Seq data [16, 17] and demonstrate its typical applications in post peak-calling ChIP-Seq data analysis, co-localization analysis in particular. The case study presented in this work addresses the analysis of typical and super ESC enhancers ([16] also see the Special term definitions). Specifically, we show PAPST's rapid applications in a) identifying ESC enhancers as regions co-localized by the master TFs Oct4, Sox2, and Nanog; b) screening for genes with co-promoter-occupancy of the three master TFs and H3K27Ac; c) comparing peak



Loaded Peaks and Features

Name	Feature Num...	Total Tags	Normalized?	Selected
c-myc	6457	6635558	false	☑
oct4	6912	4909442	false	☑
sox2	7756	4819821	false	☑
klf4	18138	3806543	false	☑
k4me3	26261	8848971	false	☐

Filters

Name	Selected
Oct4 promoter	☑
K4me3 promoter	☑

Summary 2000bp

Feature	Promoter_%	Gene_body_%	Exon_%	Intron_%	Intergenic_%	Promoter	Ger
c-myc	23.200	51.882	24.330	27.551	24.593	1498	
klf4	20.465	36.823	12.036	24.788	42.485	3712	
oct4	15.480	32.118	9.057	23.061	52.098	1070	
sox2	7.053	28.907	2.450	26.457	63.796	547	

Base Track set to Genome. Base Track: mm10_refGene.txt, 33925 Genes Loaded. 0.055601 s

Summary 2000bp | Summary 5000bp

Feature	Promote...	Gene_body_%	Exon_%	Intron_%	Intergenic_%	Promoter	Gene_body	Exon	Intron	Intergenic
c-myc	24.903	51.882	24.330	27.551	22.627	1608	3350	1571	1779	1461
klf4	23.057	36.801	12.036	24.766	39.348	4182	6675	2183	4492	7137
oct4	18.446	32.104	9.057	23.047	48.640	1275	2219	626	1593	3362
sox2	10.340	28.894	2.450	26.444	59.773	802	2241	190	2051	4636

Summary 2000bp | Summary 5000bp | Oct4 Assign

Assign -- Track: oct4 to Genome: mm10_refGene.txt

Feature_Number	Location	Gene_name	Gene_accession	Gene_Location	distance_to_tss
0	chr10:3873583-3873883	Plekhg1	NM_001033253	chr10:3872666-3967302	1067
1	chr10:3891682-3892309	Plekhg1	NM_001033253	chr10:3872666-3967302	19329
2	chr10:3949249-3949627	Mthfd1l	NM_001170785	chr10:3973074-4167081	23636
3	chr10:4242234-4242645	Akap12	NM_031185	chr10:4266328-4359471	23889
4	chr10:4243565-4243943	Akap12	NM_031185	chr10:4266328-4359471	22574
5	chr10:4254370-4254819	Akap12	NM_031185	chr10:4266328-4359471	11734
6	chr10:4333205-4333753	Zbtb2	NM_001033466	chr10:4367073-4388108	54629
7	chr10:4375157-4375663	Zbtb2	NM_001033466	chr10:4367073-4388108	12698
8	chr10:4492966-4493341	Ccdc170	NM_001195672	chr10:4509871-4561111	16718
9	chr10:6624024-6624170	Oprm1	NM_001039652	chr10:6788600-7038209	164503
10	chr10:6788646-6788969	Oprm1	NM_001039652	chr10:6788600-7038209	207
11	chr10:7589738-7590120	Lrp11	NM_172784	chr10:7589799-7625477	130
12	chr10:7662694-7663387	A630066F1...	NR_030698	chr10:7663370-7664623	330
13	chr10:8055156-8055579	Tab2	NM_138667	chr10:7905647-7956123	99244
14	chr10:8235162-8235602	Tab2	NM_138667	chr10:7905647-7956123	279259
15	chr10:8717203-8717511	Sash1	NM_175155	chr10:8722218-8886070	168713
16	chr10:8889246-8889838	Sash1	NM_175155	chr10:8722218-8886070	3472
17	chr10:9085525-9085863	Sash1	NM_175155	chr10:8722218-8886070	199624
18	chr10:9801655-9801946	Stxbp5	NM_001081344	chr10:9755546-9901040	99240
19	chr10:9803070-9803467	Stxbp5	NM_001081344	chr10:9755546-9901040	97772
20	chr10:11149349-11149...	Shprh	NM_001077707	chr10:11149429-11215...	101
21	chr10:11263530-11263...	Fbxo30	NM_027968	chr10:11281329-11297...	17636
22	chr10:11280766-11281...	Fbxo30	NM_027968	chr10:11281329-11297...	381
23	chr10:11491918-11492...	Epm2a	NM_010146	chr10:11343444-11457...	148656
24	chr10:12572692-12573...	Utrn	NM_011682	chr10:12382187-12861...	288831
25	chr10:12608631-12609...	Utrn	NM_011682	chr10:12382187-12861...	252880

1915 elements returned: (Oct4 promoter)

Gene Name	Accession	Location	oct4
Rbm48	NM_172991	chr5:35839...	20.000
Pex1	NM_00129...	chr5:35960...	20.000
Cdk14	NM_011074	chr5:48033...	8.000
Gtpbp10	NM_153116	chr5:55374...	12.000
Zfp804b	NM_00116...	chr5:67690...	7.000

186 elements returned: (Oct4 promoter) (K4me3 promoter)

Gene Name	Accession	Location	oct4	k4me3
Rbm48	NM_172991	chr5:35839...	20.000	458.000
Pex1	NM_00129...	chr5:35960...	20.000	458.000
Phtf2	NM_172992	chr5:20758...	25.000	304.000
Slc5a1	NM_019810	chr5:33104...	27.000	54.000
Btbd8	NM_00125...	chr5:10743...	21.000	200.000

Fig 1. PAPST in action. The graphic design of PAPST was modelled after familiar office productivity software. A. Using the PAPST Summarize command shows the distribution of peaks relative to genomic features. B. Modifying the promoter distance changes the results allowing customization. C. PAPST assigns genes to their nearest TSS. D & E. PAPST searches for genes having a pattern of (D) Oct4 binding in their promoters or (E) K4me3 binding in their promoters and an Oct4 tag value of greater than 20.

doi:10.1371/journal.pone.0127285.g001

intensities of multiple relevant TFs and EMs in the ESC super enhancers to those in the typical enhancers; d) assigning the ESC super enhancers to key ESC identity genes; and e) performing peak overlap based clustering of 13 TFs. Our PAPST based ESC ChIP-Seq data analysis results, all obtained through a few mouse clicks and within seconds, are highly consistent with those reported in the literature, validating the specific computational approaches implemented in the platform. To our knowledge, PAPST is the first standalone graphical desktop application of its kind for efficient and sophisticated post peak-calling data-driven ChIP-Seq data analysis. PAPST is a cross-platform Java application that is free for any use as part of the public domain. The program package may be downloaded at https://github.com/paulbible/papst, which provides the application jar, a collection of ready to use TF and EM ChIP-Seq data from [17], a comprehensive user guide, and individual tutorials for both PC and Mac users.

