Bioinformatics, 31(19), 2015, 3092–3098
doi: 10.1093/bioinformatics/btv336
Advance Access Publication Date: 1 June 2015
Original Paper

Genome analysis

Mango: a bias-correcting ChIA-PET analysis pipeline

Douglas H. Phanstiel[1], **Alan P. Boyle**[2], **Nastaran Heidari**[1] **and Michael P. Snyder**[1,*]

[1]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305 and [2]Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on January 19, 2015; revised on May 11, 2015; accepted on May 26, 2015

Abstract

Motivation: Chromatin Interaction Analysis by Paired-End Tag sequencing (ChIA-PET) is an established method for detecting genome-wide looping interactions at high resolution. Current ChIA-PET analysis software packages either fail to correct for non-specific interactions due to genomic proximity or only address a fraction of the steps required for data processing. We present Mango, a complete ChIA-PET data analysis pipeline that provides statistical confidence estimates for interactions and corrects for major sources of bias including differential peak enrichment and genomic proximity.

Results: Comparison to the existing software packages, ChIA-PET Tool and ChiaSig revealed that Mango interactions exhibit much better agreement with high-resolution Hi-C data. Importantly, Mango executes all steps required for processing ChIA-PET datasets, whereas ChiaSig only completes 20% of the required steps. Application of Mango to multiple available ChIA-PET datasets permitted the independent rediscovery of known trends in chromatin loops including enrichment of CTCF, RAD21, SMC3 and ZNF143 at the anchor regions of interactions and strong bias for convergent CTCF motifs.

Availability and implementation: Mango is open source and distributed through github at https://github.com/dphansti/mango.

Contact: mpsnyder@standford.edu

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

1 Introduction

Increasing evidence suggests that the three-dimensional structure of mammalian chromatin plays an important role in regulating gene expression and biological function (Göndör and Ohlsson, 2009; Li *et al.*, 2012; Schoenfelder *et al.*, 2010). To study chromatin structure, multiple high-throughput sequencing-based assays have been developed including 3C, 4C, 5C, Hi-C and, the newest addition, Chromatin Interaction Analysis by Paired End Tag sequencing (ChIA-PET) (Dekker, 2002; Dostie *et al.*, 2006; Fullwood *et al.*, 2009; Lieberman-Aiden *et al.*, 2009; Simonis *et al.*, 2006). Though the methods differ in scope and resolution, they all share the same basic steps including cross-linking native chromatin, ligating interacting fragments and sequencing chimeric DNA fragments to assess interaction frequency. Informatic analysis of the resulting data then entails identifying which pairs of loci exhibit more interactions than expected by random chance.

It has been well established by 3C, FISH and Hi-C studies that pairs of genomic loci exhibit random-polymer-like behavior in which genomic interaction frequencies decrease as a function of genomic distance (Supplementary Fig. S1A and B) (Dekker, 2002; Lieberman-Aiden *et al.*, 2009). Therefore, it is critical that software designed to detect interactions between pairs of loci must consider and accurately model the expected interaction frequency of pairs of loci given linear genomic distance. Indeed, most if not all available software packages to identify interactions from 5C and Hi-C datasets do consider genomic distance in the null model (Ay *et al.*, 2014; Duan *et al.*, 2010; Sanyal *et al.*, 2012). Furthermore, ChIA-PET has the unique property of specifically enriching for loci bound by a specific protein. As these regions are bound at different affinities throughout the genome, the depth of coverage at interacting peaks must be considered in any analysis paradigm.

Several analysis methods exist to process ChIA-PET data. However, these methods either fail to address the major sources of bias or require the user to write a large portion of the analysis code themselves rendering them unusable for many wet lab biologists.

ChIA-PET Tool (CPT) was the first software to address the unique problems associated with ChIA-PET data and has established an effective workflow for data processing (Li *et al.*, 2010). The authors of CPT employ both random sampling and statistical methods to filter out noise due to random ligations that occur in solution rather than *in vivo*. However, CPT fails to address the most common source of bias in 3 C-based techniques, non-specific interactions due to linear genomic distance, i.e. the random-polymer effect. CPT uses the hypergeometric distribution to assess significance of interactions. This model assumes that any two genomic loci are equally likely to be linked by a PET regardless of genomic distance. Therefore, in addition to detecting true looping interactions, this approach is likely to detect a large number of false positives that exhibit non-random interaction frequencies due to genomic proximity rather than genomic looping. Indeed, a recent study revealed that interactions determined by CPT showed poor agreement with high-resolution Hi-C datasets. In addition, CPT is extremely difficult to install as it requires a very specific OS configuration including a complex array of programming languages and environments including C, perl, python, R, MySQL, Apache web server and PHP.

