Published online 31 March 2020

NAR Genomics and Bioinformatics, 2020, Vol. 2, No. 2 **1**
doi: 10.1093/nargab/lqaa022

HIPPIE2: a method for fine-scale identification of physically interacting chromatin regions

Pavel P. Kuksa[1,†], Alexandre Amlie-Wolf[2,†], Yih-Chii Hwang[3,†], Otto Valladares[1], Brian D. Gregory[2,4] and Li-San Wang[1,2,*]

[1]Penn Neurodegeneration Genomics Center, Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA, [2]Genomics and Computational Biology Graduate Group, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA, [3]DNAnexus, Inc., Mountain View, CA 94040, USA and [4]Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA

Received June 19, 2019; Revised January 15, 2020; Editorial Decision March 10, 2020; Accepted March 16, 2020

ABSTRACT

Most regulatory chromatin interactions are mediated by various transcription factors (TFs) and involve physically interacting elements such as enhancers, insulators or promoters. To map these elements and interactions at a fine scale, we developed HIPPIE2 that analyzes raw reads from high-throughput chromosome conformation (Hi-C) experiments to identify precise loci of DNA physically interacting regions (PIRs). Unlike standard genome binning approaches (e.g. 10-kb to 1-Mb bins), HIPPIE2 dynamically infers the physical locations of PIRs using the distribution of restriction sites to increase analysis precision and resolution. We applied HIPPIE2 to *in situ* Hi-C datasets across six human cell lines (GM12878, IMR90, K562, HMEC, HUVEC, NHEK) with matched ENCODE/Roadmap functional genomic data. HIPPIE2 detected 1042 738 distinct PIRs, with high resolution (average PIR length of 1006 bp) and high reproducibility (92.3% in GM12878). PIRs are enriched for epigenetic marks (H3K27ac, H3K4me1) and open chromatin, suggesting active regulatory roles. HIPPIE2 identified 2.8 million significant PIR–PIR interactions, 27.2% of which were enriched for TF binding sites. 50 608 interactions were enhancer–promoter interactions and were enriched for 33 TFs, including known DNA looping/long-range mediators. These findings demonstrate that the novel dynamic approach of HIPPIE2 (https://bitbucket.com/wanglab-upenn/HIPPIE2) enables the characterization of chromatin and regulatory interactions with high resolution and reproducibility.

INTRODUCTION

Enhancers are non-coding DNA elements that regulate gene expression by recruiting transcription factors (TFs) that in turn mediate physical interactions with the promoters of their target genes to increase transcription of those genes. The genome-wide relationship between enhancers and their target genes depends on the 3D DNA looping associated with enhancer–promoter interactions. To capture genome-wide chromatin interactions in high-throughput chromosome conformation (Hi-C) (1), DNA physically interacting regions (PIRs) and their binding proteins are cross-linked, followed by restriction enzyme cleavage and proximity ligation of the interacting DNA fragments to localize and capture pairs of interacting DNA fragments. These ligated DNA fragments are then sequenced to identify the chromatin interaction map genome-wide. Higher resolution in localizing interacting DNA fragments has been achieved by using a restriction enzyme with more frequent sites throughout the genome (e.g., MboI, a 4-cutter with a 4-bp motif, instead of a 6-cutter such as HindIII or NcoI) and by performing the DNA–DNA proximity ligation in intact nuclei to generate denser Hi-C contact matrices (2).

Previous methods for analyzing Hi-C data (1–14) have implemented a binning-based scheme for identifying interacting genomic regions, where reads are aggregated into equally sized bins by genome coordinates and interacting regions are identified as pairs of bins with significant enrichments of reads using statistical models accounting for biases (e.g. negative correlation between linear genomic distance, number of reads and mappability) of the individual bins. While binning is effective at delineating large-scale chromatin structure, it does not capture specific DNA PIRs. The methods of Jin *et al*. and Hwang *et al*. (9,15,16) have shown that it is possible to study interactions at the level of restriction fragments [i.e. the DNA region between two consecu-

*To whom correspondence should be addressed. Tel: +1 215 746 7015; Fax: +1 215 573 3111; Email: lswang@pennmedicine.upenn.edu
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

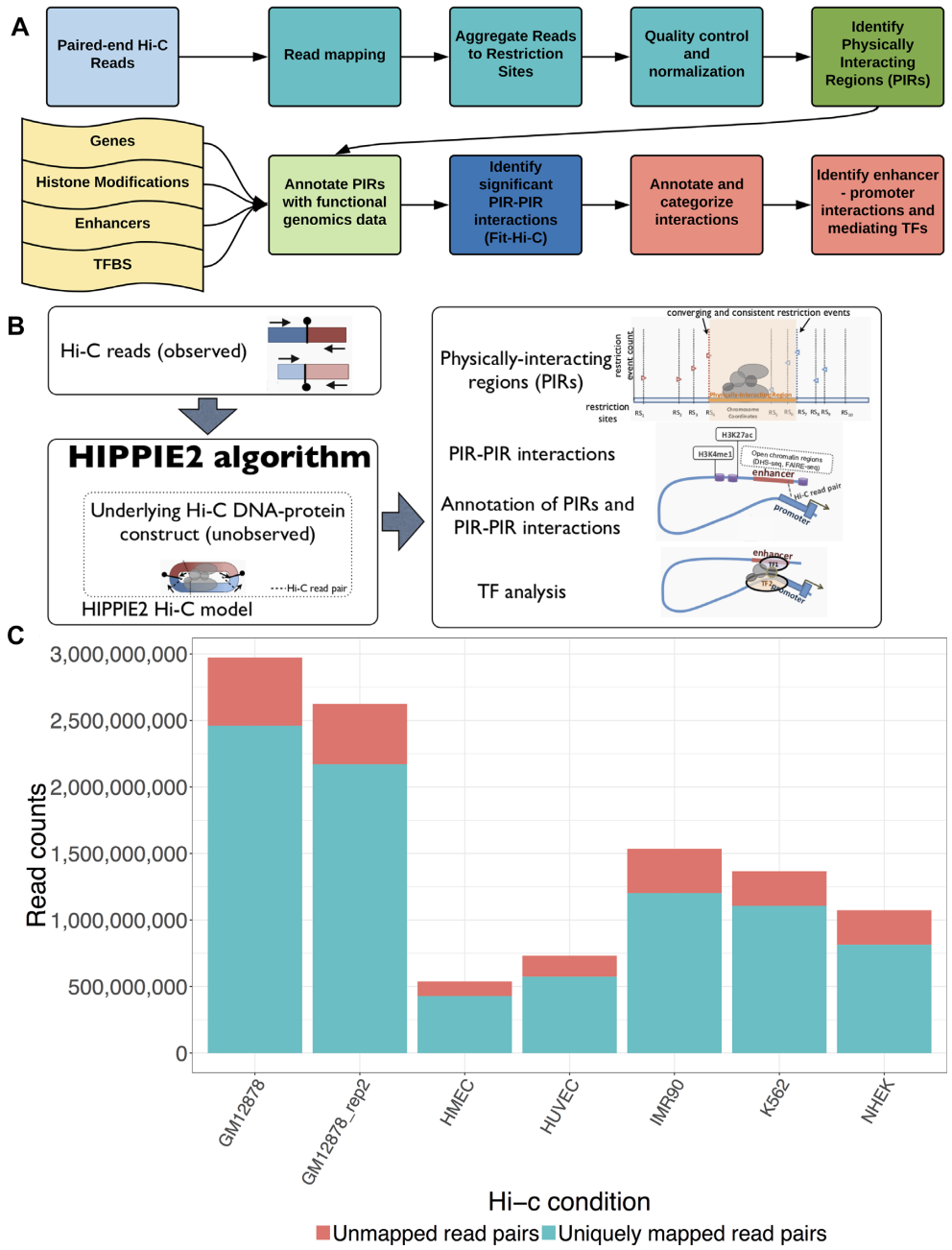

Figure 1. Description of HIPPIE2 pipeline and mapping statistics. (**A**) Detailed processing pipeline of HIPPIE2. (**B**) Overview of HIPPIE2 algorithm. (**C**) Mapping statistics across cell lines.

tive restriction sites (RSs)] rather than bins using 6-cutter restriction. Restriction fragment-based binning might be problematic for more frequent cutters such as 4-cutter (e.g. MboI) (2), since restriction fragment length is much smaller on average and interacting DNA sites are more likely to span more than one restriction fragment, thus highlighting the need for an approach to accurately infer interacting sites that span individual restriction fragments.