Results

PAPST performs common downstream ChIP-Seq analysis tasks quickly and interactively

Certain fundamental tasks are common in nature to all post peak-calling downstream ChIP-Seq analysis. Fundamentally, called ChIP-Seq peaks can be represented as a set of genomic coordinates and nearly all common ChIP-Seq analysis methods rely on determining whether peaks overlap or are near to known genomic features (gene-centric) or user-defined genomic features such as peaks from a specific TF experiment (peak-centric). At the heart of PAPST are efficient algorithms (see Methods section) to quickly process a large number of genomic locations to answer distance and overlap queries. The subsections that follow describe PAPST's solutions to the most general ChIP-Seq analysis tasks. Co-localization analysis, performed by processing the overlaps of multiple sets of significant ChIP-Seq peaks from multiple experiments is an example of one of PAPST's features presented in the following sections. All the data sets used in this section are available online as part of the tutorial, which presents step by step instructions of example applications. A detailed user guide is also included in the online distribution package of PAPST.

Location based distribution analysis. At the global genomic level, the distribution of protein binding with respect to known features provides a fundamental description of binding behavior. PAPST analyzes peak binding in the promoter, gene body, exon, intron, and Intergenic regions quickly with a single mouse click. The summary distribution information is reported as a percentage of total peaks as well as the number of significant peaks that fall into the genomic region. PAPST was used to calculate the binding distribution for four TFs. The result window of PAPST in Fig 1A shows the distributions for Oct4, Sox2, c-Myc, and Klf4 using a 2000 bp promoter. The results using 5000 bp promoter are shown in Fig 1B. Most analysis packages use a one-size-fits-all promoter size while PAPST allows the user to interactively consider multiple settings. Like most features in PAPST, this analysis can be performed simultaneously on multiple sets of ChIP-Seq peaks.

Peak assignment to proximal genes. Post peak-calling ChIP-Seq analysis of TFs invariably begins by assigning peaks to the nearest gene transcription start site (TSS) to dissect their behavior. Many TFs drive gene regulation by interacting with the upstream promoter of target genes. Although the current understanding of gene regulation has evolved to include distal TF-to-target gene regulation, assignment to the nearest TSS provides the baseline preliminary analysis that offers insight to TF behavior [16]. This feature is common to nearly all ChIP-Seq analysis packages. PAPST assigns peaks to the nearest gene and displays the results as a sortable table in the results window (Fig 1C). PAPST allows any result table generated to be exported to a spreadsheet format. The complete gene assignment for Oct4 peaks is provided in S1 Table. PAPST also offers batch gene assignment functionality that will calculate the gene assignments for multiple peak sets and place the results into separate files.

Robust gene-based search features. Labs studying specific TFs are typically interested in discovering which genes could be direct regulatory targets. PAPST offers customizable search capabilities that allow users to quickly determine which genes have particular TF binding patterns. The powerful filters feature allows users to create sets of criteria for returning genes of interest. A simple but common task would be to find all genes where a specific TF binds in the promoter. To accomplish this, a PAPST user would simply create a filter, select the promoter option (modify the length if desired), and press the search button. The S2 Table contains the search results for genes having Oct4 in their upstream promoters. A sample of these search results is shown in Fig 1D. The genes are placed in a sortable table in the results window allowing the user to prioritize genes based on total tag coverage.

The above example represents the most basic application of the search features. Much more interesting questions are addressed by combining multiple transcription factors and/or histone marks into the search criteria. A natural extension to the promoter search example described above would be to include the activating histone mark H3K4me3 and filter out lower confidence peaks by tag count. To accomplish this a second filter is added to the search that operates on significant peaks derived from an H3K4me3 ChIP-Seq experiment, and the threshold of the Oct4 filter is adjusted to filter out peaks with fewer than 20 reads. PAPST will process all filters and return only genes that meet all the specified criteria (those with both Oct4 and H3K4me3 binding at the promoter and at least 20 reads for the Oct4 peaks). Fig 1E shows a sample of the results returned using these search criteria (the complete list is available in S3 Table). Creative application of an arbitrary number of peaks and filters provides the user with a robust means of specifying complex regulatory and epigenetic patterns to select interesting genes for further analysis.

Robust peak and general interval-based search features. This powerful feature, unique to PAPST, generalizes the multiple criteria based search function to any set of genomic regions enabling peak-centric co-localization analysis. In the previous example, a search was constructed to find genes that met multiple criteria across different experiments. Setting the search mode to "Track as Base" enables the selection of a peak set to serve as the genomic locations that are subjected to multiple user-defined filters. This feature allows complex search criteria to be applied to any set of arbitrary genomic regions of interest. As the current understanding of gene regulation evolves, gene-distal regulatory regions will become increasingly important. PAPST provides easy to use tools to explore protein co-localization patterns at any user-defined set of genomic regions. The Example Applications section below describes a case study that utilizes this feature to detect enhancers. Another example of peak-based search features is presented in the tutorial at https://github.com/paulbible/papst.

Performance and Usability. Efficient performance and usability are the two main design principles of PAPST. All PAPST's functions can be selected and executed through a graphical user interface with a few mouse clicks, and the results will return within seconds. In addition, we have also implemented specific features to increase its usability and enhance its ability for efficient data handling. A few examples of PAPST's usability features are: 1) imbedded customizable parsers simplify importing data from a wide range of file formats, 2) an analysis session of any number of peak sets can be saved into a single file to revisit later or share with a colleague without reloading individual data sets, 3) analysis results can be saved as a csv file for further analysis by spreadsheet programs, 4) specific entries may be easily copied and pasted into the UCSC genome browser for viewing, and 5) analysis results may be quickly reformatted within PAPST and used as a new input for the next round of co-localization analysis, greatly simplifying iterative analyses.

To demonstrate its efficient performance and ease of use, we have timed a complete analysis session in which we loaded Oct4, Sox2, and Nanog peaks, determined regions of co-localized binding, loaded the literature derived normal and super enhancer regions [16], and calculated their agreement with PAPST derived regions in terms of overlapping peaks. PAPST completed the entire 7-step analysis session within 19 seconds (wall clock time), from data loading to result generation. The detailed data are given in Table 1, including the timings for each specific step and the number of mouse clicks needed. We have also tested PAPST, using a typical laptop PC (IBM ThinkPad, 64 bit Windows 7, 8 GB RAM, Intel Core i7 2620M CPU @ 2.70 GHz), on a hypothetical data set of 50 factors (each with 26,261 peaks 1.3 MB each, 65 MB total size) and a co-localization analysis was completed within 1.39 seconds against a peak set (26,261 genomic locations, peak-centric analysis) and 1.54 seconds against RefGene (34150 transcripts, gene-centric analysis), respectively.