Recently, Paulsen *et al* described a novel method to process ChIA-PET data that does consider genomic distance in the null model. This method, ChiaSig, uses the non-central hypergeometric to identify statistically significant interactions {Paulsen:2014bn}. As expected, this method identifies far fewer interactions than CPT. While this approach yields accurate interactions, the software only executes the very final step in ChIA-PET data analysis, interaction scoring. Therefore, users must write their own software to find and remove linker sequences, align PETs, remove duplicates, call peaks, group PETs into interactions and determine the lower bound cutoff for PET distances. As such, this software is only useful to researchers with significant programming skills. Other software packages have been described but either are not publicly available or have similar limitations to CPT and ChiaSig (Niu and Lin, 2014; Reeder and Gifford, 2013).

To address these shortcomings, we introduce Mango an open source ChIA-PET data analysis pipeline. Mango models the likelihood of interactions between genomic loci as a function of both distance and peak depth and uses this model to assign statistical confidence to interactions. This software is simple to install, requires only fastq files as input and completes all steps required to analyze ChIA-PET datasets with only a single command.

To evaluate the applicability of Mango, we compared the interactions determined by Mango, ChiaSig and CPT for three publicly available ChIA-PET datasets (Heidari *et al.*, 2014). We demonstrate that interactions detected by Mango show better agreement with recently published high-resolution Hi-C data and independently recapitulate several recently discovered characteristics of long-range chromatin interactions. First, the anchor regions of interactions detected by Mango were strongly enriched for the proteins CTCF, RAD21, SMC3 and ZNF143 with between 92 and 98% of anchor regions overlapping a CTCF binding site. Second, the majority (73 to 94%) of interactions linked two loci that harbored CTCF motifs that had a convergent orientation.

2 Methods

The general structure of the Mango workflow is divided into five steps that can be executed all at once or one at time (Supplementary Fig. S2). Step 1 involves finding and removing linker sequences from reads as well as filtering reads based on the combination of linkers observed. Only PETs that have the same linker sequences at both ends are kept for further processing. In step 2, reads are aligned to the genome using the widely used Bowtie software suite (Ben Langmead *et al.*, 2009). Step 3 removes reads that may be due to polymerase chain reaction duplication and organizes the data for peak calling and interaction analysis. Step 4 uses MACS2 to call binding peaks, which are subsequently used as anchor regions for the detection of interactions in step 5 (Zhang *et al.*, 2008). In step 5, statistical confidence estimates are assigned to interactions based on comparison to a model that considers both genomic distance and the read depth of each peak. The resulting *P* values are corrected to account for multiple hypothesis testing and filtered to a user defined false discovery rate (FDR).

2.1 Calculating statistical confidence estimates of interactions

Calculating the statistical confidence estimates of interactions in step 5 begins with two stages of filtering. Since previous studies have shown that most interactions, especially functional interactions, take place within megabase scale topological domains (TADs: Supplementary Fig. S1A and B), Mango removes inter-chromosomal PETs and PETs with a distance greater than a user-defined value (1 Mb for this article) (Dixon *et al.*, 2012; Phillips-Cremins *et al.*, 2013; Sanyal *et al.*, 2012). This both reduces computational burden and increases statistical power by minimizing the effect of multiple hypothesis testing. The second step of filtering involves setting a lower bound distance cutoff for PETs to eliminate bias introduced by PETs that result from self-circularization rather than inter-ligation of interacting loci. Mango uses the orientation of the strands on either end of a PET to estimate the percent of reads due to self-circularization as a function of distance and allows the user to define an acceptable cutoff (5% for this article; see Supplementary Methods and Fig. S1C). The remaining ‘mid-range’ PETs are used to determine interactions.

PETs are next grouped into putative interactions. Since all ligated fragments in a ChIA-PET experiment must be captured during the chromatin immunoprecipitation step, there should be enrichment of reads at both ends of each interaction. Therefore, Mango uses the peaks detected in step 4 or a user supplied Browser Extensible Data file of genomic regions, as the anchor regions for putative interactions.

