C3D: A tool to predict 3D genomic interactions between cis-regulatory elements

Tahmid Mehdi[1,2,*], Swneke D. Bailey[1], Paul Guilhamon[1] and Mathieu Lupien[1,3,4,*]

[1] Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada.

[2] Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada.

[3] Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada.

[4] Ontario Institute for Cancer Research, Toronto, Ontario, Canada.

*Corresponding authors: tfmehdi@edu.uwaterloo.ca; mlupien@uhnres.utoronto.ca

ABSTRACT

Motivation: The 3D genome architecture influences the regulation of genes by facilitating chromatin interactions between distal cis-regulatory elements and gene promoters. We implement Cross Cell-type Correlation based on DNA accessibility (C3D), a highly customizable computational tool that predicts chromatin interactions using an unsupervised algorithm that utilizes correlations in chromatin measurements, such as DNaseI hypersensitivity signals.

Results: C3D accurately predicts 32.7%, 18.3% and 24.1% of interactions, validated by ChIA-PET assays, between promoters and distal regions that overlie DNaseI hypersensitive sites in K562, MCF-7 and GM12878 cells, respectively.

Availability: Source code is open-source and freely available on GitHub (https://github.com/LupienLabOrganization/C3D) under the GNU GPLv3 license. C3D is implemented in Bash and R; it runs on any platform with Bash (≥4.0), R (≥3.1.1) and BEDTools (≥2.19.0). It requires the following R packages: GenomicRanges, Sushi, data.table, preprocessCore and dynamicTreeCut.

INTRODUCTION

Genome-wide association studies (GWAS) have highlighted that non-coding variants play a large role in the etiology of common diseases by targeting cis-regulatory elements (CREs), mainly enhancers as opposed to promoters (Cowper-Sal·lari et al., 2012; Maurano et al., 2012; Schaub et al., 2012). However, it can be difficult to identify the target genes of these CREs. The challenge arises because enhancer-type CREs are often distal to their target genes relying on chromatin interactions to come in physical proximity with promoters, often skipping over the most proximal gene (Dekker et al., 2013; Sanyal et al., 2012). Means to delineate the 3D architecture of the genome are therefore required.

There are several assays for detecting chromatin interactions based on Chromosome Conformation Capture (3C) (Dekker et al., 2002), such as Chromatin Interaction Analysis with Paired-End-Tag sequencing (ChIA-PET) (Fullwood et al., 2009), Hi-C, 4C and 5C (Göndör et al., 2008; Dostie et al., 2006; van

Berkum et al., 2010). These methods require large amounts of material, not available for each cell or tissue of interest, and can be limited in resolution. Computational approaches to predict chromatin interactions present an attractive alternative or complementary approach. Ernst et al. (2011) and Corradin et al. (2013) proposed methods that utilize Pearson correlation coefficients between chromatin modification levels at CREs across a collection of cell types. Sheffield et al. (2013) applied a similar correlation method but comparing chromatin accessibility signals with gene expression levels. Thurman et al. (2012) predicted interactions from correlations based solely on chromatin accessibility across DNaseI hypersensitive sites (DHSs) using aggregated DNaseI hypersensitivity signals across 79 ENCODE cell lines (ENCODE Project Consortium, 2012). This allowed the pairing of distal CREs with active promoters as opposed to genes, which facilitated direct comparison with in vivo 3C-based assays. This method reported a significant enrichment of chromatin interactions called by ChIA-PET ($p < 10^{-15}$) and 5C ($p < 10^{-13}$) in K562 cells with highly correlated ($r > 0.7$) promoter-distal predicted interactions. Here, we expanded this approach with our Cross Cell-type Correlation in DNA accessibility (C3D) tool to predict chromatin interactions between CREs by expanding the analysis to an additional 100 biological samples [47 cell lines from Duke University (ENCODE Project Consortium, 2012) and 53 cell-types from the University of Washington (UW) (Roadmap Epigenomics Consortium et al., 2015)] into its correlation calculations and by applying statistical tests to identify significant correlations.

METHODS

C3D computes correlations between genomic measurements to predict chromatin interactions between genomic regions of interest. For instance, these genomic measurements can be DNase-seq (Boyle et al., 2008) or ATAC-seq (Buenrostro et al., 2013) signal intensities. It requires a reference catalogue of genomic coordinates found in a cell or tissue of interest (ie. all regions of accessible chromatin in a given cell or tissue type). It also requires an anchor file containing genomic coordinates of interest such as the promoters of genes of interest. Lastly, it requires signal files from a set of biological samples; these files should include genomic coordinates and their corresponding measurements.