Example Applications of PAPST

The following sections describe case applications of PAPST, with pubic ChIP-Seq datasets, to demonstrate its typical uses and its great potential in cutting-edge research. The results as compared to those published also validate the algorithms and approaches specifically implemented in PAPST.

Co-localization analysis to identify enhancers and screen for genes with common TF bindings. Efficient co-localization analysis, both peak-centric and gene-centric, is the key

Table 1. A Complete Analysis Session in PAPST with Timing Information.

Task	Number of mouse clicks	Time
Load peak files of Oct4, Sox2, and Nanog (23986, 16056, 22463 peaks, respectively)	3	3 sec
Assign Oct4 peaks as the base	4	4 sec
Define search criteria by creating a filter	4	4 sec
Search for regions co-localized by the 3 TFs	1	1 sec (the search operation itself took 0.051 sec)
Reformat the search results into an input file	4	4 sec
Load genomic locations of super-enhancers and typical enhancers	2	2 sec
Calculate overlaps for all 20 pairwise comparisons	1	1 sec
Total	19 mouse clicks	19 seconds

PAPST is a very fast and easy to use platform for exploratory research as illustrated by the number of clicks to accomplish a task as well as the time.

doi:10.1371/journal.pone.0127285.t001

feature of PAPST. ESC enhancers are characterized by the co-occupancy of master transcription factors Oct4, Sox2, and Nanog [17]. Young's lab has recently used this co-occupancy of the master TFs as a main criterion to identify 8794 enhancers including 231 super enhancers [16]. We applied PAPST, using peak-centric analysis, to the same mouse ESC ChIP-Seq data (after BOWTIE mapping and MACS peak calling) and identified 10110 regions that are co-localized by the three master TFs. After merging close-by regions as previously described [16], a total of 8787 enhancer regions were identified by PAPST, which overlap 95% and 99%, respectively, of the 8563 typical enhancers and 231 super enhancers reported by the paper (Fig 2). In general, given a set of genomic coordinates such as those defined by the binding peaks of a transcription factor, PAPST can be used to quickly identify unlimited number of co-localized transcription factors and/or histone modifications, or any other defined genomic regions of interest.

PAPST can also perform customizable (user defined promoter, gene body, and extended gene body) gene-centric co-localization analysis to screen for genes with a specific occupancy profile of TFs and/or EMs. For example, we have applied PAPST to identifying 1182 non-redundant ESC genes with co-promoter-occupancy of Oct4, Sox2, Nanog, and H3K27Ac (S4 Table). The gene-centric co-localization analysis can be particularly useful to address specific questions related to gene regulation (see the tutorials for more examples).

Analysis of peak intensity within the defined genomic regions of super enhancers vs typical enhancers. PAPST can generate quantitative data for extended co-localization analysis. ESC super enhancers have been shown to have higher levels of active enhancer epigenetic mark H3K27Ac and binding of key TFs as compared to typical enhancers [16]. We used PAPST to perform total tags based peak signal normalization of ChIP-Seq peaks for Oct4, Sox2, Nanog, Mediator (Med1), and H3K27Ac. Next PAPST was used to generate normalized read signals for these factor's peaks in super enhancer regions and in typical enhancer regions respectively. The comparative results are shown in Fig 3, which indicate significantly higher levels of these key factors in the super enhancer associated peaks than those associated with the typical enhancers (p-values are: Oct4 2.26E-33, Sox2 1.96E-30, Nanog 1.30E-19, H3K27Ac 1.78E-32, and Med1 4.17E-12 using Welch's t-test). We also used PAPST to quickly generate the co-localization data showing a significantly higher percentage of super enhancers are occupied by

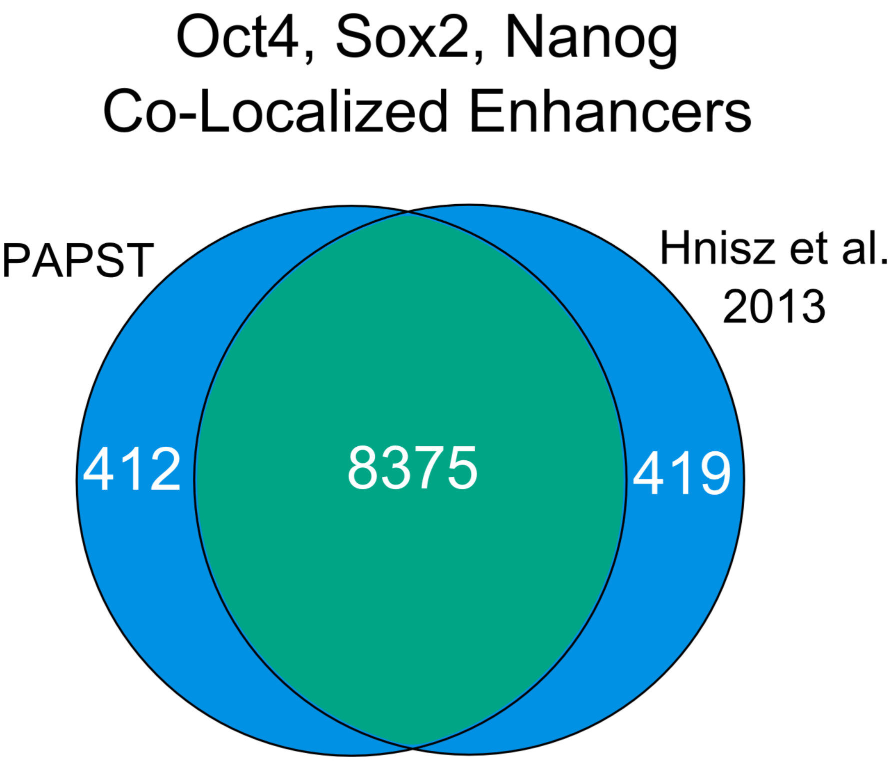


Fig 2. Comparison of PAPST Derived Enhancers to the Literature. Comparison of PAPST identified ESC enhancers with published ESC enhancers [16] shows that PAPST identified over 95% of those found by Hnisz et al. The enhancers were identified through the calculation of genomic regions co-localized by Oct4, Sox2, and Nanog.

doi:10.1371/journal.pone.0127285.g002

H3K27Ac and Med1 compared to typical enhancers (Fig 4). In these rapid applications of PAPST (see the Performance and Usability section above for the timings), co-localized peaks are not only easily identified, but they can also be investigated quantitatively.