Mango then models the probability of observing a single PET linking two loci as a function of their genomic distance of separation and the product of their read depths (see Supplementary Methods and Fig. S1C–F). These models are built empirically for each data and are thus robust to differences in antibodies, sequencing depth and experimental variation. A comparison of models built for three different datasets is shown in Supplementary Figure S3. Application of Bayes theorem shows that the probability of observing a PET linking two loci separated by distance *L* and characterized by joint peak depth *D* is a product of two other probabilities:

(i) the probability of observing a PET with distance L connecting regions with depth D and (ii) the probability of connecting these two specific regions given length L and depth D (for details see Supplementary Methods). That is:

$$P(I) = P(L,D) \times P(I|L,D) \tag{1}$$

where $P(I)$ is the probability of observing a PET connecting two specific loci, $P(L,D)$ is the probability of observing a PET with a given L and D and $P(I|L,D)$ is the probability of the PET linking these two specific loci given L and D.

L and D are independent (Supplementary Fig. S4) and therefore this equation, can be represented by:

$$P(I) = \frac{P(I|L) \times P(I|D)}{P(C|L) \times P(C|D) \times C_T} \tag{2}$$

where $P(I|L)$ represents the probability of observing a PET linking loci with distance L, $P(I|D)$ represents the probability of observing a PET linking loci with depth D, $P(C|L)$ represents the probability of observing a pair of loci with distance L (regardless of whether any PETs link the two loci), $P(C|D)$ represents the probability of observing a pair of loci with depth D (regardless of whether any PETs link the two loci) and C_T is the total number of pairwise combinations of loci.

All the terms on the right side of this equation can be modeled from the data itself (Supplementary Fig. S1D and E) allowing for the determination of $P(I)$ for each pair of loci (Supplementary Fig. S1F). According to the binomial distribution, this probability can be used to calculate the P value of observing exactly k PETs using the following equation:

$$P(K = k) = \binom{N}{k} P(I)^k (1 - P(I))^{N-k} \tag{3}$$

where N is the total number of mid-range PETs in the experiment. The P value of observing k or more PETs can be calculated as

$$P(K \geq k) = \sum_{i=K}^{N} P(K = i) \tag{4}$$

Finally, P values for all possible pairs of interacting loci, not just the ones connected by PETs, are corrected for multiple hypothesis testing using the Benjamini–Hochberg method (Benjamini and Hochberg, 1995) and filtered to a user-defined FDR.

3 Results

3.1 Comparison to existing ChIA-PET analysis methods

To evaluate Mango, we processed three publicly available ChIA-PET datasets with Mango and two publicly available software packages, CPT and ChiaSig (Li *et al.*, 2010; Paulsen *et al.*, 2014). The datasets targeted histone H3K4Me3, POLR2A and RAD21 in a human myelogenous leukemia cell line (K562) (Heidari *et al.*, 2014). ChiaSig only performs the very final step in ChIA-PET data analysis (Table 1). Therefore, we used Mango to perform linker parsing, read alignment, duplicate removal, peak calling and PET grouping and applied ChiaSig to resulting data. The exact commands required to replicate the Mango results are available in the Supplementary Information. Application of Mango to the H3K4Me3, POLR2A and RAD21 data resulted in 1259, 4040 and 9168 significant interactions at an FDR of 0.05, respectively (Fig. 1A, Supplementary Fig. S5). ChiaSig produced 1360, 2345 and 5869 interactions. CPT resulted in far more significant interactions than either Mango or ChiaSig at the same FDR (66 787, 16 961 and 50 725). This large decrease in interactions compared with CPT is expected since virtually all pairs of local intra chromosomal loci should interact non-randomly according to the hypergeometric model used by CPT. Mango and ChiaSig, in contrast, correct for non-specific interaction frequencies due to genomic distance and therefore report only those interactions with interaction frequencies that significantly exceed those expected for a given distance. Despite this sharp decrease in identified interactions, interactions detected by Mango and ChiaSig show better agreement with alternative methodologies for detecting chromatin loops (*vide infra*).

We next compared the sizes of the interactions detected by determining the number of base pairs between the interacting anchor regions (Fig. 1B). Since Mango and ChiaSig account for random interactions due to genomic proximity, the median interaction distances detected were significantly larger than the interactions detected by CPT (Mango: 100 507, 117 861 and 177 846 bp; ChiaSig: 106 213, 140 738 and 158 898 bp; CPT: 1371, 120 and 1060 bp; Wilcoxon test, P value $< 10^{-16}$). The minimum interaction distances for CPT, ChiaSig and Mango were 1, 3398 and 13 425 bp. This strong skew toward short interaction distances in CPT results helps explain the poor overlap observed in Figure 1A as 83–98% of CPT interactions were shorter than the shortest Mango interaction (excluding intra-chromosomal interactions longer than 1 Mb and all inter-chromosomal interactions).