As a case example, we used DNase-seq from ENCODE and the Roadmap Epigenomics Project (ENCODE Project Consortium, 2012; Roadmap Epigenomics Consortium et al., 2015) for the signal files. For each DNase-seq signal file, the maximal DNaseI signals are computed at each region in the reference catalogue. Quantile normalization (Bolstad et al., 2003) is then applied to ensure identical scaling of signals across signal files. The DynamicTreeCut algorithm (Langfelder et al., 2008) is then used to cluster the signal files based on the similarity between their signals across the reference catalogue. Each region in the reference catalogue receives a vector of consensus genomic measurements where each measurement corresponds to a cluster. The consensus genomic measurement of a cluster is the average signal from its respective signal files. This limits biases introduced by an overrepresentation of a particular tissue/cell-type. C3D then iterates through the sites listed in the reference catalogue, which overlap a region from the anchor file. For each of these sites, the Pearson correlations between the consensus genomic measurements of each reference catalogue genomic coordinates and those in the anchor file within a user-defined window are computed. For each pair of sites, a two-sided t-test is performed to determine whether they are significantly correlated and the p-values are adjusted through the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). Instructions for installing and running C3D are provided in the supplementary information along with a tutorial.

We validated the predictive accuracy of C3D using RNA polymerase II (POL2) ChIA-PET datasets for K562, MCF-7 (ENCODE Project Consortium, 2012) and GM12878 (Tang et al., 2015) cells. ChIA-PET called chromatin interactions are reported as pairs of genomic regions (or paired-end-tags) mapped to GRCh37/hg19. Interactions in these datasets are filtered to focus on interactions between promoters and distal sites that were 10 to 500 kilobase pairs (kbp) apart and overlapped DHSs reported in the matched cell-type (ENCODE Project Consortium, 2012; Roadmap Epigenomics Consortium et al., 2015). We define promoters as tags that are at most 1 kbp upstream from a transcription start site defined in GENCODEv19 (Harrow et al., 2012); all other tags are considered distal. Shuffling the distances between tags and rearranging them accordingly generated a matched control set of paired-end-tags (matched Control-PETs), for each interaction dataset, to test C3D's ability to control for false positives. We calculated correlations between promoter and distal DHSs in each of the three cell-types. For comparison, correlations were computed across DNase-seq databases of 79 cell lines from UW, 47 cell lines from Duke University (ENCODE Project Consortium, 2012), 53 cell-types from the Roadmap Epigenomics Project [also from UW] (Roadmap Epigenomics Consortium et al., 2015) and a combined database consisting of all 179 samples.

RESULTS

In K562, MCF-7 and GM12878 cells, interacting DHSs tend to have significantly higher correlations, across ENCODE and Roadmap Epigenomics DNase-seq data, than DHS-pairs in the matched Control-PETs (Figure 1A). Correlation thresholds for each cell are chosen as the value that 5% of matched Control-PETs have correlations above. With a cut-off of 0.62 ($p < 5.74 \times 10^{-7}$, $q < 1.68 \times 10^{-5}$) C3D predicted 32.7% and 18.3% of the ChIA-PET connections in K562 and MCF-7, respectively. In GM12878, a cut-off of 0.63 ($p < 3.32 \times 10^{-7}$, $q < 1.10 \times 10^{-6}$) allowed it to have a true positive rate (TPR) of 24.1% (Figure 1A).

To compare the influences of the four aforementioned databases on the predictive abilities of C3D, receiver operating characteristic (ROC) curves were generated and areas under the curves (AUCs) are measured (Figure 1B). Calculating correlations across the combined DNase-seq database was more effective than the individual datasets in predicting interactions in K562 as it generated the highest AUC. It produced the same AUC as the Roadmap Epigenomics Project database for GM12878 but it had a greater Matthews Correlation Coefficient (MCC) (Matthews, 1975) at high thresholds ($r > 0.63$) (Supplementary Table 1). In MCF-7, the Roadmap Epigenomics Project database had the highest AUC but its predictive accuracy was similar to that of the combined database at high thresholds (Figure 1B). With a threshold of 0.63, C3D achieved higher MCCs in all three cells with the combined database than the individual datasets (Supplementary Table 1). Thus, C3D can predict chromatin interactions and control for false positives if adequate databases and correlation thresholds are applied.

While we present a method to predict chromatin interactions, this method will also report correlations dependent on other biological processes, such as gene co-expression across cell-types reflected by the signals at DHSs within their promoters, stressing the need for their functional validation.