Assignment of super enhancers to key ESC genes by proximity. PAPST can quickly assign a set of genomic regions such as ChIP-Seq peaks to genes by proximity to their transcription start site (TSS). Using PAPST, we assigned ESC super enhancers to their closest genes. The complete assignment results are presented in the supplemental materials (S5 Table). Importantly, among these PAPST assigned super enhancer associated genes are the 14 key ESC cell identity genes (Oct4, Sox2, NANOG, Tet1, Tet2, Dppa5a, n-Myc, Tbx3, Utf1, Esrrb, Prdm14, Klf4, Sall4, and Zfp42) that were also identified as the super enhancer associated genes in the original paper [16]. We have also assigned these super enhancers to genes using HOMER [13]. A comparison of the gene assignment results obtained with PAPST and the other two methods is given in Fig 5, and shows that PAPST and HOMER assigned these super enhancers to the same set of genes with an agreement of greater than 98%. The difference between the PAPST and HOMER assigned genes and those assigned by the original paper is most likely due to the specific version of RefGene used in the analysis. The minor differences between PAPST and HOMER assigned genes reflect the specific assignment algorithm's implementation details and likely represent differences in the transcript sort order of genes with multiple isoforms.

Peak overlap calculation and clustering reveals known ESC transcription factor relationships. PAPST has implemented a novel algorithm for symmetric peak overlap calculation,

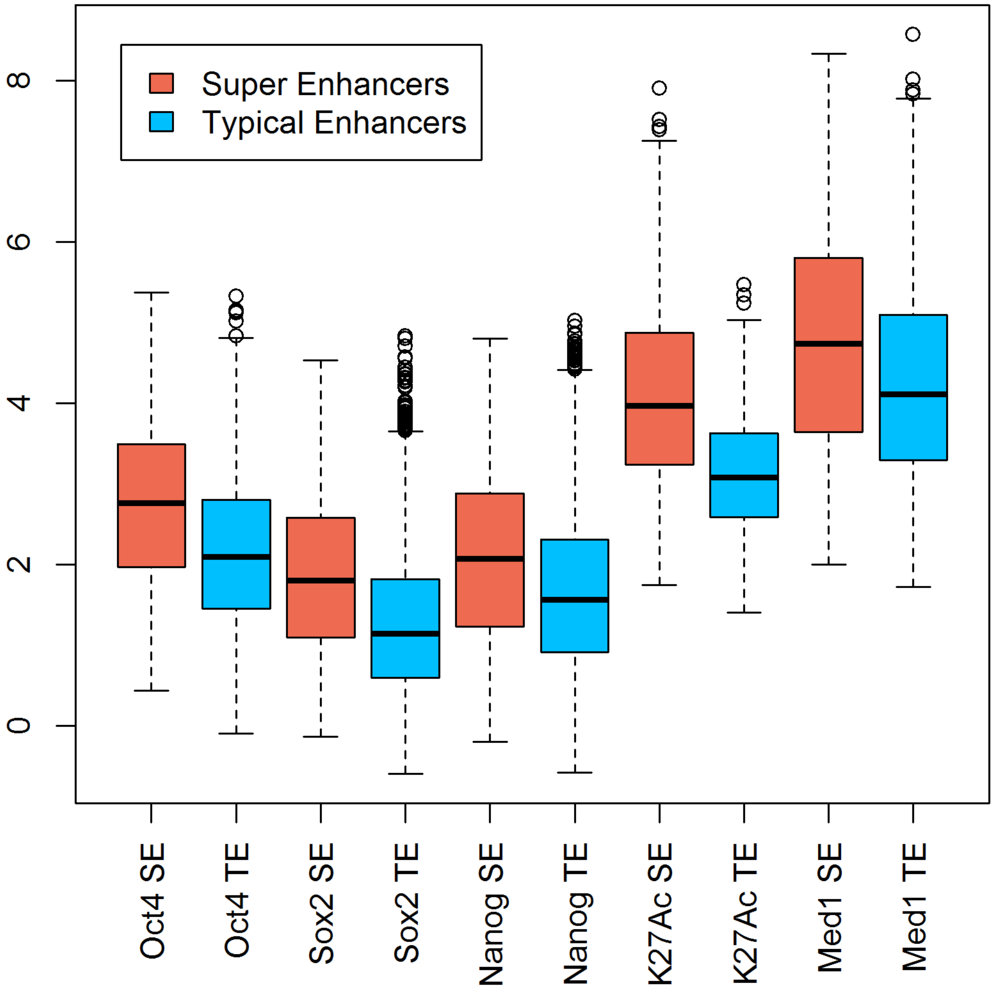


Fig 3. Signal Distribution of Key Transcription Factors in Super Enhancers and Typical Enhancers. Quantitative applications of PAPST compared peak signals within super-enhancers (SEs) to those within typical enhancers (TEs) of five factors showing significantly stronger signal in SEs. Red: peak signal distribution within super-enhancers. Green: peak signal distribution within typical enhancers. Peak signals are expressed as normalized read counts. All comparisons are significant with p < 1e-10 (Welch's t-test).

doi:10.1371/journal.pone.0127285.g003

determining the total number of overlapping peaks for all possible pairwise comparisons of TFs and EMs included in an analysis, with one single mouse click (see Methods for details). The symmetric overlapping relationships table may then be used for binding profile based clustering. We have used PAPST to calculate the overlapping peaks of 78 pairwise comparisons among 13 ESC transcription factors (S6 Table), using another ChIP-Seq data set publicly available [17]. The results were used to cluster these factors (see S1 Appendix for details). The clustering patterns obtained with PAPST generated data (Fig 6) are virtually identical to those published, which were created with a different set of mostly command-line based tools and algorithms [17].

Discussion

We have demonstrated PAPST's main features by applying it to two publicly available ChIP-Seq data sets from embryonic stem cell research [16, 17]. Our results, which were obtained