Since pairs of loci interact non-specifically as a function of genomic distance, we reasoned that short-range interactions should require more PETs to achieve significance compared with long-range interactions. To test this, we binned significant interactions from the RAD21 dataset by genomic distance ranging from 25 kb to 1 Mb and plotted the percent of interactions in each bin that were supported by PETs ranging from 2 to ≥ 20 (Fig. 1C–E). CPT interactions were fairly uniform with respect to distance. Across all bins, the majority of significant interactions were supported by only two PETs (Fig. 1C). In contrast, Mango and ChiaSig interactions showed the expected trend in which shorter distance interactions required more PETs to achieve a significant P value (Fig. 1E and F). Conducting the same analysis for interactions detected by 5C revealed a trend very similar to that observed in the Mango and ChiaSig datasets (Supplementary Fig. S6) (Sanyal *et al.*, 2012).

3.2 Comparison to Hi-C

To determine the accuracy and biological relevance of interactions detected by Mango, we intersected our results with deeply sequenced Hi-C data. Rao *et al.* (2014) recently published deeply sequenced high-resolution Hi-C data for multiple cell lines including

Table 1. Comparison of ChIA-PET software packages

Software	Linker parsing	Read alignment	Duplicate removal	Peak calling	Self-ligation cutoff	PET grouping	Statistical estimate
CPT	✓	✓	✓	✓	✓	✓	✓
ChiaSig	✗	✗	✗	✗	✗	✗	✓
Mango	✓	✓	✓	✓	✓	✓	✓

