A global high-density chromatin interaction network reveals functional long-range and trans-chromosomal relationships

Ruchi Lohia[1], Nathan Fox[1], Jesse Gillis[2]
1 Cold Spring Harbor Laboratory, 2 University of Toronto

Abstract

Chromatin contacts are essential for gene-expression regulation, however, obtaining a high-resolution genome-wide chromatin contact map is still prohibitively expensive owing to large genome sizes and the quadratic scale of pairwise data. Chromosome conformation capture (3C) based methods such as Hi-C have been extensively used to obtain chromatin contacts. However, since the sparsity of these maps increases with an increase in genomic distance between contacts, long-range or trans chromatin contacts are especially challenging to sample.

Here, we created a high density reference genome-wide chromatin contact map using a meta-analytic approach. We integrate 3600 Human, 6700 Mouse, and 500 Fly 3C experiments to create species-specific meta-3C contact maps with 304 billion, 193 billion, and 19 billion contacts in respective species. We validate that meta-3C are uniquely powered to capture functional chromatin contacts in both cis and trans. Unlike individual experiments, meta-3C gene contacts predict gene coexpression for long-range and trans chromatin contacts. Similarly, for long-range cis-regulatory interactions, meta-3C contacts outperform both all individual experiments, providing an improvement over the conventionally used linear genomic distance-based association. Assessing between species, we find patterns of chromatin contacts conservation in both cis and trans and strong associations with coexpression even in species for which 3C data is lacking.

We have generated an integrated chromatin interaction network which complements a large number of methodological and analytic approaches focused on improved specificity or interpretation. This high-depth “super-experiment” is surprisingly powerful in capturing long-range functional relationships of chromatin interactions, which are now able to predict coexpression, expression quantitative trait loci (eQTL), and cross-species relationships.

Main

Introduction

The physical associations generated by chromatin contacts are a critical factor to regulate and determine gene-expression patterns (Diament and Tuller 2019; Delaneau et al. 2019; Xu et al. 2020). Functional chromatin contacts can form across a wide range of genomic distances within a chromosome (cis) or across a chromosome (trans). Although trans contacts are non-random (Sarnataro et al. 2017) and there is evidence of trans-regulatory interactions (Dekker and Misteli 2015; Maass et al. 2019), studying the functional role of these interactions is difficult due to the high sparsity of available contact maps in trans.

Obtaining high-density contact maps at all genomic distances and in trans is not yet feasible with most existing maps being essentially probabilistic in nature, capturing some fraction of likely-present contacts in a distance-dependent manner. Genome-wide contact maps can be obtained using chromosome conformation capture (3C)-based technologies such as Hi-C (Lieberman-Aiden et al. 2009). However, due to large genome sizes and the quadratic scale of pairwise data, obtaining these maps at high resolution would require prohibitively expensive sequencing at even 1X depth in the pairwise space. Capturing long-range and trans

chromatin contacts is made more difficult since the frequency of contacts decreases with an increase in genomic distance between contacting loci in cis (Lieberman-Aiden et al. 2009). And in trans, the contacts are at least 2 orders of magnitude less frequent (Sarnataro et al. 2017) while also having a larger search space than cis.

To overcome the sequencing-depth barrier targeted 3C-based techniques such as ChiA-PET (Fang et al. 2016) and Capture-Hi-C (Dryden et al. 2014) are widely used to obtain high-resolution contacts maps for specific proteins or selected loci respectively. Alternatively, several in-silico methods have taken the advantage of existing limited resolution contact maps to either generate higher resolution maps using machine learning approaches (Fudenberg et al. 2020; Zhang et al. 2019; Zhang et al. 2018; Schwessinger et al. 2020) and/or detect statistically significant interactions by background fitting (Ay et al. 2014; Rao et al. 2014). However, with a few exceptions (Bulathsinghalage and Liu 2020; Xiong and Ma 2019), most of the available methods are only tested to enhance cis interactions because longer range interactions are essentially unavailable within any given data set.