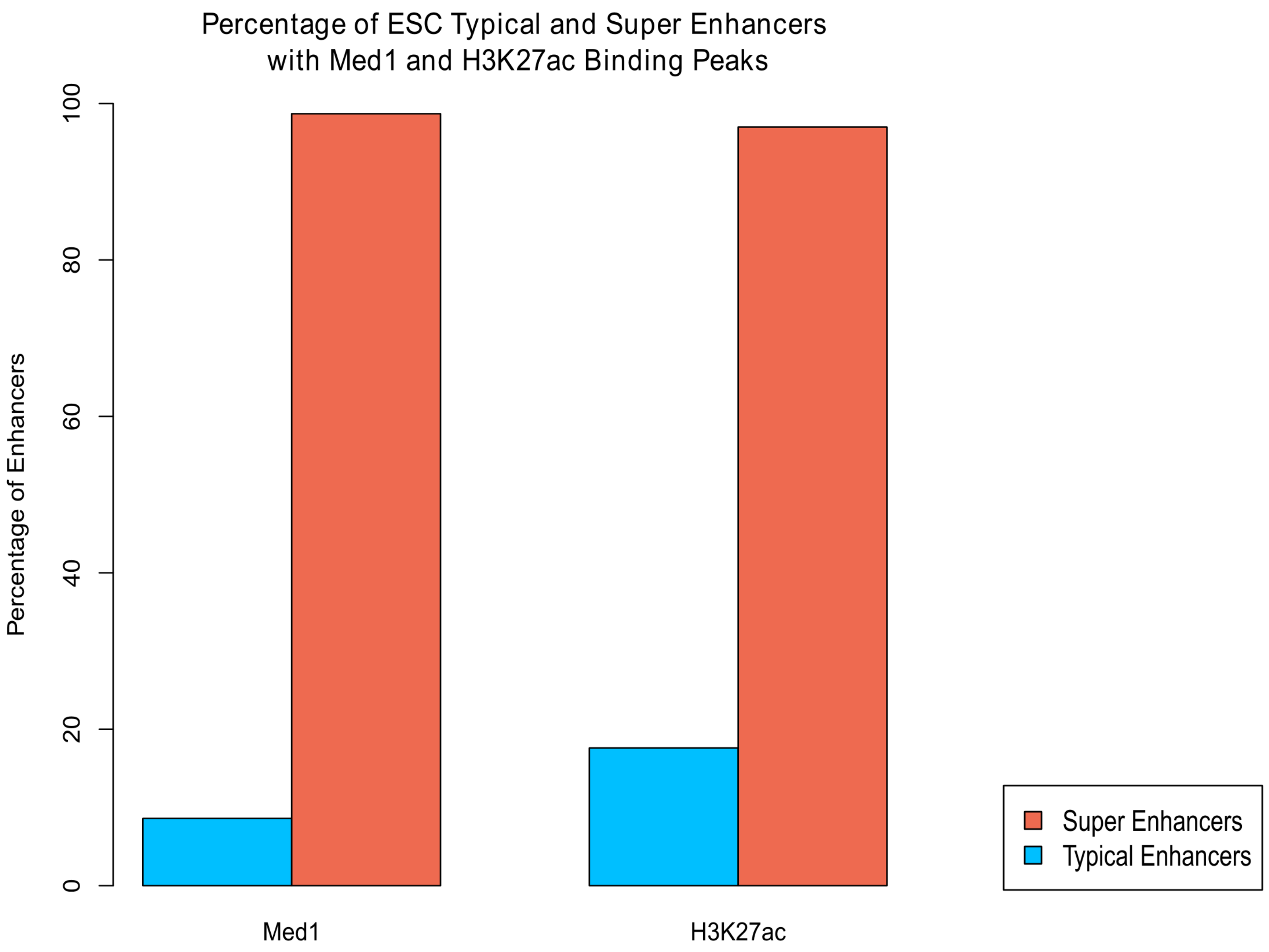


Fig 4. Comparison of Mediator (Med1) and H3K27ac occupancy in Super Enhancers and Typical Enhancers. PAPST quickly identified that significantly more SEs are occupied by Med1 and H3K27Ac than TEs. Quickly answering questions of this nature make PAPST ideal for exploratory research.

doi:10.1371/journal.pone.0127285.g004

with PAPST on the scale of seconds, are highly consistent with those presented in the original research publications, validating the specific computational algorithms implemented in PAPST.

PAPST has been developed based on our extensive experience in ChIP-Seq data analysis [18–21], which has helped us to realize the need for an easy to use yet powerful software tool for post peak-calling ChIP-Seq data analysis that is more accessible to the researchers in the field, including bench scientists. And as such, PAPST has been designed to offer a focused set of powerful and convenient features that are very easy to use for researchers with or without computational expertise.

Two of the most powerful and unique features of PAPST are: 1) simultaneous co-localization analysis of a large number of TFs and EMs (limited only by a computer's memory), and 2) the extremely fast return of the analysis results on the scale of seconds. A co-localization calculation of 50 factors took only 1.39 seconds and 1.54 seconds for peak-centric analysis and gene-centric analysis, respectively. As most functions of PAPST may be completed basically as

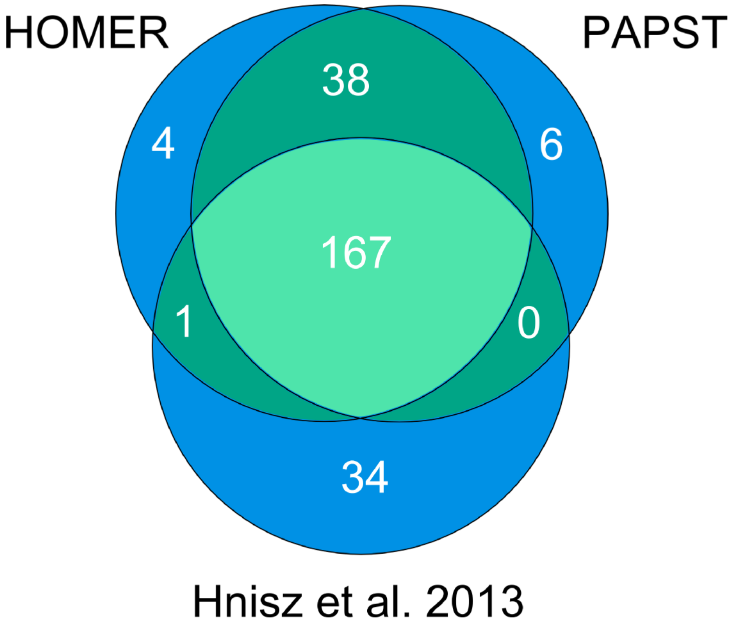


Fig 5. PAPST Gene Assignment of Super Enhancers Compared to HOMER and the Literature. PAPST was compared to two others gene assignments for super enhancer regions. PAPST and HOMER are 98% consistent with each other. The discrepancy with the Hnisz et al. [16] assignment is most likely due to the different versions of RefGene used (mostly new non-coding gene annotations).

doi:10.1371/journal.pone.0127285.g005

fast as one can click the mouse (Table 1), and the analysis results can be easily reformatted into an input file within the package for a new round of analysis, PAPST can serve as a powerful and efficient tool for data-driven exploratory research. This is particularly true when combined with the flexible feature of PAPST that allows easy and highly customizable parameter adjustment.

PAPST offers both gene centric and peak centric features in a single package, capable of analyzing ChIP-Seq peaks both on gene-defined genomic regions (promoters, exons, introns, gene bodies) and on peak-defined genomic regions, making it a true genome-wide data analysis platform. Its peak-centric feature can be creatively and easily extended to include the co-localization analysis of TFs and EMs on any set of user-defined genomic regions of interest, such as the ENCODE determined locations of regulatory sequences [22], regions of evolutionary conservation [23], or non-coding RNAs [24]. PAPST can thus be used to address both types of general questions in ChIP-Seq data analysis: a) which genes have a specific co-localization profile of TFs and EMs? and b) which genomic locations of interest are co-localized by a set of TFs and EMs? The interesting genes identified by PAPST can be further analyzed in combination with other data types such as those from RNA-Seq or subject to a pathway analysis. Those co-localized regions identified by PAPST can either be used for extended co-localization analysis or for further downstream analysis such as motif and composition analysis. In addition, as the core analysis performed by PAPST is genomic interval based, it may also potentially find its applications in any genome data analysis that involves genomic interval based calculations.