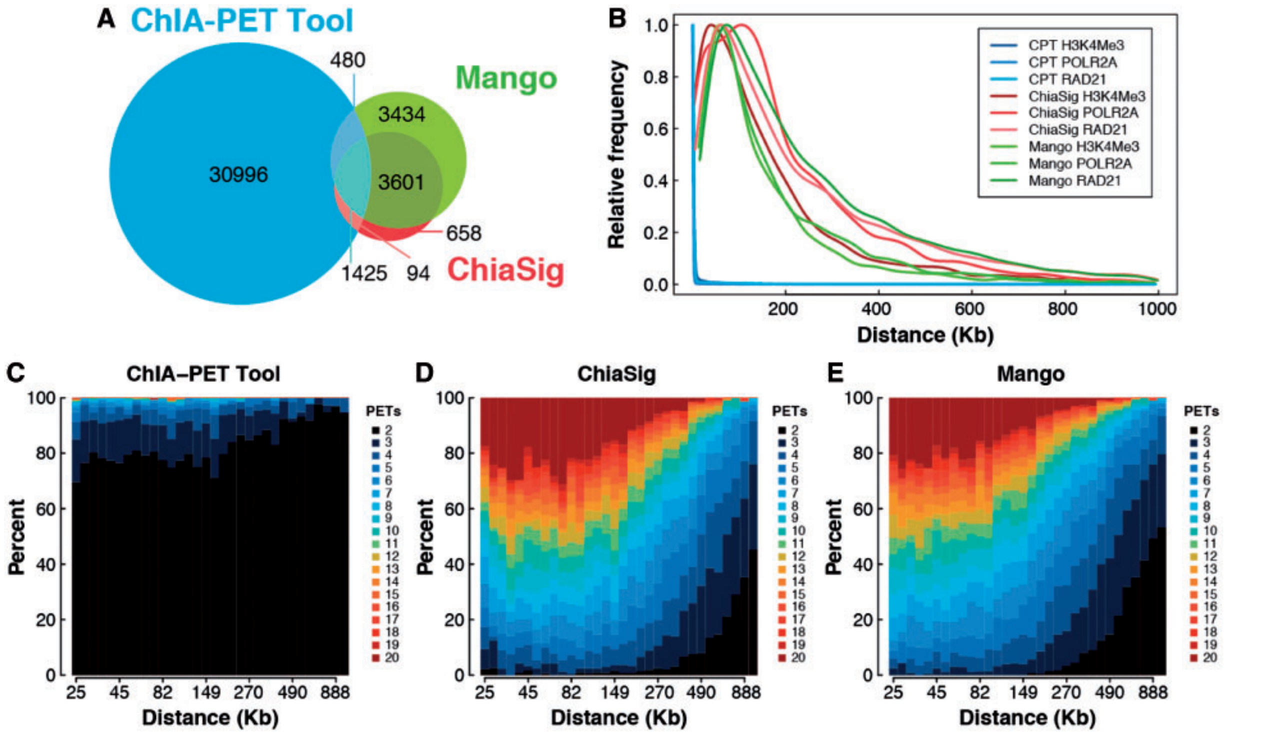

Fig. 1. Comparison of interactions reported by CPT, ChiaSig and Mango. (**A**) Venn diagram depicting overlap between interactions reported by CPT, ChiaSig and Mango for the RAD21 dataset. (**B**) Density plots of interaction lengths for each dataset. (**D–F**) Barplots depicting the percentages of significant interactions supported by PETs ranging from 2 to ≥20 for interactions determined by CPT, ChiaSig and Mango for the RAD21 dataset

K562. We plotted normalized Hi-C contact matrices and interaction calls for each of the three algorithms (Fig. 2A).

Visual inspection of ChIA-PET interactions overlaid on top of Hi-C contact matrices revealed a strong agreement between Hi-C interactions and ChIA-PET interaction calls made by both ChiaSig and Mango. A contact matrix for a 1 kb region on chromosome generated by Rao *et al.* is shown in Figure 2A. The resolution of the Hi-C dataset permitted clear identification of genome-wide looping interactions. Loops detected by Hi-C are characterized by islands of high interaction frequency, which are surrounded on all sides by lower interaction frequencies. Red squares mark interaction calls made by each of the three algorithms. Loops called by Rao *et al.* are shown in Supplementary Figure S7. Interactions detected by ChiaSig and Mango exhibited good agreement with Hi-C contact matrices as shown in Figure 2A. In contrast, although CPT interactions do overlap some interactions that were detected by Mango, ChiaSig or Rao *et al.*, the majority linked extremely close genomic loci.

To quantify how well the Hi-C data supported the ChIA-PET interactions, we generated aggregate peak analysis (APA) plots as described by Rao *et al.* (2014) (Fig. 2B). These plots aggregate the signal in pixels surrounding anchor regions across all interactions. Rao *et al.* demonstrated that even low-sequencing depth Hi-C datasets can be used to evaluate the quality of interaction calls. Moreover, this method does not depend heavily on the methods used to process the Hi-C data as it uses contact matrices rather than interaction calls to build the plots. To generate APA plots, interaction counts are summed for all pairs of loci in 5-kb bins spanning 50 kb up- and downstream of both interacting loci. These values are plotted as a 21 × 21 pixel matrix colored by interaction count. Since true looping interactions should interact more frequently than intervening pairs of loci, they should be characterized by a dark center pixel. To generate these plots, only interactions linking regions separated by greater than 150 kb and less than 1 Mb were used. The level of support for each set of interactions can be quantified by calculating an APA score, which is simply the value of the center pixel divided by the mean of pixels 15–30 kb downstream of the upstream loci and 15–30 kb upstream of the downstream loci. Scores of 1 indicate no evidence of a loop. Higher scores indicate stronger evidence. For all three datasets, APA scores for mango and ChiaSig are similar and both are greater than APA scores for CPT (Supplementary Fig. S8). However, the different algorithms produced vastly different quantities of significant interactions. Therefore, two equivalent interaction-ranking algorithms could exhibit different APA scores simply based on the cutoff applied to control FDR. To address this, we developed an extension of APA methods that disentangles interaction-ranking from these thresholding effects.

The quality of interaction detection between the three methods was assessed using cumulative APA (CAPA) plots. To generate CAPA plots, we ranked interactions by *P* values and calculated APA score in a cumulative fashion adding 100 interactions at a time (Fig. 2C). CAPA plots reveal that Mango interactions are better supported by Hi-C matrices than either ChiaSig or CPT across all three datasets. Interestingly, sets of interactions deemed significant by either ChiaSig or Mango had very similar APA scores for all three datasets but since Mango's interaction ranking is superior to ChiaSig, Mango reported nearly twice as many interactions as ChiaSig for both the POL2 and RAD21 datasets. Both Mango and ChiaSig provided far better interaction ranking capabilities compared with CPT.