In this work, we propose a meta-analysis approach where we leverage several hundreds of available CC-maps generated from 3C-based experiments to create a dense genome-wide CC-map for three species; Human, Mouse, and Fly. We show that these maps are valuable for capturing long-range and trans-chromosomal interactions. We evaluated the effectiveness of contact maps using three criteria; CC-maps were used to predict 1] gene-expression profiles, 2] target genes for eQTLs and 3] conservation across pairs of species (Human-Mouse, Human-Fly, and Mouse-Fly). Our reference networks complement a very diverse array of efforts in genomics, from those focused on more targeted experiments in 3C which now have an overall “null” with which to compare individual results, to genome interpretation methods, whether interpreting variants, expression patterning, or regulatory sequence.

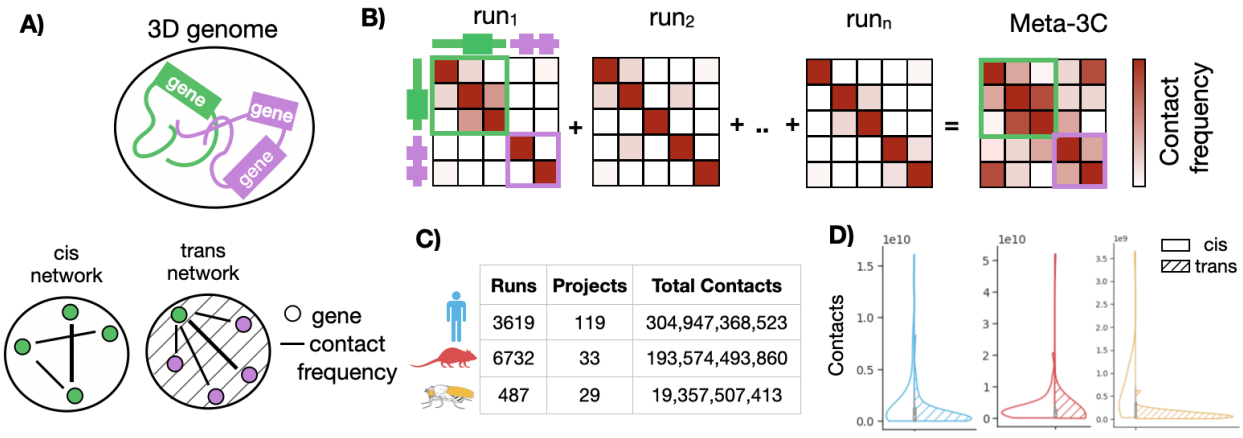

Fig 1: Creating meta-3C network. A) Genes are co-localized in chromatin 3D structure through frequent chromatin contacts. The structure can be represented with networks where nodes are genes and edges are interaction frequencies. The cis and trans networks consist of intra-chromosome and inter-chromosome edges respectively. B) Individual 3C experiments are aggregated to create a meta-3C network. C) Runs, projects, and contacts in Human, Mouse, and Fly meta-3C networks. D) Total contacts (sequencing depth) distribution across projects in cis and trans for each species.

Results

Meta-3C network predicts coexpression at greater resolution and scale than individual networks

In brief, for building the meta-3C network, we uniformly processed 3619, 6732, and 487 3C runs for Human, Mouse, and Fly respectively (Figure 1C). The runs were obtained after querying Sequence Read Archive (SRA) with field limitations of given species and Hi-C as experiment strategy. A genome-wide interaction matrix was created for each run after mapping the reads to the same reference genome for each species. Within SRA, all the runs (SRR) belonging to a study are grouped together as project (SRP). A project can consist of multiple runs, which can include biological or technical replicates across multiple tissues or cell types. All the interaction matrices within a project were aggregated to create a project-level aggregate. There were 119, 33, and 29 projects for Human, Mouse, and Fly respectively. The meta-3C map was created after further aggregating all processed runs within their respective species (Figure 1B). For subsequent analysis, the genome-wide contact matrix was mapped to genes (see Methods) to create networks where nodes are genes and edges are the interaction frequency between genes. The genome-wide networks were divided into cis and trans depending on if the edge connects two genes in the same chromosome or different chromosome respectively (Figure 1A). To validate the predictive power of the meta-3C network, we benchmarked it against networks inferred from individual projects for each species.