Finally, PAPST can also generate quantitative data matrices in the form of normalized peak reads within regions defined by genes or peaks, making it possible to study epigenetic and regulatory data with other quantitative statistical methods such as PCA, K-means and hierarchical clustering, and statistical learning based modeling.

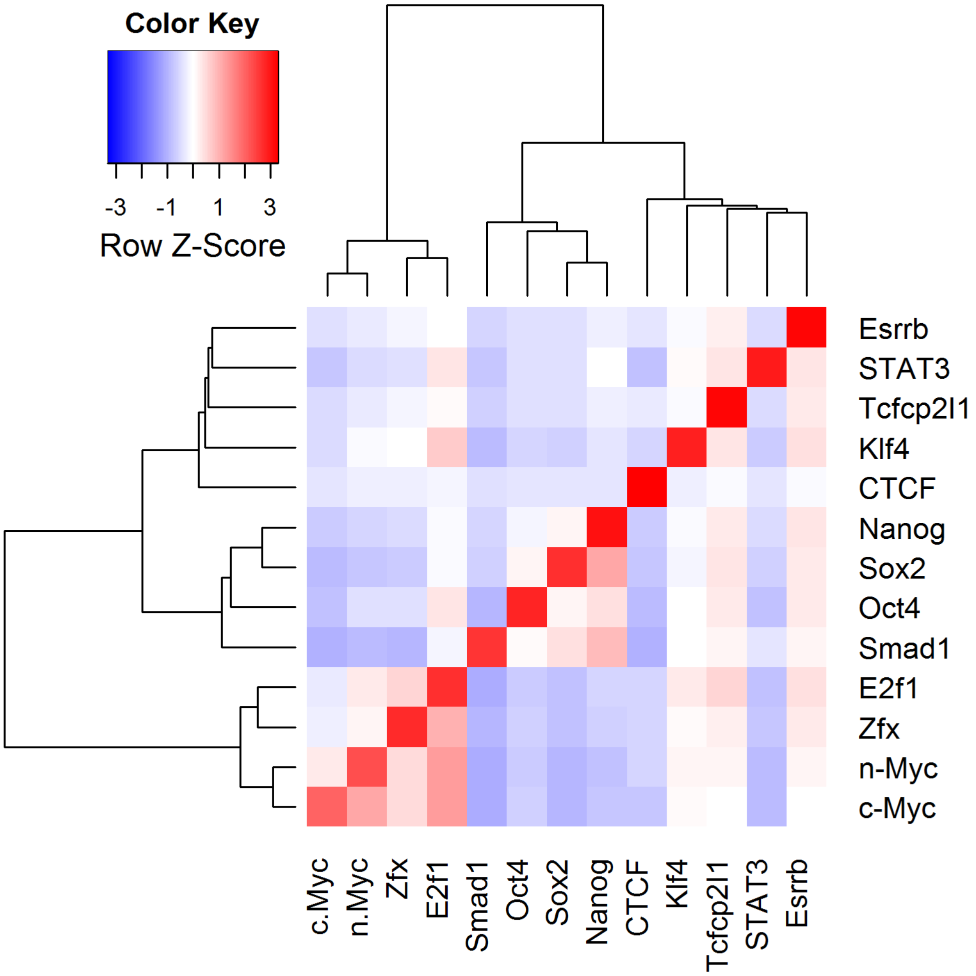


Fig 6. Overlap Based Clustering of 13 ESC Transcription Factors. Using PAPST's symmetric overlap algorithm, 13 ESC Transcription were clustered. The heatmap shows the association of the 3 key reprogramming factors Oct4, Sox2, and Nanog. These groupings closely match those of Chen et al. [17].

doi:10.1371/journal.pone.0127285.g006

Materials and Methods

Data, read mapping, and peak calling

The raw data was collected from NCBI GEO using accessions GSE44288 [16] and GSE11431 [17]. Reads were mapped with Bowtie 0.12.8 [2] against mouse genome mm9. For the clustering, mm9 mapping data from the original paper were lifted to mm10 using liftOver from UCSC Genome Browser [25]. Peaks were called with MACS 1.4.2 [8]. The specific parameters used with Bowtie and MACS were chosen as close to those used in the original papers as practically possible. The boxplots and the clustering were made with R. Complete details are available in S1 Appendix.

Design and Implementation

PAPST is built on Java 1.7 (available at http://java.com/en/download/). It operates on the statistically significant peaks called from ChIP-Seq experiments. Any BED-like file is accepted by PAPST. An in-app input wizard allows users to create parsers for custom formatted files as long as the chromosome, start and end location, and tag count or score value are present. Gene