3.3 Mango interactions recapitulate known characteristics of DNA loops

To assess the biological relevance of Mango interaction calls, we determined if the interaction calls made by Mango could

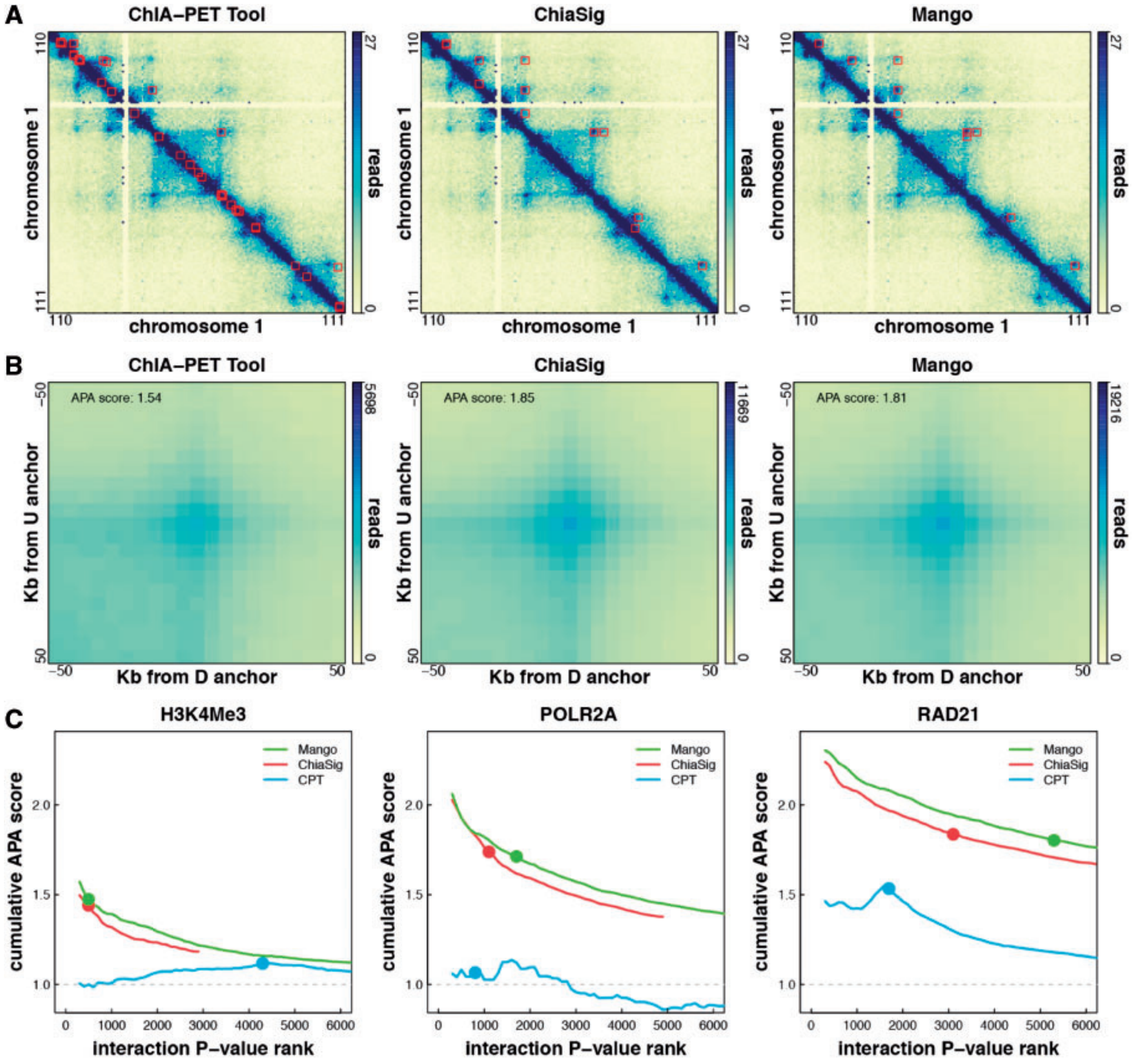

Fig. 2. Intersection of CPT, ChiaSig and Mango interactions with Hi-C contact matrices. (**A**) 5 kb resolution normalized Hi-C contact matrix for K562 cells generated by Rao *et al.* Red squares depict interactions determined by CPT, ChiaSig and Mango for the RAD21 dataset. (**B**) APA plots depicting normalized Hi-C counts for all pairs of loci ±50 kb summed across all interactions determined by CPT, ChiaSig and Mango for the RAD21 dataset. (**C**) CAPA plots depicting cumulative APA scores as a function of interaction rank. Points represent significance thresholds determined by each application for reported interactions

recapitulate previous findings regarding 3D chromatin structure. We and others have shown that anchor regions of 3D loops are very strongly enriched for CTCF, members of the cohesion complex and ZNF143. Intersecting these interactions with ChIP-Seq datasets available from the ENCODE consortium revealed that these proteins were indeed enriched at anchor regions detected by Mango (Fig. 3A) (ENCODE Project Consortium, 2012).