As our first performance test, we assessed the tendency for spatially co-localized to be co-expressed (Varrone et al. 2020; Babaei et al. 2015), using previously derived shared patterns of expression in independent data (Lee et al. 2020). The underlying hypothesis is that spatial proximity may be a useful way to organize regulatory relationships, as in the case of linear sequence, thus yielding shared spatial relationships for genes that are co-expressed. Thus, while perfect performance at predicting coexpression is not expected, the genome-wide scale of the assessment makes it useful for assessing cis and trans effects. For each gene, we measure the ability of interaction frequency to predict the gene's top 1% coexpression partners (Figure 2A). We call this measure "contact coexpression" and is expressed as an AUC (Area Under the ROC Curve) with possible values ranging between 0 and 1. A score of 1 indicates that interaction frequency perfectly predicts coexpression; 0.5 indicates no relationship. We evaluated the contact coexpression as a function of the sequencing depth of the 3C network (Figure 2B and Figure 2C). We find that performance is linearly dependent on the log of sequencing depth and meta-analysis provides additional coverage. We find that in cis the best powered individual experiments are close to the saturation depth that maximizes performance (Figure 2B), although performing substantially worse in trans. In trans, the meta-3C network acts like a "super-experiment", where the additional coverage fully converts into substantial additional performance (Figure 2C). We found similar results for Mouse (Figure S1) and Fly (Figure S2).