To address these limitations, we propose HIPPIE2 (Figure 1A and B), a novel computational method that uniquely infers the locations of DNA PIRs by identifying regions enclosed by restriction events observed on both sides of DNA–protein Hi-C construct (Figure 2; 'Materials and Methods' section; Supplementary Figure S6) using multiple sources of information including read-pileup information, patterns of observed restriction events, ligation constraints, and read strand and orientation (Figure 2A; Supplementary Figures S6–S8; Table 1).

The main novelty of HIPPIE2 algorithm is the development of an approach for interaction detection that does not use binned data or fragment-based mapping in contrast to HIPPIE and most of the existing methods (Table 1; Figures 1 and 2; Supplementary Figures S6–S8) to increase analysis resolution and precision.

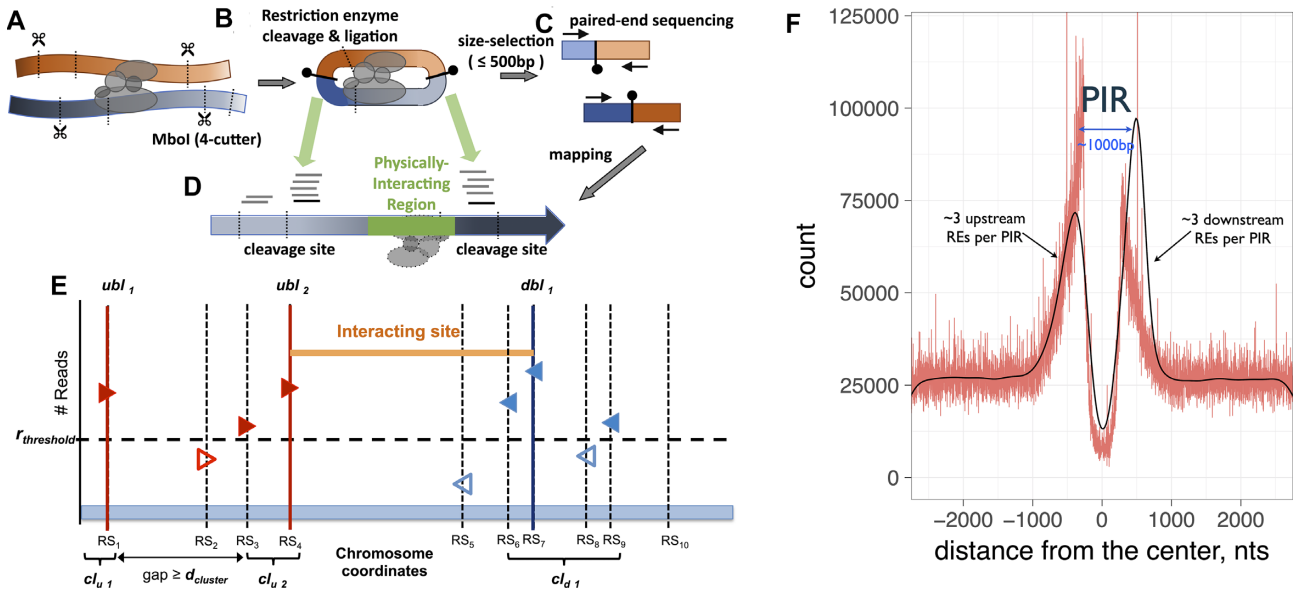

Figure 2. Hi-C model and identification of DNA PIR. (**A**) Interacting DNA regions are cut by the MboI restriction enzyme. (**B**) Cut fragments are ligated together and size selected. (**C**) Paired-end sequencing is performed on the ligated fragments. (**D**) Read pileups around the cleavage sites inform the identification of the PIR. (**E**) Genomic view of read pileups, RSs and PIR locations. Upstream (*ubl*) and downstream (*dbl*) boundary locations for PIRs correspond to most consistently cut (as evidenced by the number of reads) restriction/ligation sites. (**F**) Distribution of restriction events around DNA PIRs identified by HIPPIE2. Shown is the distribution of the restriction events for PIRs in GM12878 cell line.

Table 1. Comparison of HIPPIE2 with other Hi-C methodologies

Method	Input	Resolution	Output	Downstream analyses
HIPPIE2	Raw reads	Dynamic, restriction event-based (1 kb average)	High-resolution RS-based PIRs, PIR–PIR interactions	Genic and cell type-specific epigenetic annotation of interactions, identification of mediating TFs, identification of enhancer–promoter interactions
HIPPIE (15)	Raw reads	Restriction fragment-based (4 kb average)	Full restriction fragment-based PIRs	Annotated interactions, enhancer–promoter interactions
HiCCUPS (2,12)	Raw reads	Fixed bins	Loops (bin–bin interactions)	NA

This strategy allows HIPPIE2 to identify individual interacting DNA elements with better specificity than binning. HIPPIE2 uses cell type-matched functional genomic data to characterize the interacting PIRs into functional categories, including enhancers and promoters. This enables the high-resolution identification of cell type-specific enhancer–promoter interactions, and we show a corresponding enrichment in PIR–PIR interactions of TF binding sites (TFBSs) for TFs known to be involved in enhancer–promoter interactions. HIPPIE2 is open source (https://bitbucket.com/wanglab-upenn/HIPPIE2) and freely available as a full pipeline to automate analysis from raw Hi-C reads to identification of PIRs, significant PIR–PIR interactions, functional genomic annotations and TF analysis.

MATERIALS AND METHODS

Hi-C data acquisition and genome mapping

For our analysis, we used the Hi-C datasets for GM12878 (primary and secondary replicates), HMEC, HUVEC, IMR90, K562 and NHEK cell lines from (2) (GEO database accession number GSE63525). For each condition, we acquired FASTQ files from the SRA files available on GEO corresponding to sequencing libraries within each condition. Each library was mapped separately and then combined for downstream analyses. HIPPIE2 first aligns the paired-end reads to the human genome (GRCh37/hg19 assembly) using the STAR aligner (17) allowing only unique mapping (full parameters available in HIPPIE2 open-source repository, `starMappingToBam.sh` script). Each of the single-end reads from a read pair was first mapped separately and then re-associated with the corresponding second read in a read pair. To improve mapping, both contiguously mapped and chimeric reads are identified and paired. Both halves of a chimeric read were required to map uniquely and have a minimum mapped length of 22 nt. For those paired-end reads with a chimeric read involved, we required that the pairing partner of the chimeric read (a single-end read) mapped in the proximity of one of the two split halves spanned by the chimeric read.

Hi-C read normalization

To remove potential random ligation events, including uncut, self-ligated, or re-ligated read pairs, we filtered out the read pairs that are <5000 bp apart from each other as

suggested in (9,18). In addition, to correct for all possible Hi-C experimental biases including length of the cross-linked DNA fragments, RS accessibility or ligation rate of the restriction enzyme digested fragments, we normalized the read counts using the matrix normalization method by Knight and Ruiz (19) as used in (2). Additionally, to avoid any biases on detecting the region that cannot be mapped as a unique genomic locus, we also removed from the analysis RSs that have mappability <0.8. We found that 96% of the RSs have mappability >0.8; i.e. most of RSs had high mappability given a relatively long read length (101 nt).

Identification of PIRs

To identify DNA PIRs, we utilized the idea that each single-end Hi-C read is always located in the proximity of an RS that serves as both the restriction enzyme cleavage and ligation site in the Hi-C protocol. The RSs correspond to sites in the genomic DNA containing sequence that can be recognized by the restriction enzyme, e.g. 'GATC' for restriction enzyme MboI. After HIPPIE2 maps reads, it first determines corresponding RSs (cleavage/ligation sites), and infers the relative position (upstream or downstream from the RSs) for the DNA PIR.

The cleavage/ligation sites are identifiable from the mapping information of Hi-C paired-end reads because (i) a proper DNA ligation forms a phosphodiester bond between the 5′ phosphate of the donor DNA and the 3′ hydroxyl of the acceptor DNA and (ii) the strand orientation pattern reported by Illumina sequencer is restricting the combinations of upstream or downstream cleavage/ligation site of each read pairs. The workflow of identifying all PIRs and PIR–PIR interactions along the genome includes three major phases: (i) finding ligation junctions for read pairs; (ii) identifying PIRs; and (iii) finding PIR–PIR interactions. Each phase is described next.

Identifying ligation junctions for read pairs. Each mapped read has two candidate (nearest upstream or downstream) RSs to be assigned as the restriction enzyme cut-and-ligation site. To determine the cut-and-ligation site, we first determine which type of interaction has happened based on the mapped strand orientation (Supplementary Figures S7 and S8). We enumerated all possible ligation types: head/tail, tail/head, head/head or tail/tail ligations, where head is the end with smaller genome coordinate and tail is the end that is on a larger coordinate of the chromosome.