Anchor regions identified from H3K4Me3 and POLR2A datasets showed enrichment for virtually all TFs. This finding is consistent with Hi-C interaction calls from Rao *et al.* (2014). As we have previously shown, interacting loci are enriched for HOT regions, regions bound by many TFs, which would explain this enrichment for all TFs (Heidari *et al.*, 2014). Removing interacting loci which overlap HOT regions reveals only four strongly enriched chromatin-associated proteins: CTCF, RAD21, SMC3 and ZNF143 (Fig. 3B). These findings indicate that virtually all (between 97 and 99%) of interacting loci overlap a HOT region, a CTCF binding site or both (Supplementary Fig. S9). Further analysis of these two types of interacting loci revealed unique characteristics of each subtype such as differences in interaction distance and additive strength of interactions (Supplementary Figs. S9 and S10).

Finally, we asked if interactions detected by Mango were characterized by an orientation bias in CTCF motifs. Rao *et al.* revealed that the majority of interactions contained CTCF motifs that were oriented toward the intervening region between the two loci. Indeed, all three datasets exhibited a strong bias for inward pointing motifs and a virtual absence of outward pointing motifs (Fig. 3C). These strong trends and overwhelming agreement with high-resolution Hi-C data provides powerful validation of Mango results as true looping chromatin interactions.

3.4 Implementation and availability

Mango was designed with both accuracy and ease of use in mind. Mango runs on both Linux and OSX operating systems. Installation of Mango requires only a single command and its dependencies are limited to four very commonly used software suites: R, bedtools, bowtie and MACS2 (Ben Langmead *et al.*, 2009; Quinlan and Hall, 2010;

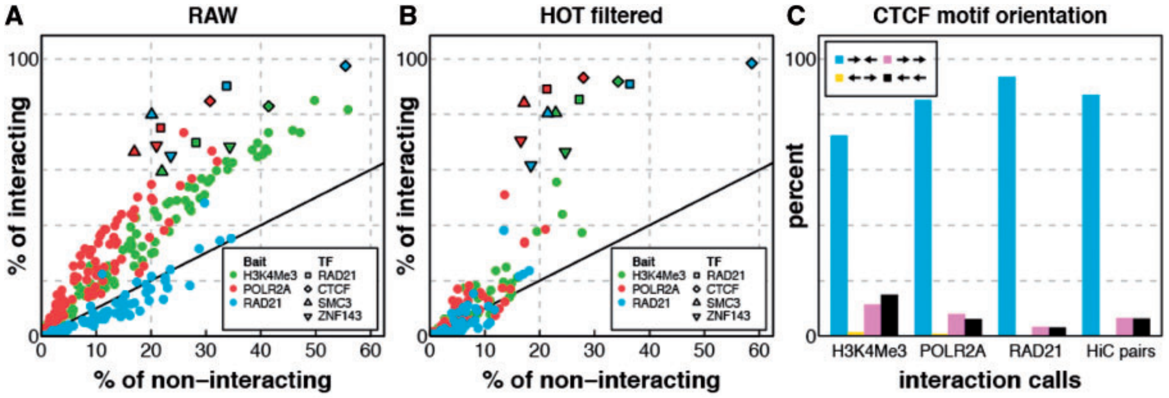

Fig. 3. Biological basis of chromatin interactions. (**A**) The percent of TF binding sites in sets of anchor regions that are observed (*y*-axis) versus expected (*x*-axis) in the ChIA-PET interaction calls. The black line represents a slope of 1 which indicates no enrichment. The cohesion complex members are highlighted. (**B**) The same plots as shown in (A) after removal anchor regions that overlap a HOT region. (**C**) A barplot depicting the percentage of interactions with CTCF motifs in each of the four possible orientations

R Core Team, 2013; Zhang *et al.*, 2008). Mango is open source, distributed through github, and can be downloaded at https://github.com/dphansti/mango.