FUNDING

This work was supported by Prostate Cancer Canada and is proudly funded by the Movember Foundation (grant #RS2014-04 to M.L.), with the additional support of the Canadian Cancer Society and the Princess Margaret Cancer Foundation (M.L.). M.L holds an Investigator Award from the Ontario Institute for Cancer Research; a Canadian Institutes of Health Research (CIHR) New Investigator Award; and a Movember

Rising Star Award from Prostate Cancer Canada. P.G is supported by a CIHR Fellowship (MFE 338954). S.D.B. is a Knudson and CIHR postdoctoral fellowship recipient.

ACKNOWLEDGEMENTS

We thank Dr. Michael M. Hoffman for helpful comments and suggestions on the manuscript.

Conflict of Interest: none declared.

REFERENCES

Benjamini,Y. and Hochberg,Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Series B Stat. Methodol.*, **57**, 289–300.

Bolstad,B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Boyle,A.P. *et al.* (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311–322.

Buenrostro,J.D. *et al.* (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, **10**, 1213–8.

Corradin,O. *et al.* (2014) Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res.*, **24**, 1–13.

Cowper-Sal·lari,R. *et al.* (2012) Breast cancer risk–associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat. Genet.*, **44**, 1191–1198.

Dekker,J. *et al.* (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.*, **14**, 390–403.

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

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.

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

Ernst,J. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–9.

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

Göndör,A. *et al.* (2008) High-resolution circular chromosome conformation capture assay. *Nat. Protoc.*, **3**, 303–313.

Harrow,J. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760–1774.

Langfelder,P. *et al.* (2008) Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*, **24**, 719–720.

Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442-451.

Maurano,M.T. *et al.* (2012) Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*, **337**, 1190–1195.

Roadmap Epigenomics Consortium *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.

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

Schaub,M.A. *et al.* (2012) Linking disease associations with regulatory information in the human genome. *Genome Res.*, **22**, 1748–1759.

Sheffield,N.C. *et al.* (2013) Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Res.*, **23**, 777–788.

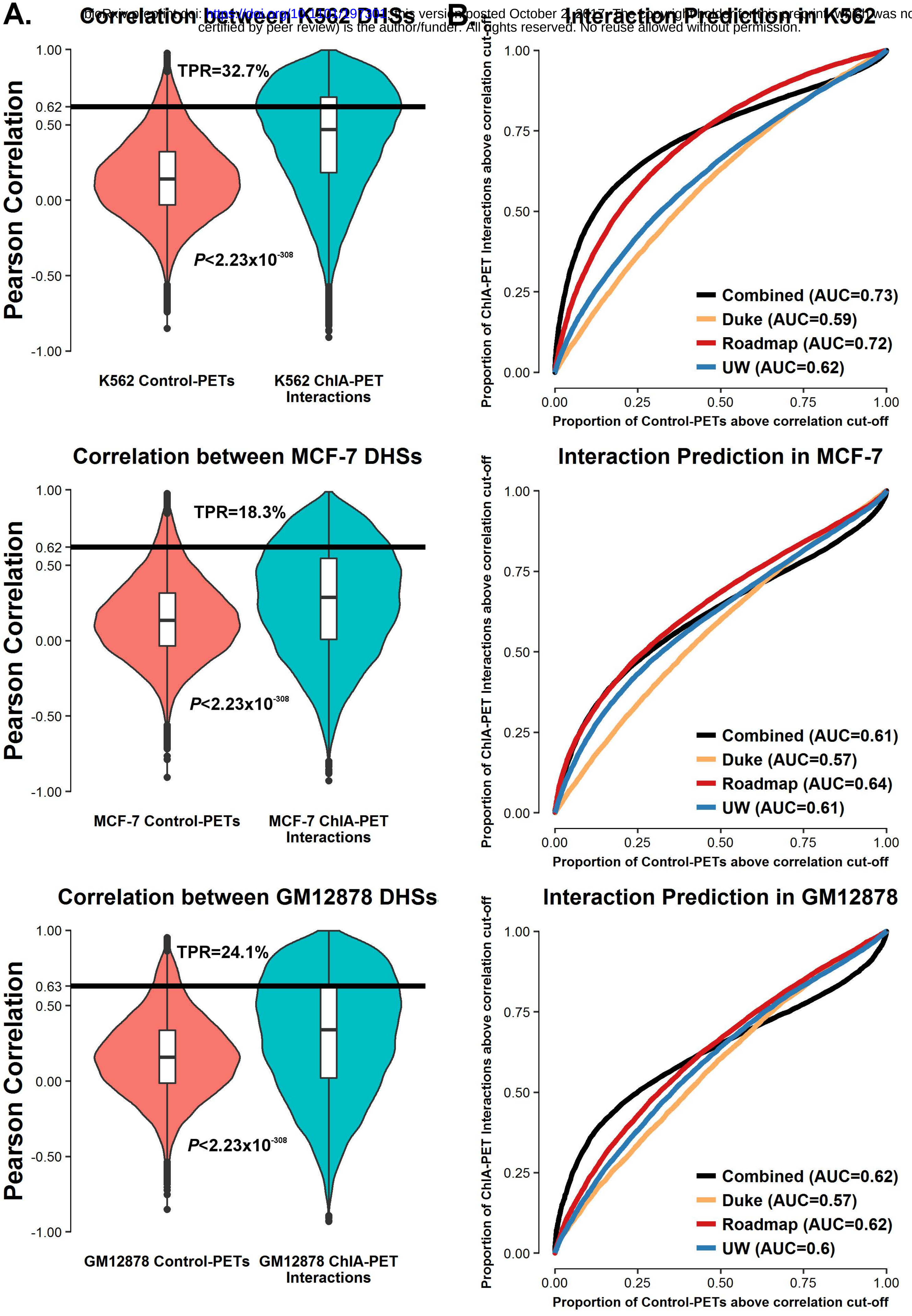
Tang,Z. *et al.* (2015) CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell*, **163**, 1611–1627.