Because (i) the ligation of two DNA fragments is formed by a phosphodiester bond between a 3′ hydroxyl and a 5′ phosphate and (ii) Illumina paired-end sequencing reads are generated from opposite strands from the sequenced DNA fragments, we can narrow down four possible ligation types for each paired-end reads to two scenarios using its strand orientation. For the read strand combinations of +/− or −/+ (different strand), the two possible ligation types are either head/tail or tail/head ligations (Supplementary Figure S7, left). Similarly, for the read strand combinations of +/+ or −/− (same strand), the two possible ligation types are either head/head or tail/tail ligations (Supplementary Figure S7, right). Next, because of the size-selection step in the Hi-C protocol, cut-and-ligation events are expected to generate read pairs within 500 bp of the restriction enzyme (MboI) cutting sites due to the size selection; to resolve the two possible cases of head/tail or tail/head for +/− and −/+, we calculated the two possible sums of the two distances to the nearest cutter sites, and ruled out the ligation event that made the sum >300 bp, which would be result from ligation of non-specific cleavage product (8) in the Hi-C experiment (Supplementary Figure S8). As shown in Supplementary Table S2, the observed fractions of strand orientation combinations for sequenced Hi-C read pairs are close to uniform as expected from the stochastic nature of the proximity ligation reaction.

Identifying PIRs. With the identified RSs that form DNA–DNA ligation junctions, we further identify PIRs (algorithm in Supplementary Figure S6B). First, we note the sum of upstream and downstream read counts (single-end reads from read pairs) for each RS identified in the previous step. To group restriction events corresponding to the same interaction (interacting region), we clustered RSs separately for upstream and downstream read counts by thresholds of the maximum gap (d_{cluster}) and the minimum read ($r_{\text{threshold}}$). The maximum gap is defined as the third quantile of the restriction fragment distance distribution, and the minimum read requirement is defined as the median of the normalized read distribution for each chromosome. Within each corresponding cluster, we identify the RSs with the maximum read count (i.e. most *consistently* cut site) as the candidate flanking ends for PIRs. Finally, we matched the nearest upstream and downstream candidate flanking ends with a max-gap algorithm (in this study, the max gap is 4000 bp), and report the PIRs as regions that are enclosed by the upstream and downstream RSs with the maximum read count in the upstream and downstream restriction clusters.

Finding all PIR–PIR interactions. We find the interactions between PIRs by tracing the Hi-C read pairs that participated in the identification of PIRs (algorithm in Supplementary Figure S6C). For each PIR identified in the previous step, IDs of single-end reads in the left and right RS clusters are used to identify PIRs containing mate reads (i.e. other single-end reads from read pairs) as interacting partners. All such PIR–PIR interactions are then reported along with the read counts.

Identifying significant PIR–PIR interactions genome-wide

To identify significant intrachromosomal PIR–PIR interactions, we applied the Fit-Hi-C method (11) in R v3.2.3. For each of the autosomal chromosomes (1–22) and chromosome X, we split all observed PIR–PIR interactions into 2000 distance groups according to the linear distance (in nucleotides) between interacting PIRs. We filtered out the PIR pairs that are <5000 nt apart. For each distance group, we calculated the average distance and the average normalized read counts of the interacting PIRs. With the 2000 aggregated data points, we fit the normalized read counts by the function of distance using *smooth.spline* function in R. After the first spline fitting, we removed the outliers as described in (11) and fit the second spline function. We then reported PIR–PIR interactions that are significant after Benjamini–Hochberg correction (adjusted P-value $\leq$0.05).

Overlap with DNA loops from Rao *et al.*

We compared PIR–PIR interactions in our study with the set of DNA loops identified in (2) using HiCCUPS. We downloaded the set of loops and Hi-C loci (DNA regions that are participating in significant DNA loops) and filtered to include only those with the highest 10 kb resolution from GEO database under accession number GSE63525. We then overlapped these loop anchors and interactions with HIPPIE2-identified PIRs and PIR–PIR interactions using a custom script (available in the HIPPIE2 software repository) using awk, bedtools v2.25.0 (20) and Python v2.7.9.

Functional and genomic annotation data

We downloaded the cell type-specific ChIP-seq peak data for histone modifications (H3K4me1, H3K4me2, H3K4me3, H3K27ac and H3K36me3), DNase I hypersensitive sites, and TFs or DNA-binding proteins (RNA polymerase II, p300 and CTCF) from the 2011 freeze of the UCSC Genome Browser (21) for ENCODE datasets and directly from the web portal (https://egg2.wustl.edu/roadmap/web_portal/index.html) for Roadmap datasets including combinatorial epigenomic states from ChromHMM (22) that we used to identify enhancer states. Functional and genomic annotation data integrated into HIPPIE2 are summarized in Supplementary Table S1.

Enrichment analysis of functional genomic overlaps

To estimate the extent of overlap between PIRs and regulatory and epigenetic marks genome-wide, we calculated the sum of overlapped nucleotides between PIRs and each signal track (regulatory/epigenetic mark) genome-wide as the observed value. We sampled (1000 times) random genomic regions from the genome with length distribution matched with the length of PIRs. We calculated the average of 1000 sums of overlaps between the sampled regions and each of the signal tracks. We then reported the percentage differences between the observed value and the averaged value from the background as the enrichment of the PIRs for each of the signal tracks. All region intersections were performed with bedtools v2.25.0 (20).

Regulatory and genetic annotation of the interacting PIRs

HIPPIE2 annotates PIRs as enhancers, promoters, exons, introns or intergenic elements. To do this, we used the cell type-matched enhancer annotations described above and gene models downloaded from RefSeq (23). We annotate as enhancers the promoter-interacting PIRs that overlapped the enhancer or weak enhancer annotation from the genome segmentation track [ChromHMM (22)]. We also annotated all promoter-interacting PIRs as an enhancer if they overlapped an open chromatin region with H3K4me1 or H3K27ac ChIP-seq peak, while not overlapping H3K4me3 and H3K27me3 peaks. The rest of the PIRs were annotated as promoters, exons, introns and intergenic elements using RefSeq gene models (GRCh37/hg19 assembly). The promoters were defined as 500-bp-long regions upstream of the RefSeq TSS of protein-coding genes. We then annotated PIRs as promoter, exonic, intronic or intergenic elements (in this prioritized order) based on their overlap with RefSeq gene models. To calculate the background expectations of interactions between annotations a and b, we used the product of the proportion of individual PIRs in annotation a and the proportion in annotation b.

Transcription factor binding analysis of PIR–PIR interactions

To identify PIRs with evidence of TF binding, we used Factorbook data (24) that integrate ENCODE ChIP-seq experimental data with computational prediction of TFBSs to comprehensively survey protein–DNA binding genome-wide. The Factorbook data were obtained from UCSC hg19 database (factorbookMotifPos table, release 4). The Factorbook data contain 161 factors and the motifs were discovered from 91 cell types. We focused on 133 known DNA-binding TFs. We filtered out the TFs with <10 binding sites within PIRs genome-wide. For each PIR, we reported all TFs that have at least one binding site within that PIR. We reported enrichment for each of the surveyed binding motifs in PIR–PIR interactions. To do this, we categorized PIR–PIR interactions according to the classes of interacting PIR elements (enhancers, promoters, exons, introns or intergenic elements). We estimated binding motif enrichment as observed/expected frequency odds ratio. We computed the expected probability as the probability of the first PIR class (C_i) having a motif M_k times the probability of the second PIR class (C_j) having another motif (M_l) as follows:

$$\begin{aligned}\text{Prob}\left(M_k,\ M_l \text{ observed in } C_i,\ C_j\right) &= P\left(M_k|C_i\right) \times P\left(M_l|C_j\right)\\ &= \frac{P(M_k,C_i)}{P(C_i)} \times \frac{P(M_l,C_j)}{P(C_j)}\\ &= \frac{\#(C_i \text{ containing } M_k)}{\#C_i}\\ &\quad \times \frac{\#(C_j \text{ containing } M_l)}{\#C_j}.\end{aligned}$$

We performed a binomial distribution test to report the significance of observed binding motifs in each type of PIR–PIR interaction. To compare against the BioGRID database (25,26), we downloaded the list of TF–TF interactions across cell lines and searched for TF matches by name and by alias.