4 Discussion

Here we describe Mango a ChIA-PET data analysis pipeline that corrects for non-specific interactions as a function of genomic proximity and peak depth. We demonstrate that Mango exhibits increased accuracy compared with both CPT, the only existing ChIA-PET analysis pipeline, and ChiaSig, a software package that provides statistical confidence estimates for ChIA-PET interactions. Application of Mango to multiple ChIA-PET datasets allowed for the independent replication of findings regarding the nature of 3D chromatin loops including strong enrichment for CTCF binding sites with inward oriented motifs.

In addition to improved accuracy, one of the key benefits of Mango is usability. Mango was designed to be usable by all researchers even those with minimal computer competency. Mango is easily installed and completes all steps from fastq to interactions with a single command. It relies on only four widely used and easily installed software packages. In contrast, CPT requires a very specific OS configuration including a complex array of programming languages and environments including C, perl, python, R, MySQL, Apache web server and PHP and is accompanied by a seven page installation guide. ChiaSig can be installed easily yet only performs a single step required for the analysis of ChIA-PET data. Users are therefore required to write their own code to perform the majority of processing steps including linker parsing, read alignment, duplicate removal, peak calling and distance filtering.

The software presented here improves upon methods used in our previous work (Heidari *et al.*, 2014). Most notably Mango replaces the computationally expensive distance matched rewiring method with a simple and robust Bayesian approach.

Because of improvements in ease of use and accuracy, Mango will drastically improve our ability to uncover the characteristics and function of 3D chromatin structure through the analysis of ChIA-PET datasets.

Funding

This project was supported by NIH grant U54HG006996 to M.P.S. and R00HG007356 to A.P.B. D.H.P. is a Damon Runyon fellow supported by the Damon Runyon Cancer Research Foundation (DRG-2122-12).

Conflict of Interest: M.P.S. is a cofounder and scientific advisory board (SAB) member of Personalis. MPS is on the SAB of Genapsys.

References

Ay,F. *et al.* (2014) Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res.*, **24**, 999–1011.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B Methodol.*, 289–300.

Dekker,J. (2002) Capturing chromosome conformation. *Science*, **295**, 1306–1311.

Dixon,J.R. *et al.* (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.

Dostie,J. *et al.* (2006) Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.*, **16**, 1299–1309.

Duan,Z. *et al.* (2010) A three-dimensional model of the yeast genome. *Nature*, **465**, 363–367.

ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Fullwood,M.J. *et al.* (2009) An oestrogen-receptor-α-bound human chromatin interactome. *Nature*, **462**, 58–64.

Göndör,A. and Ohlsson,R. (2009) Chromosome crosstalk in three dimensions. *Nature*, **461**, 212–217.

Heidari,N. *et al.* (2014) Genome-wide map of regulatory interactions in the human genome. *Genome Res.*, **24**, 1905–1917.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Li,G. *et al.* (2010) ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome Biol.*, **11**, R22.

Li,G. *et al.* (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, **148**, 84–98.

Lieberman-Aiden,E. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.

Niu,L. and Lin,S. (2014) A Bayesian mixture model for chromatin interaction data. *Stat. Appl. Genet. Mol. Biol.*, **14**, 53-64.

Paulsen,J. *et al.* (2014) A statistical model of ChIA-PET data for accurate detection of chromatin 3D interactions. *Nucleic Acids Res.*, **42**, e143.

Phillips-Cremins,J.E. *et al.* (2013) Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*, **153**, 1281–1295.

Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. **26**, 841–842.

R Core Team (2013) *R: A Language and Environment for Statistical Computing*, http://www.R-project.org.

Rao,S.S.P. *et al*. (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.

Reeder,C. and Gifford,D. (2013) High resolution modeling of chromatin interactions. In: Deng,M. *et al*. (eds) *Research in Computational Molecular Biology, Lecture Notes in Computer Science*. Springer Berlin Heidelberg, Berlin, Germany, pp. 186–198.

Sanyal,A. *et al*. (2012) The long-range interaction landscape of gene promoters. *Nature*, **489**, 109–113.

Schoenfelder,S. *et al*. (2010) The transcriptional interactome: gene expression in 3D. *Curr. Opin. Genet. Dev.*, **20**, 127–133.

Simonis,M. *et al*. (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet*., **38**, 1348–1354.

Zhang,Y. *et al*. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol*., **9**, R137.