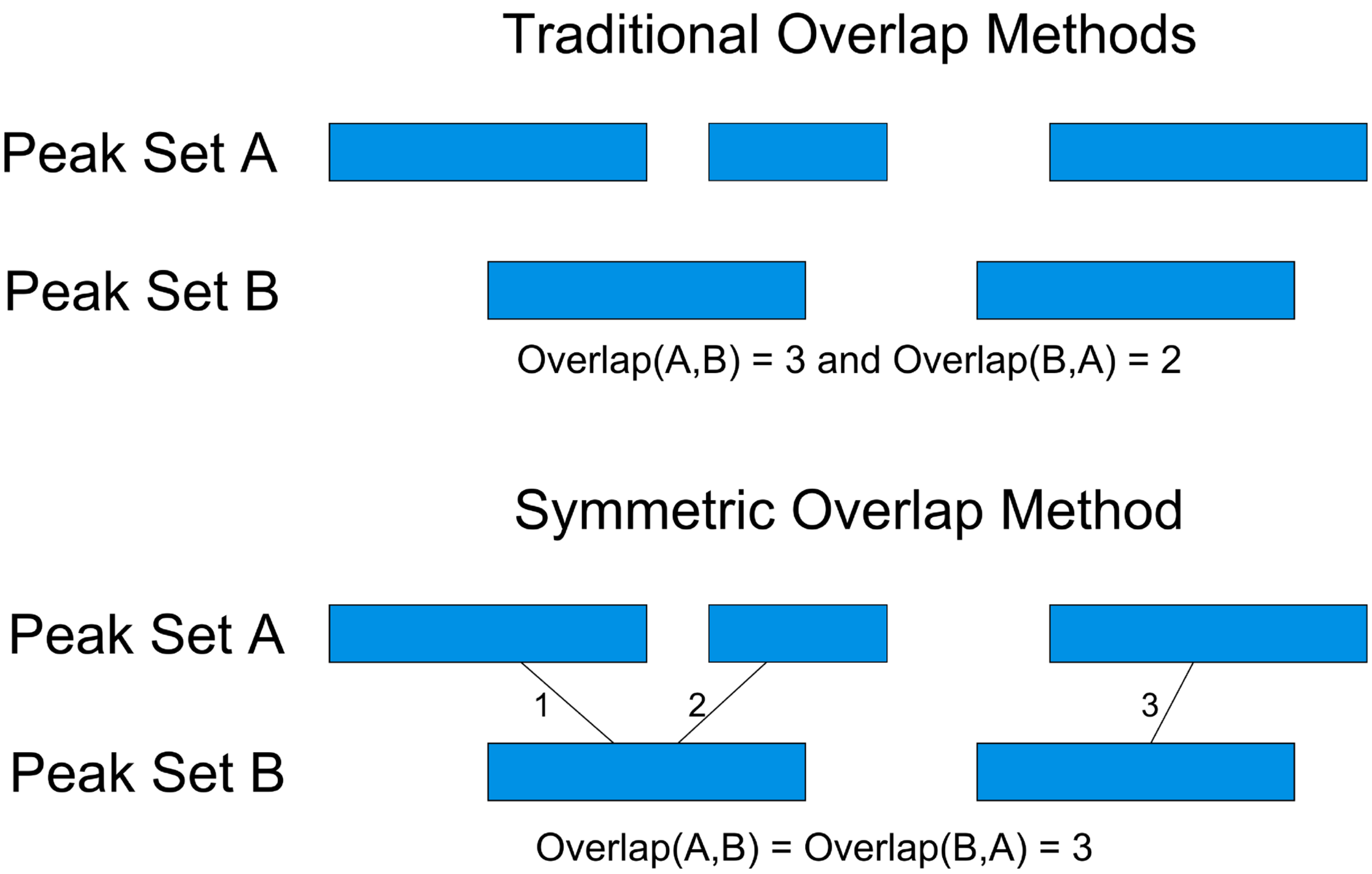

Traditional Overlap Methods
Peak Set A
Peak Set B
Overlap(A,B) = 3 and Overlap(B,A) = 2
Symmetric Overlap Method
Peak Set A
Peak Set B
1
2
3
Overlap(A,B) = Overlap(B,A) = 3

Fig 7. PAPST's Symmetric Overlap of Peak Regions. PAPST uses a novel overlap algorithm which counts the number of edges drawn between overlapping peaks. This method ensures that two sets, A and B, produce a symmetric overlap relationship where Overlap(A,B) = Overlap(B,A).

doi:10.1371/journal.pone.0127285.g007

and exon information is loaded into PAPST using RefGene files, and PAPST supports all genomes available from the UCSC Genome Browser (GB). The specific RefGene data used in this study were downloaded from GB site on October 13, 2014.

Internally, sets of peaks are represented by hashmaps of nested containment lists (NCL) [26]. Given a chromosome, a hashmap returns the NCL in constant time. The NCL answers range intersection queries in O(log n) time. For PAPST, this data structure makes operating on data from a large number of experiments fast and efficient.

Peak Overlap Calculation and Clustering

Most ChIP-Seq analysis programs calculate the overlap of peak regions in a non-symmetric way. Problems arise when a single peak region from one experiment overlaps multiple peaks in another experiment. For applications such as clustering, having a symmetric overlap relationship is desirable. PAPST uses a novel overlap algorithm that ensures the binary relationship between two peaks sets is symmetric (i.e. Overlap(A,B) = Overlap(B,A)). The algorithm accomplishes this by counting the number of edges that would be drawn to connect a peak with its overlapping peaks in the opposite set (Fig 7). When applied to multiple peak sets, PAPST generates a symmetric matrix of overlap values that is suitable for clustering. R was used to cluster 13 ChIP-Seq experiments based on their binding profiles using Ward's method. Before clustering, the data were normalized to have a mean of zero and unit variance. Details are available in the supplementary materials and tutorials.

Special term definitions

Gene-centric: ChIP-Seq data analysis on genomic regions defined by genes, such as promoters, exons, introns, and gene bodies. The analysis result usually is a set of genes with shared ChIP-Seq features, such as genes with their promoters co-occupied by two transcription factors

Peak-centric: ChIP-Seq data analysis on genomic regions defined by a set of ChIP-Seq peaks. The analysis result usually is a set of genomic regions with shared ChIP-Seq features, such as those regions co-occupied by a transcription factor and a histone modification

Typical enhancer: normal enhancer

Super enhancer: a cluster of adjacent typical enhancers

Supporting Information

S1 Appendix. The appendix contains supplementary methods that describe data collection, generation, preprocessing, and normalization methods.
(DOC)

S1 Table. Assignment of 6912 Oct4 ChIP-Seq Peaks to the nearest TSS. The data presented in this table provides an example of the gene assignment function of PAPST.
(CSV)

S2 Table. Search Results for Genes Bound by Oct4 in the 2000 bp Upstream Promoter. This file contains the table for all unique genes (isoforms excluded) that have a significant peak located in the promoter regions 2000 up stream of the TSS.
(CSV)

S3 Table. Search Results for Genes Bound by both Oct4 and K4me3 in the 2000 bp Upstream Promoter. This file contains the table for all unique genes (isoforms excluded) that have a significant peak for both Oct4 and K4me3 located in the promoter regions 2000 up stream of the TSS.
(CSV)

S4 Table. Non-redundant ESC Genes with Promoter Co-occupancy of Oct4, Sox2, Nanog, and H3K27ac. PAPST searched for gene promoters (+/- 5000 bp from TSS) that contained a binding profile that included all 3 transcription factors and the H3K27ac epigenetic mark. This table contains the exhaustive list of identified target genes.
(XLSX)

S5 Table. Assignment of Super Enhancers to their Nearest Gene. Using PAPST's gene assignment feature, all super enhancer regions were assigned to the nearest TSS. This table contains all the assigned genes' accessions, both the enhancer and gene coordinates and the distance from the enhancer to the TSS.
(XLSX)

S6 Table. Pairwise overlapping matrix generated by PAPST for 13 ESC transcription factors. PAPST was used to generate the all-pairs overlap data. Data generated from overlap comparisons in PAPST is symmetric making it suitable for clustering as well as other multidimensional analyses.
(CSV)

Acknowledgments

The authors would like to thank Dr. Kan Jiang for aiding in MacOS testing and Drs. Han-Yu Shih, Hossein Zare, and Vittorio Sartorelli for their valuable suggestions. We would also like to thank the anonymous reviewers for their constructive comments that have helped us to improve our manuscript. This study utilized the high-performance computational capabilities of the Helix Systems at the National Institutes of Health, Bethesda, MD (http://helix.nih.gov). This work was supported by the Intramural Research Program and the Office of Science and Technology of the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of Health.