RESULTS

HIPPIE2 identifies fine-scale DNA PIRs

The HIPPIE2 method presented in this paper further develops our HIPPIE method (15): HIPPIE2 applies a newer read mapping protocol to resolve chimeric reads, uses an interaction calling algorithm and introduces novel algorithms to dynamically identify fine-scale interacting regions (Figures 1 and 2; Supplementary Figures S6–S8) instead of binning reads into full restriction fragments used in HIPPIE (15). To illustrate our method, we used HIPPIE2 to analyze high read depth Hi-C sequencing datasets (2) using the 4-bp cutter MboI across six human cell lines that had matching functional genomic data from ENCODE or Roadmap (5,27), including K562, HMEC, HUVEC, IMR90, NHEK and GM12878, with two replicates for GM12878 (Figure

1A and B). For each cell line, we mapped the raw Hi-C read pairs using STAR (17) (see 'Materials and Methods' section), uniquely mapping between 73.6% and 85.4% of Hi-C reads across the 51 separate libraries for these cell lines (Figure 1C; Supplementary Figure S1). Following (2), we normalized the read counts using matrix normalization by Knight and Ruiz (19) (see 'Materials and Methods' section). Using normalized counts, HIPPIE2 identifies PIRs as the DNA regions flanked on both sides by RSs that were observed to be consistently cleaved/ligated in a given Hi-C sequencing library (Figure 2D–F), using information from the Hi-C sequencing read-out including the read mapping coordinates, distances from reads to their nearest RSs, DNA ligation constraints and strand orientations (+/−) of the mapped read pairs, and relative locations of DNA interaction sites with respect to mapped reads (see 'Materials and Methods' section; Figure 2). This dynamic PIR-based approach enables finer scale identification of specific interacting DNA regions compared to the genomic binning-based approaches (Table 1).

In total, HIPPIE2 called between 1584 000 and 1886 000 PIRs from chromosomes 1–22 and X across cell lines (Figure 3A). These PIRs had an average length of 1006 bp consistent across cell lines (Supplementary Figure S2A), which corresponds to 2.4 average restriction fragment length. Across libraries, these identified PIRs covered 53.2–59.3% of the genome (Supplementary Figure S2B). HIPPIE2 annotated these PIRs with gene annotations (28) including promoters, exons and introns, which found that a majority of PIRs in all cell types were intergenic and the next largest class of overlaps were in mRNA introns, supporting the regulatory roles of these PIRs (Figure 3B). Comparing overlaps between HIPPIE2-called PIRs and DNase-seq-based regions of open chromatin in each of cell types from Roadmap/ENCODE, we found that 73.84–79.04% of open chromatin regions overlapped with at least one PIR across cell types, with an average of 69.97% of the open chromatin regions covered by PIRs (Figure 3C). PIR identification by HIPPIE2 is highly robust, with 92.3% PIRs (1649 417) found in both of the GM12878 replicates.

HIPPIE2 detects fine-scale chromatin interactions

To identify which PIR–PIR pairs are significantly interacting, HIPPIE2 applies the Fit-Hi-C algorithm (11) using the normalized read counts and linear genomic distance between pairs of potentially interacting PIRs (see 'Materials and Methods' section). Across cell lines, HIPPIE2 identified between 42 500 and 1194 010 intrachromosomal significant PIR–PIR interactions (>5 kb apart, adjusted P-value $\leq$0.05, Figure 4A). To investigate robustness of interaction calling, we compared the two GM12878 replicates. Consistent with the lower sequencing depth of the replicate library (2.5 billion versus 3 billion reads), we identified fewer significant interactions and PIRs involved in significant interactions in the replicate library (Figure 4A), but found a significant overlap between PIRs and PIR–PIR interactions (Figure 4B and C), with majority (66.2%; 274 445/414 343; Jaccard index = 0.3696) of PIRs and 31.1% (196 343/631 610; Jaccard index = 0.1205) of PIR–PIR interactions in the replicate found in the primary library. This level of replication is consistent with prior studies of Hi-C replication (14), which found similar levels of replicated interactions across Hi-C datasets [e.g., Jaccard index <0.1 on GM12878 Rao data for all tested methods in (14)].

To interrogate the relationship between sequencing depth and interaction replication, we binned the interactions from the GM12878 primary library into deciles by read coverage (analogous to downsampling) and compared their replication rates. We found a striking positive correlation between read coverage and replication rate ($R^2 = 0.9398$, Figure 4D), suggesting that reproducibility between replicates may be increased with a higher sequencing depth.

Comparison with uniform binning-based approach

HIPPIE2 provides a more accurate approach for identifying fine-scale interacting sites by design: previous methods that use a binning-based approach maximize their statistical power to detect interactions, with a trade-off of accuracy for identifying the interacting site (Table 1). Due to the fundamentally different natures of the binning-based algorithms compared to HIPPIE2 and the lack of a 'ground truth' dataset of expected Hi-C interactions, it is challenging to directly compare the HIPPIE2 interactions with these previous methods. However, to explore the differences between the binning approaches and HIPPIE2 PIR-based approach, we compare the HIPPIE2 results with the results from (2) obtained using the HiCCUPS method [we note that a detailed comparison among Hi-C methods has been reported in the recent study by Forcato *et al*. (14)].

We compared our HIPPIE2-identified PIRs with the HiCCUPS-identified loop anchors and interactions (bin size = 10 000) (2). Across cell lines, the set of PIRs identified by HIPPIE2 is consistent with and is complementary to the previously identified set of interacting genomic regions: we found that HIPPIE2 PIRs covered an average of 60.2% of HiCCUPS-identified loop anchors across cell lines, with the highest proportion (91.4%) in the primary, most deeply sequenced GM12878 library (see 'Materials and Methods' section; Figure 4E). The HMEC, HUVEC and NHEK cell lines were the only ones with a proportion <50%, corresponding to their shallower sequencing depth (Pearson R^2 = 0.862 between sequencing depth and replication proportion).

Next considering HIPPIE2 PIR–PIR interactions and HiCCUPS loops, we found that 37.4% of HiCCUPS loops in GM12878 primary library, with an average of 16.6% of HiCCUPS loops across cell lines, were supported by significant HIPPIE2 PIR–PIR interactions across cell lines (Figure 4E). When we matched the bin size (10 000) used in HiCCUPS analysis by expanding HIPPIE2 PIR–PIR interactions so that each PIR covered at least 10 kb, we found that the majority (55.8%) of HiCCUPS loops replicated in the primary GM12878 library (highest sequencing depth) and the average proportion of replicated HiCCUPS loops across cell lines increased to 28.88% from 16.6%, with an average of 68% of HiCCUPS loop anchors replicated (Supplementary Figure S3C). Interestingly, each PIR overlapping a HiCCUPS-identified loop anchor was involved in an average of between 5.62 and 21.52 significant PIR–PIR interactions across cell lines compared to a single interac-