Thurman,R.E. *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75–82.

van Berkum,N.L. *et al.* (2010) Hi-C: a method to study the three-dimensional architecture of genomes. *J. Vis. Exp.*, doi: (2010).10.3791/1869.

FIGURE LEGEND

Fig. 1. Performance of C3D on ChIA-PET and control datasets. (A) For the ChIA-PET data, C3D is applied to catalogues of K562, MCF-7 and GM12878 DHSs individually and calculates correlations across the combined panel of DNase-seq signals. Since some promoters and distal tags contain multiple DHSs, each paired-end tag was assigned the maximal correlation between a DHS in its promoter tag and one in its distal tag. For the control sets, C3D is applied to a merged catalogue of DHSs consisting of all DNase-seq peaks from the samples in the database. Each Control-PET received the correlation between a DHS in its promoter tag and a random one in its distal tag. For each cell-type, a Mann-Whitney U test compared the correlations of its interacting regions and control-PETs. TPRs are calculated at the correlation cut-offs that 5% of control-PETs are above. **(B)** C3D is run on the catalogues of K562, MCF-7 and GM12878 DHSs individually using four different databases: 79 samples from UW, 47 samples from Duke University, 53 samples from the Roadmap Epigenomics Project and a combined database of all 179 samples. For each of the three cell-types, the proportion of control-PETs and ChIA-PET connections classified as interactions are plotted at different correlation thresholds ranging from -1 to 1. AUCs are computed.


A.
Correlation between K562 DHSs
Pearson Correlation
1.00
0.62
0.50
0.00
-0.50
-1.00
TPR=32.7%
P<2.23x10^{-308}
K562 Control-PETs
K562 ChIA-PET Interactions
B.
Interaction Prediction in K562
Proportion of ChIA-PET Interactions above correlation cut-off
Proportion of Control-PETs above correlation cut-off
0.00
0.25
0.50
0.75
1.00
Combined (AUC=0.73)
Duke (AUC=0.59)
Roadmap (AUC=0.72)
UW (AUC=0.62)
Correlation between MCF-7 DHSs
Pearson Correlation
1.00
0.62
0.50
0.00
-0.50
-1.00
TPR=18.3%
P<2.23x10^{-308}
MCF-7 Control-PETs
MCF-7 ChIA-PET Interactions
Interaction Prediction in MCF-7
Proportion of ChIA-PET Interactions above correlation cut-off
Proportion of Control-PETs above correlation cut-off
0.00
0.25
0.50
0.75
1.00
Combined (AUC=0.61)
Duke (AUC=0.57)
Roadmap (AUC=0.64)
UW (AUC=0.61)
Correlation between GM12878 DHSs
Pearson Correlation
1.00
0.63
0.50
0.00
-0.50
-1.00
TPR=24.1%
P<2.23x10^{-308}
GM12878 Control-PETs
GM12878 ChIA-PET Interactions
Interaction Prediction in GM12878
Proportion of ChIA-PET Interactions above correlation cut-off
Proportion of Control-PETs above correlation cut-off
0.00
0.25
0.50
0.75
1.00
Combined (AUC=0.62)
Duke (AUC=0.57)
Roadmap (AUC=0.62)
UW (AUC=0.6)