Author Contributions

Conceived and designed the experiments: PWB HWS YK LW. Performed the experiments: PWB. Analyzed the data: PWB HWS SRB. Wrote the paper: HWS, PWB. Edited the manuscript: YK SRB LW JJO MIM RL. Supervised the project: HWS JJO MIM RL.

References

1. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012; 9(4):357–9. doi: 10.1038/nmeth.1923 PMID: 22388286; PubMed Central PMCID: PMC3322381.
2. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology. 2009; 10(3):R25. doi: 10.1186/gb-2009-10-3-r25 PMID: 19261174; PubMed Central PMCID: PMC2690996.
3. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25(14):1754–60. doi: 10.1093/bioinformatics/btp324 PMID: 19451168; PubMed Central PMCID: PMC2705234.
4. 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(11):1293–300. doi: 10.1038/nbt.1505 PMID: 18978777; PubMed Central PMCID: PMC2596672.
5. Jothi R, Cuddapah S, Barski A, Cui K, Zhao K. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. Nucleic acids research. 2008; 36(16):5221–31. doi: 10.1093/nar/gkn488 PMID: 18684996; PubMed Central PMCID: PMC2532738.
6. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, et al. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nature biotechnology. 2009; 27(1):66–75. doi: 10.1038/nbt.1518 PMID: 19122651; PubMed Central PMCID: PMC2924752.
7. 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):1952–8. doi: 10.1093/bioinformatics/btp340 PMID: 19505939; PubMed Central PMCID: PMC2732366.
8. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome biology. 2008; 9(9):R137. doi: 10.1186/gb-2008-9-9-r137 PMID: 18798982; PubMed Central PMCID: PMC2592715.
9. Ashoor H, Herault A, Kamoun A, Radvanyi F, Bajic VB, Barillot E, et al. HMCan: a method for detecting chromatin modifications in cancer samples using ChIP-seq data. Bioinformatics. 2013; 29(23):2979–86. doi: 10.1093/bioinformatics/btt524 PMID: 24021381; PubMed Central PMCID: PMC3834794.
10. Chojnowski K, Goryca K, Rubel T, Mikula M. jChIP: a graphical environment for exploratory ChIP-Seq data analysis. BMC research notes. 2014; 7:676. doi: 10.1186/1756-0500-7-676 PMID: 25260876; PubMed Central PMCID: PMC4189168.
11. Huang W, Loganantharaj R, Schroeder B, Fargo D, Li L. PAVIS: a tool for Peak Annotation and Visualization. Bioinformatics. 2013; 29(23):3097–9. doi: 10.1093/bioinformatics/btt520 PMID: 24008416; PubMed Central PMCID: PMC3834791.
12. Chen TW, Li HP, Lee CC, Gan RC, Huang PJ, Wu TH, et al. ChIPseek, a web-based analysis tool for ChIP data. BMC genomics. 2014; 15:539. doi: 10.1186/1471-2164-15-539 PMID: 24974934; PubMed Central PMCID: PMC4092222.
13. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular cell. 2010; 38(4):576–89. doi: 10.1016/j.molcel.2010.05.004 PMID: 20513432; PubMed Central PMCID: PMC2898526.

14. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26(6):841–2. doi: 10.1093/bioinformatics/btq033 PMID: 20110278; PubMed Central PMCID: PMC2832824.

15. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nature biotechnology. 2011; 29(1):24–6. doi: 10.1038/nbt.1754 PMID: 21221095; PubMed Central PMCID: PMC3346182.

16. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-Andre V, Sigova AA, et al. Super-enhancers in the control of cell identity and disease. Cell. 2013; 155(4):934–47. doi: 10.1016/j.cell.2013.09.053 PMID: 24119843; PubMed Central PMCID: PMC3841062.

17. Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell. 2008; 133(6):1106–17. doi: 10.1016/j.cell.2008.04.043 PMID: 18555785.

18. Kanno T, Kanno Y, LeRoy G, Campos E, Sun HW, Brooks SR, et al. BRD4 assists elongation of both coding and enhancer RNAs by interacting with acetylated histones. Nature structural & molecular biology. 2014; 21(12):1047–57. doi: 10.1038/nsmb.2912 PMID: 25383670.

19. Barlow JH, Faryabi RB, Callen E, Wong N, Malhowski A, Chen HT, et al. Identification of early replicating fragile sites that contribute to genome instability. Cell. 2013; 152(3):620–32. doi: 10.1016/j.cell.2013.01.006 PMID: 23352430; PubMed Central PMCID: PMC3629730.

20. Vahedi G, Takahashi H, Nakayamada S, Sun HW, Sartorelli V, Kanno Y, et al. STATs shape the active enhancer landscape of T cell populations. Cell. 2012; 151(5):981–93. doi: 10.1016/j.cell.2012.09.044 PMID: 23178119; PubMed Central PMCID: PMC3509201.

21. Wei L, Vahedi G, Sun HW, Watford WT, Takatori H, Ramos HL, et al. Discrete roles of STAT4 and STAT6 transcription factors in tuning epigenetic modifications and transcription during T helper cell differentiation. Immunity. 2010; 32(6):840–51. doi: 10.1016/j.immuni.2010.06.003 PMID: 20620946; PubMed Central PMCID: PMC2904651.

22. Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, Kirkup VM, et al. ENCODE data in the UCSC Genome Browser: year 5 update. Nucleic acids research. 2013; 41(Database issue):D56–63. doi: 10.1093/nar/gks1172 PMID: 23193274; PubMed Central PMCID: PMC3531152.

23. Hubisz MJ, Pollard KS, Siepel A. PHAST and RPHAST: phylogenetic analysis with space/time models. Briefings in bioinformatics. 2011; 12(1):41–51. doi: 10.1093/bib/bbq072 PMID: 21278375; PubMed Central PMCID: PMC3030812.

24. Hu G, Tang Q, Sharma S, Yu F, Escobar TM, Muljo SA, et al. Expression and regulation of intergenic long noncoding RNAs during T cell development and differentiation. Nature immunology. 2013; 14(11):1190–8. doi: 10.1038/ni.2712 PMID: 24056746; PubMed Central PMCID: PMC3805781.

25. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, et al. The UCSC Genome Browser Database: update 2006. Nucleic acids research. 2006; 34(Database issue):D590–8. doi: 10.1093/nar/gkj144 PMID: 16381938; PubMed Central PMCID: PMC1347506.

26. Alekseyenko AV, Lee CJ. Nested Containment List (NCList): a new algorithm for accelerating interval query of genome alignment and interval databases. Bioinformatics. 2007; 23(11):1386–93. doi: 10.1093/bioinformatics/btl647 PMID: 17234640.