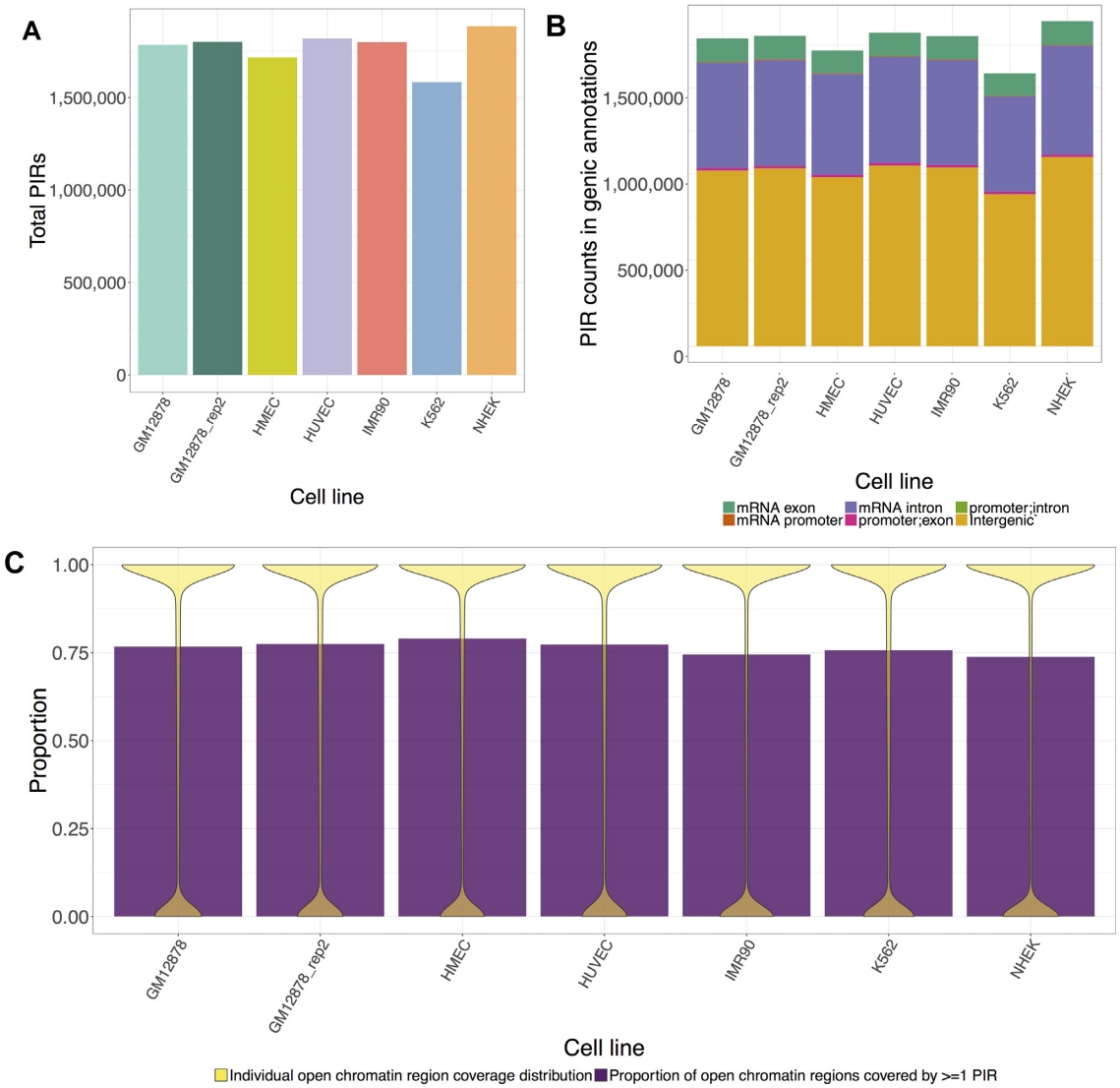

Figure 3. PIR characteristics. (**A**) Total PIRs identified across cell lines. (**B**) Localization patterns of PIRs in various genomic annotations. (**C**) Overlap patterns of PIRs with cell type-matched open chromatin annotations. For each cell type, cell-specific open chromatin regions were overlapped with HIPPIE2-called PIRs. Shown are (i) proportion of open chromatin regions overlapping with PIRs in each cell type (purple bars) and (ii) fraction of individual open chromatin regions covered by PIRs (distribution in yellow).

tion (loop) reported by HiCCUPS, corresponding to an average fold enrichment of 10.07 more interactions identified by HIPPIE2 than by the HiCCUPS binning/loop detection approach. Overall, HIPPIE2 identified about two orders of magnitude more interactions than the bin-based approach across all cell lines (2794 123 versus 27 827). This illustrates that HIPPIE2 identifies more, finer scale regulatory interactions than the bin-based approach to maximize power to detect large-scale genomic architecture rather than a multiplicity of fine-scale regulatory interactions. For example, in the 1-Mb locus on chr14 investigated in Rao *et al.* (2) (chr14:94 000 000–95 000 000), there were 1082 HIPPIE2 significant PIR–PIR interactions within four bin–bin interactions that were called in the original study (Supplementary Figure S4). We have also compared HIPPIE2 with the recent Binless method proposed by Spill *et al.* (29) (Supplementary Figure S9) with 86/102 (84%) HIPPIE2-called PIRs overlapping with Binless loci. As can be seen from the figure, while HIPPIE2 does not recover all the interacting sites or indeed the smaller interactions from the Binless method, HIPPIE2 recovers a lot more longer range interactions especially closer to the HCF gene, suggesting that HIPPIE2 and Binless are complementary methods where Binless may recover shorter range interactions while HIPPIE2 may more reliably identify more long-range interactions (indeed by design with the 5-kb threshold; see 'Materials and Methods' section).

When comparing power to recall validated interactions from 5C data (Supplementary Tables S6 and S7), using the benchmark results from (14) on the same GM12978 Rao *et al.* data [as reported in figure 2 in (14)], HIPPIE2 recovers more interactions (99 5C-based loops) than most of other benchmarked methods (14) including HiCCUPS, which recover an average of 11.33 5C-based loops (14).

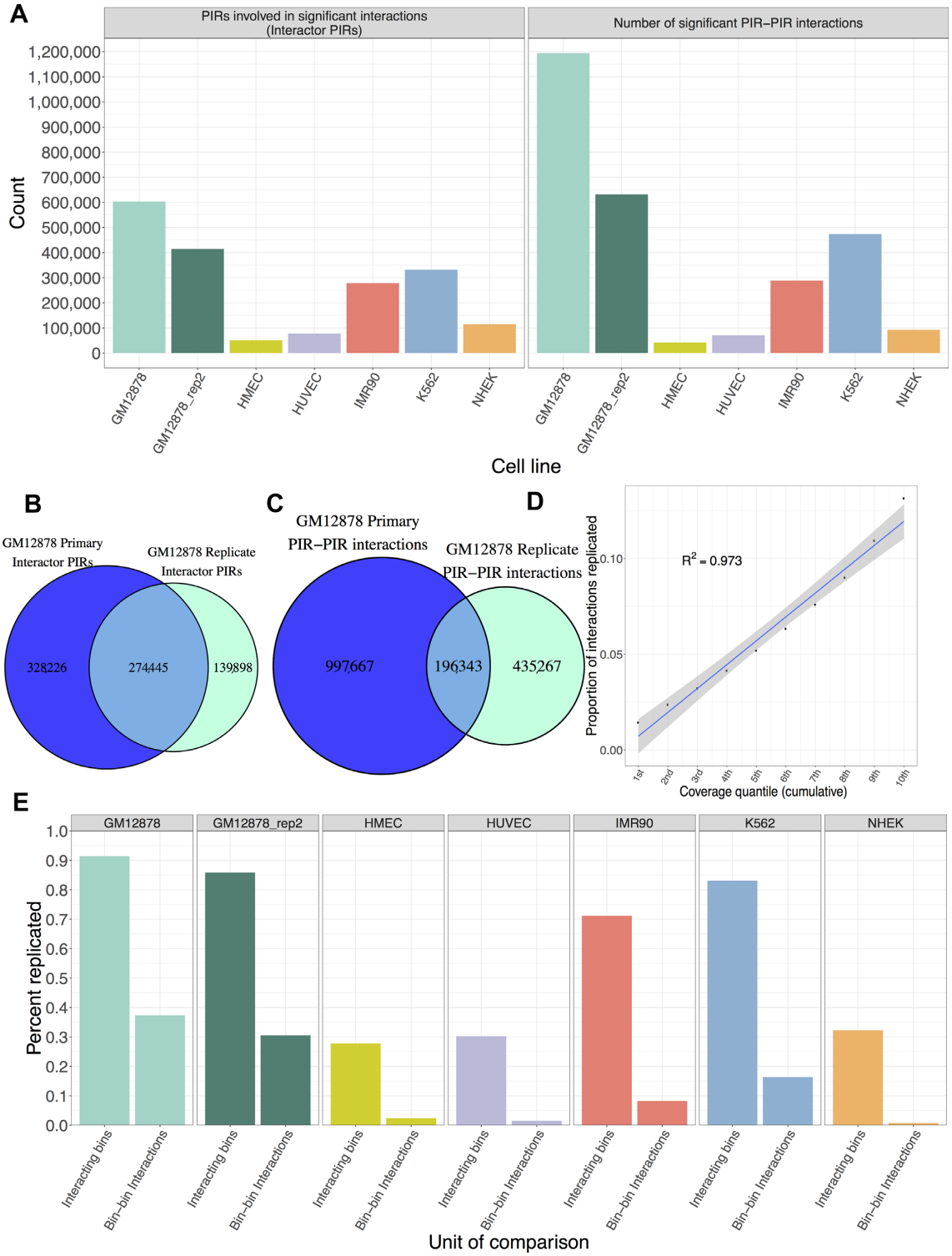

Figure 4. Characteristics of significant PIR–PIR interaction identification and replication. (**A**) Counts of PIRs involved in significant interactions (left) and number of significant PIR–PIR interactions (right) across cell lines. (**B**) Number of PIRs involved in significant interactions replicated between the primary and secondary GM12878 libraries. (**C**) Number of PIR–PIR interactions replicated between the primary and secondary GM12878 libraries. (**D**) Plot of replication rate against PIR read coverage quantiles. Correlation is the Pearson correlation. (**E**) Replication of Rao data by HIPPIE2. Interacting bins refer to 10-kb Rao bins involved in significant interactions and bin–bin interactions are significant interactions. Across cell lines, an average of 60.2% of HiCCUPS loop anchors and 16.6% of loops are reproduced by HIPPIE2, while for most deeply sequenced GM12878 library 91.4% of loop anchors and 37.4% of loops were reproduced by HIPPIE2 (see also main text for other comparisons).

Table 2. Enrichment of all GM12878 PIRs in transcriptional and architectural protein binding sites relative to randomly sampled background genomic regions

ChIP-seq dataset	Number of base pairs overlapped by PIRs	Number of ChIP-seq peaks overlapped by PIRs	% increase in number of base pairs overlapped by PIRs	% increase in number of ChIP-seq peaks overlapped by PIRs	Empirical *P*-value
CTCF	10 214 990	41 134	20.69	18.51	<0.001
P300	4662 562	16 509	23.99	19.12	<0.001
PolII	2758 845	9678	23.07	18.21	<0.001

HIPPIE2 PIR–PIR interactions are enriched in regulatory genomic features

To evaluate how PIRs co-locate with binding of factors known to be involved in genome architecture and transcriptional mechanisms, we overlapped HIPPIE2-identified PIRs from the primary GM12878 library, the most deeply sequenced library, with ENCODE ChIP-seq binding sites for CTCF, PolII and P300 (Table 2) from the same cell type (27). We found that HIPPIE2 PIRs overlapped 92% (41 134 out of 44 597) CTCF sites, consistent with studies that suggest CTCF has a role in mediating chromatin interactions (30). Similarly, for PolII and P300, associated with transcriptional and enhancer activity (31), we found high overlaps at 96.5% (9678 out of 10 026) and 96.3% (16 509 out of 17 150 sites), respectively. We randomly sampled 1000 sets of background genomic regions matched to the distribution of PIR lengths and calculated percent enrichment of the GM12878 PIRs relative to these background sites (see 'Materials and Methods' section; Table 2). We found that the GM12878 PIRs had increases of 20–23% base pairs of overlap over background and overlapped 18–19% more ChIP-seq sites, suggesting that the GM12878 PIRs are involved in genomic architecture and regulatory function.

To characterize the function of PIRs involved in significant interactions, HIPPIE2 automatically annotates PIRs with DNase-based open chromatin regions, enhancers defined by combinatorial epigenomic status using ChromHMM (22), the enhancer-associated histone modifications H3K4me1 and H3K27ac (32,33), the inactive or poised enhancer histone modification H3K27me3 (34) and the promoter-associated histone modification H3K4me3 (35), all in the matching cell types from ENCODE or Roadmap (5,27) (Figure 5A and B; Supplementary Tables S4 and S5). Between 18.35% and 42.26% of PIRs involved in significant PIR–PIR interactions overlapped open chromatin sites across cell lines, with the lowest proportions in the shallowest sequencing libraries, while the other annotations encompassed between 7.4% (H3K4me3) and 17.2% (Roadmap ChromHMM enhancers) of PIRs on average across cell lines. By comparing against samples of length-matched background intervals (see 'Materials and Methods' section), we found that the PIRs involved in significant interactions were enriched for overlaps with all of the active epigenomic marks in every cell line except for NHEK, where PIRs were depleted of overlaps with all annotations except H3K4me3 (Figure 5B). The repressive mark H3K27me3 had the smallest average enrichment (40.75%) across cell lines, suggesting that significant PIR interactions are associated with active regulatory elements (2,16), which showed much stronger enrichments.

We have also assessed degree of overlap between HIPPIE2-called PIRs and open chromatin regions in cell type-specific manner using Jaccard indexes and odds ratios (Supplementary Table S4). Across cell lines, we observe an average enrichment odds ratio of 2.1 indicating high degree of overlap between PIRs and open chromatin regions.

Similarly, we have also assessed an overlap between PIRs involved in significant interactions (interactor PIRs) and cell type-specific functional genomic annotations (Supplementary Table S5). Across cell lines, we observed significant overlaps with all functional genomic features compared, with the strongest enrichments in enhancer- and promoter-associated histone modifications including H3K27ac and H3K4me1/3 (Supplementary Table S5).

HIPPIE2 PIR–PIR interactions are enriched for enhancer–promoter mechanisms

HIPPIE2 identifies specific enhancer–promoter interactions by classifying PIRs as enhancers if they overlap (i) an open chromatin region and a shared H3K4me1/H3K27ac peak and/or (ii) a ChromHMM (22) epigenomic enhancer (see 'Materials and Methods' section). For enhancer elements, HIPPIE2 further requires that the putative enhancer PIR display at least one significant interaction with a promoter-overlapping PIR and does not overlap any H3K4me3 (active promoter mark) or H3K27me3 (repressive mark) peaks in the matching tissue. We found that the percentage of regulatory interactions (enhancer–promoter, enhancer–enhancer or promoter–promoter pairs) accounted for an average of 2.36% (ranging from 0.36% to 3.78%) of significant interactions across cell lines, a significant enrichment compared to the background expectation of an average of 0.076% (ranging from 0.0049% to 0.2%) of interactions. For enhancer–promoter interactions specifically, we detected an average of 51.95× enrichment over the background expectation (ranging from 26.79× to 86.47×), and these were the most enriched interactions in all cell types, suggesting that the interactions identified by HIPPIE2 are indeed reflective of transcriptional regulatory processes (see 'Materials and Methods' section; Figure 5C).

HIPPIE2 recovers a repertoire of known regulatory TFs mediating chromatin interactions

In order to elucidate the mechanisms underlying the observed enhancer–promoter interactions, HIPPIE2 annotates interacting PIRs with TFBSs from the Factorbook database that contains TFBSs for 133 DNA-binding proteins identified by ChIP-seq experiments (24). Combined

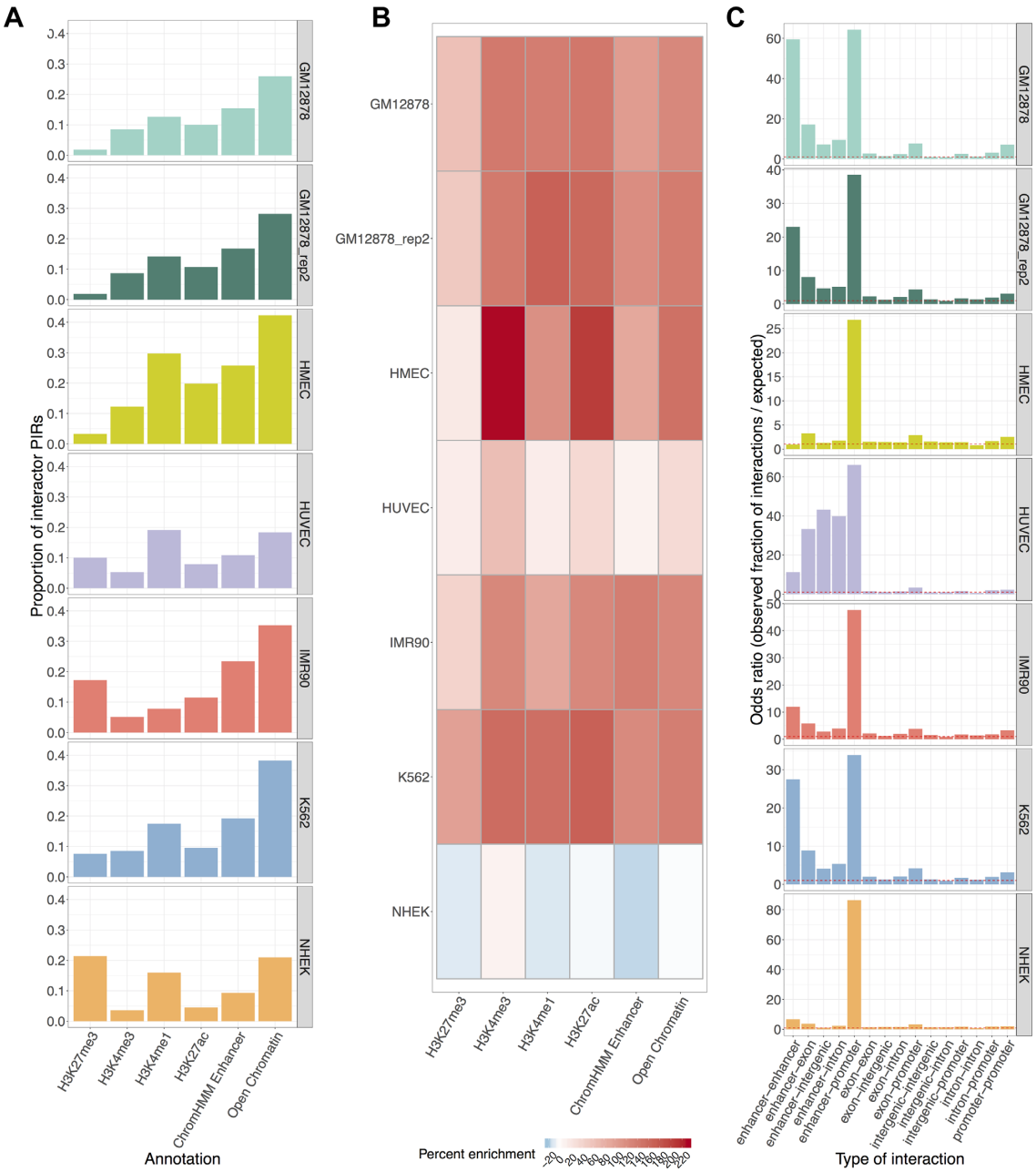

Figure 5. Regulatory annotation of PIR–PIR interactions. (**A**) Proportion of PIRs involved in significant interactions (interactor PIRs) overlapping cell type-matched functional annotations. (**B**) Enrichment of interactor PIRs relative to background expectation calculated by sampling. (**C**) Ratio of number of observed annotation–annotation PIR–PIR interactions relative to background expectations.

with our HIPPIE2-identified fine-scale PIR annotations, this approach enables HIPPIE2 to identify the TFs mediating enhancer–promoter interactions with high resolution. We found that an average of 14.11% of PIRs involved in all significant interactions had overlaps with known TFBS across cell lines, while an average of 39.2% of HIPPIE2-identified enhancer–promoter interactions across cell lines had an evidence of known TF binding (Supplementary Figure S5A).

To determine whether enhancer–promoter interactions were enriched in TFBSs, we quantified the observed/expected ratio for binding motif enrichment and used a binomial model to identify significant enrichments of TFs involved in enhancer–promoter interactions (see 'Materials and Methods' section). We found significant enrichments for 31 TFs in enhancers and 29 in promoters for a total of 33 unique TFs across all cell lines except for HUVEC and NHEK (Supplementary Figure S5B; Supplementary Table S3). To test whether these putative HIPPIE2-identifed TF–TF interactions correspond to known protein–protein interactions, we compared HIPPIE2 TF–TF interactions to the BioGRID database (25,26). To do this, for each cell line, we identified all the TFs involved in significant enhancer–promoter in-

teractions, quantified all their interactions in BioGRID and determined the proportion of BioGRID interactions involving these TFs that were recapitulated by HIPPIE2. We found that HIPPIE2-identified TF–TF interactions in GM12878 recapitulated most (78%) of known physical TF–TF interactions reported in BioGRID, with an average of 57.5% of known physical interactions between TFs in BioGRID (31–78%) recovered across cell lines. These proportions were strongly correlated with the sequencing depth of each cell line (Pearson $R^2 = 0.88$), suggesting that increased read depth may recover more BioGRID interactions.

We then stratified these TFs by how many different cell lines they were enriched in to identify regulatory mechanisms common across cellular contexts (Supplementary Figure S5C). This identified several TFs enriched in several cell lines that were consistent with known enhancer and chromatin architecture biology, including SP1, AP1, MYC, CEBPB, YY1 and CTCF. SP1 has been shown to function as a link of both sides of DNA, and is able to form a tetrameric structure and assemble multiple tetramers that facilitate a DNA looping structure (36). AP1 is a TF involved in cellular proliferation, transformation and apoptosis that forms heterodimers with the Jun oncogene (37). MYC is an oncogene involved in several different cancer types and exerts widespread transcriptional regulatory effects (38). CEBPB is another major enhancer-binding protein family that can aid the transition of enhancer elements from closed chromatin to a primed or poised state and is involved in immune and inflammatory responses (39). CTCF is a major architectural protein with a role in defining megabase-scale topologically associated domains as well as regulating smaller scale enhancer–promoter interactions such as those observed here (30,40). YY1 is another major architectural protein that cooperates with CTCF to mediate looping interactions involved in developmental processes and enhancer–promoter interactions (41,42).

DISCUSSION

In this paper, we introduce a novel method for Hi-C data analysis, HIPPIE2, which dynamically discovers fine-scale PIRs of the genome with increased resolution compared to previous methods, detects fine-scale chromatin interactions and provides functional and mechanistic characterization of these interactions. HIPPIE2 uses the pattern of restriction events as evidenced by sequencing read pileups relative to RSs to fine-map interacting DNA regions. Our results suggest that HIPPIE2 detects more specific, finer scale interactions at the gene regulatory level of chromatin architecture (average PIR length of 1006 bp), offering a complementary approach to the binning-based procedures (2,11,12,14,43,44).

Our method also complements restriction fragment-based methods (9,15) for mapping Hi-C data as an alternative approach for analyzing data from more frequent cutters with much smaller fragment length and interaction regions spanning more than one fragment. While fragment-based mapping methods were successfully used with data from low-frequency cutters (relatively large fragment size) where interactions were likely to be contained within the fragments themselves, HIPPIE2 addresses the main difficulty of using the data from more frequent cutters with much smaller restriction fragment sizes when interactions are likely to span more than one restriction fragment necessitating accurate inference of interacting sites/boundaries such as those provided by the HIPPIE2 algorithm.

HIPPIE2 algorithm, differently from restriction fragment-based or fixed-size genomic bin-based methods, introduces a dynamic algorithm for identifying precise location of interacting sites using multiple sources of information, including read-pileup information, patterns of observed restriction events, ligation constraints, and read strand and orientation.

While an original (2) normalization procedure (Knight–Ruiz method) was used for all Hi-C datasets in this study, the HIPPIE2 approach is designed to work independent of the normalization procedure (i.e. downstream of the data normalization steps), and as such our method can be easily paired with an appropriate normalization method that captures specific biases and characteristics of a particular experimental protocol or with an improved normalization procedure for high-resolution, non-binned data, which is an ongoing effort (29,45). Additionally, HIPPIE2 can be paired with other methods for calling significant interactions such as (46).

With our approach designed to work at the inherent resolution of the data (as determined by restriction enzyme cutting frequency and restriction efficiency), our method will prove useful in the analysis of chromosome conformation capture experiments with further increased sequencing depth or improved restriction protocols. Another natural application in which our approach will prove useful is the analysis of the data generated by the assays targeting particular types of interactions, such as Capture-C and Capture Hi-C (43,47) that capture promoter-centric interactions.

Furthermore, the fine-scale resolution of our method for detecting interacting regions enables analysis, identification and interpretation of specific proteins/TF complexes mediating these interactions. This TF analysis can be improved by *de novo* motif discovery in PIR sequences, incorporation of protein–protein interaction networks to identify protein/TF complexes and protein domain compatibility information. Using the identified interacting sequences and mediating TFs can help build predictive models for regulatory interactions such as (48,49). Another direction in which our fine-mapping HIPPIE2 method will prove useful is in comparison and analysis of changes in fine-scale regulatory networks during development or between different conditions. HIPPIE2 is freely available as an open-source pipeline (https://bitbucket.org/wanglab-upenn/HIPPIE2). HIPPIE2-generated interaction data are also available in UCSC Genome Browser hub (https://genome.ucsc.edu/s/alexamlie/HIPPIE2%20vs%20Rao%20all%20cell%20lines%20darker%20interaction%20lines).

DATA AVAILABILITY

HIPPIE2 software is freely available at https://bitbucket.org/wanglab-upenn/HIPPIE2. All of the functional genomic data and annotations used by HIPPIE2 are available at https://tf.lisanwanglab.org/

GADB/full_HIPPIE2_annotations.tar.gz. The corresponding interaction data tracks are available on the UCSC Genome Browser: https://genome.ucsc.edu/s/alexamlie/HIPPIE2%20vs%20Rao%20all%20cell%20lines%20darker%20interaction%20lines.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

FUNDING

National Institute of General Medical Sciences [R01-GM099962]; National Institute on Aging [U54-AG052427, U24-AG041689, T32-AG00255]. Funding for open access charge: National Institute on Aging [U24-AG041689, U54-AG052427].
Conflict of interest statement. None declared.

REFERENCES

1. Lieberman-Aiden,E., van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* , **326**, 289–293.
2. Rao,S.S.P., Huntley,M.H., Durand,N.C., Stamenova,E.K., Bochkov,I.D., Robinson,J.T., Sanborn,A.L., Machol,I., Omer,A.D., Lander,E.S. *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.
3. Norton,H.K., Emerson,D.J., Huang,H., Kim,J., Titus,K.R., Gu,S., Bassett,D.S. and Phillips-Cremins,J.E. (2018) Detecting hierarchical genome folding with network modularity. *Nat. Methods*, **15**, 119–122.
4. Yang,D., Jang,I., Choi,J., Kim,M.-S., Lee,A.J., Kim,H., Eom,J., Kim,D., Jung,I. and Lee,B. (2018) 3DIV: a 3D-genome interaction viewer and database. *Nucleic Acids Res.*, **46**, D52–D57.
5. Consortium,R.E., Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
6. Ma,W., Ay,F., Lee,C., Gulsoy,G., Deng,X., Cook,S., Hesson,J., Cavanaugh,C., Ware,C.B., Krumm,A. *et al.* (2015) Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. *Nat. Methods*, **12**, 71–78.
7. Imakaev,M., Fudenberg,G., McCord,R.P., Naumova,N., Goloborodko,A., Lajoie,B.R., Dekker,J. and Mirny,L.A. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods*, **9**, 999–1003.
8. Yaffe,E. and Tanay,A. (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.*, **43**, 1059–1065.
9. Jin,F., Li,Y., Dixon,J.R., Selvaraj,S., Ye,Z., Lee,A.Y., Yen,C.-A., Schmitt,A.D., Espinoza,C.A. and Ren,B. (2013) A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, **503**, 290–294.
10. Kaplan,N. and Dekker,J. (2013) High-throughput genome scaffolding from *in vivo* DNA interaction frequency. *Nat. Biotechnol.*, **31**, 1143–1147.
11. Ay,F., Bailey,T.L. and Noble,W.S. (2014) Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res.*, **24**, 999–1011.
12. Durand,N.C., Shamim,M.S., Machol,I., Rao,S.S.P., Huntley,M.H., Lander,E.S. and Aiden,E.L. (2016) Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.*, **3**, 95–98.
13. Lun,A.T.L. and Smyth,G.K. (2015) diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics*, **16**, 1–11.
14. Forcato,M., Nicoletti,C., Pal,K., Livi,C.M., Ferrari,F. and Bicciato,S. (2017) Comparison of computational methods for Hi-C data analysis. *Nat. Methods*, **14**, 679–685.
15. Hwang,Y.-C., Lin,C.-F., Valladares,O., Malamon,J., Kuksa,P., Zheng,Q., Gregory,B.D. and Wang,L.-S. (2014) HIPPIE: a high-throughput identification pipeline for promoter interacting enhancer elements. *Bioinformatics*, **31**, 1290–1292.
16. Hwang,Y.-C., Zheng,Q., Gregory,B.D. and Wang,L.-S. (2013) High-throughput identification of long-range regulatory elements and their target promoters in the human genome. *Nucleic Acids Res.*, **41**, 4835–4846.
17. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
18. Lajoie,B.R., Dekker,J. and Kaplan,N. (2015) The Hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods*, **72**, 65–75.
19. Knight,P.A. and Ruiz,D. (2012) A fast algorithm for matrix balancing. *IMA J. Numer. Anal.*, **33**, 1029–1047.
20. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
21. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,A.D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
22. Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.
23. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **33**, D501–D504.
24. Wang,J., Zhuang,J., Iyer,S., Lin,X.-Y., Greven,M.C., Kim,B.-H., Moore,J., Pierce,B.G., Dong,X., Virgil,D. *et al.* (2013) Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic Acids Res.*, **41**, D171–D176.
25. Chatr-Aryamontri,A., Breitkreutz,B.-J., Oughtred,R., Boucher,L., Heinicke,S., Chen,D., Stark,C., Breitkreutz,A., Kolas,N., O'Donnell,L. *et al.* (2015) The BioGRID interaction database: 2015 update. *Nucleic Acids Res.*, **43**, D470–D478.
26. Tyers,M., Breitkreutz,A., Stark,C., Reguly,T., Boucher,L. and Breitkreutz,B.-J. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–D539.
27. Bernstein,B.E., Birney,E., Dunham,I., Green,E.D., Gunter,C. and Snyder,M. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
28. Pruitt,K.D., Brown,G.R., Hiatt,S.M., Thibaud-Nissen,F., Astashyn,A., Ermolaeva,O., Farrell,C.M., Hart,J., Landrum,M.J., McGarvey,K.M. *et al.* (2014) RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.*, **42**, 756–763.
29. Spill,Y.G., Castillo,D., Vidal,E. and Marti-Renom,M.A. (2019) Binless normalization of Hi-C data provides significant interaction and difference detection independent of resolution. *Nat. Commun.*, **10**, 1938.
30. Phillips,J.E. and Corces,V.G. (2009) CTCF: master weaver of the genome. *Cell*, **137**, 1194–1211.
31. Shlyueva,D., Stampfel,G. and Stark,A. (2014) Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.*, **15**, 272–286.
32. Calo,E. and Wysocka,J. (2013) Modification of enhancer chromatin: what, how, and why? *Mol. Cell*, **49**, 825–837.
33. Thurman,R.E., Rynes,E., Humbert,R., Vierstra,J., Maurano,M.T., Haugen,E., Sheffield,N.C., Stergachis,A.B., Wang,H., Vernot,B. *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75–82.
34. Zhu,Y., Sun,L., Chen,Z., Whitaker,J.W., Wang,T. and Wang,W. (2013) Predicting enhancer transcription and activity from chromatin modifications. *Nucleic Acids Res.*, **41**, 10032–10043.
35. Bernstein,B.E., Mikkelsen,T.S., Xie,X., Kamal,M., Huebert,D.J., Cuff,J., Fry,B., Meissner,A., Wernig,M., Plath,K. *et al.* (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, **125**, 315–326.
36. Mastrangelo,I.A., Courey,A.J., Wall,J.S., Jackson,S.P. and Hough,P.V (1991) DNA looping and Sp1 multimer links: a mechanism for transcriptional synergism and enhancement. *Proc. Natl. Acad. Sci. U.S.A.*, **88**, 5670–5674.
37. Shaulian,E. and Karin,M. (2002) AP-1 as a regulator of cell life and death. *Nat. Cell Biol.*, **4**, E131–E136.

38. Dang,C.V. (2012) MYC on the path to cancer. *Cell*, **149**, 22–35.
39. Heinz,S., Romanoski,C.E., Benner,C. and Glass,C.K. (2015) The selection and function of cell type-specific enhancers. *Nat. Rev. Mol. Cell Biol.*, **16**, 144–154.
40. Ong,C.T. and Corces,V.G. (2014) CTCF: an architectural protein bridging genome topology and function. *Nat. Rev. Genet.*, **15**, 234–246.
41. Beagan,J.A., Duong,M.T., Titus,K.R., Zhou,L., Cao,Z., Ma,J., Lachanski,C.V., Gillis,D.R. and Phillips-Cremins,J.E. (2017) YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. *Genome Res.*, **27**, 1139–1152.
42. Weintraub,A.S., Li,C.H., Zamudio,A.V., Sigova,A.A., Hannett,N.M., Day,D.S., Abraham,B.J., Cohen,M.A., Nabet,B., Buckley,D.L. *et al.* (2017) YY1 is a structural regulator of enhancer–promoter loops. *Cell*, **171**, 1573.e28–1588.e28.
43. Mifsud,B., Tavares-Cadete,F., Young,A.N., Sugar,R., Schoenfelder,S., Ferreira,L., Wingett,S.W., Andrews,S., Grey,W., Ewels,P.A. *et al.* (2015) Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.*, **47**, 598–606.
44. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
45. Stansfield,J.C., Cresswell,K.G., Vladimirov,V.I. and Dozmorov,M.G. (2018) HiCcompare: an R-package for joint normalization and comparison of HI-C datasets. *BMC Bioinformatics*, **19**, 279.
46. Kaul,A., Bhattacharyya,S. and Ay,F. (2020) Identifying statistically significant chromatin contacts from Hi-C data with FitHiC2. *Nat. Protoc.*, **15**, 991–1012.
47. Hughes,J.R., Roberts,N., Mcgowan,S., Hay,D., Giannoulatou,E., Lynch,M., De Gobbi,M., Taylor,S., Gibbons,R. and Higgs,D.R. (2014) Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat. Genet.*, **46**, 205–212.
48. Whalen,S., Truty,R.M. and Pollard,K.S. (2016) Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.*, **48**, 488–496.
49. Schreiber,J., Libbrecht,M., Bilmes,J. and Noble,W. (2018) Nucleotide sequence and DNaseI sensitivity are predictive of 3D chromatin architecture. bioRxiv doi: https://doi.org/10.1101/103614, 30 January 2017, preprint: not peer reviewed.