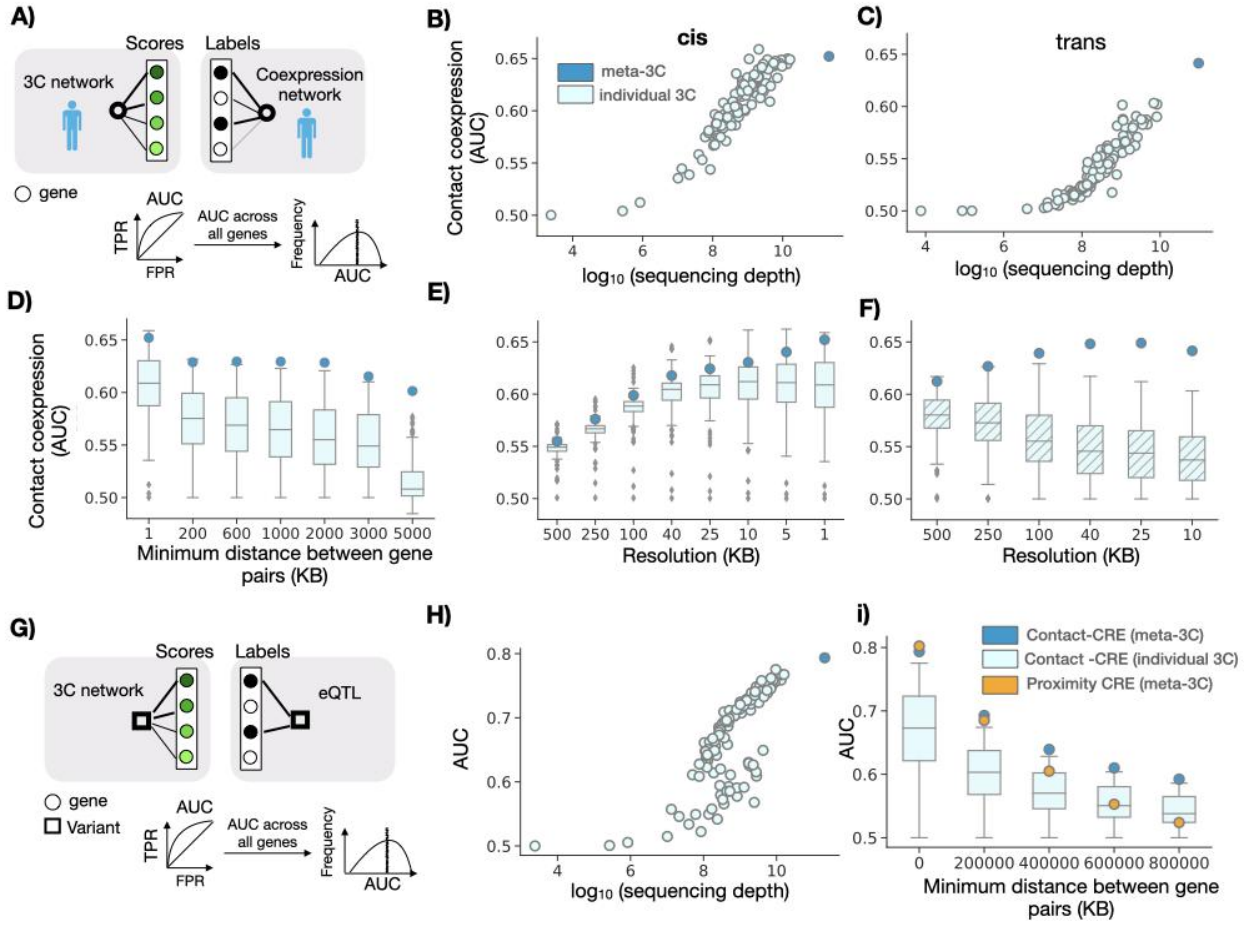

Fig 2: Meta-3C network benchmarking. A) Contact coexpression metric schematic. Circles represent genes and lines represent edges of that gene in respective networks. For each target gene, we use its ranked edges in 3C network to predict the top 1% of its edges in the coexpression network. We perform this task for every gene and then report the average across all genes. B) The circles are contact coexpression for individual and meta3C network in cis as a function of sequencing depth at 1KB resolution C) Same as B) but in trans and at 10KB resolution. D) The boxplot shows the distribution of contact coexpression in cis for each project at various distance thresholds. E) The boxplot shows the distribution of contact coexpression for each project at various resolutions in cis. Circles represent the performance of cis meta-3C network. F) Same as E) but using trans networks. G) Contact-CRE and proximity CRE metric schematic. For contact-CRE and proximity-CRE, for each variant, the edges are ranked by contact frequency or inverse of the genomic distance from the variant respectively. The labels are obtained from eQTL associations (Methods). We perform this task for every variant and then report the average across all variants. H) Contact-CRE for individual project and meta-3C network i) Contact-CRE and Proximity-CRE for meta-3C network, distribution across individual networks at various minimum distance thresholds.

In order to validate that the meta-3C network has more uniform coverage, we compared the contact coexpression of individual 3C networks and meta-3C networks at various linear distance thresholds in cis. We find that for long-range contacts meta-3C network performs better than every individual 3C network (Figure 2D). For both individual networks and meta-3C network, the performance decreases in the absence of short-range contacts. This could be due to a higher number of short-range regulatory interactions or due to the similarity of the chromatin environment for nearby genes.

The contact coexpression is dependent on the resolution of the 3C network used and therefore we compared the performance of individual 3C and meta-3C networks at various resolutions. We find that for the individual networks performance increases with an increase in resolution, plateaus, and then slightly falls off in cis (Figure 2D). In essence, improved resolution is useful in cis because the coverage is adequate for it to provide useful signal until the very finest resolution where most experiments begin to decline, although the meta-3C network continues to slightly increase, as might be expected. In contrast, in trans (Figure 2E) the performance monotonically falls with an increase in resolution for individual experiments. However, the in trans pattern for meta-3C networks strongly resembles that of individual experiments in cis, increasing and then plateauing with improvements in resolution (Figure 2D, Figure 2E). This suggests unlike individual networks, meta-3C networks are dense enough to be analyzed at high resolutions even in trans. We found similar results for Mouse (Figure S1) and Fly (Figure S2).

Meta-3C network effectively capture more eQTL interactions

For our second performance assessment, we tested the hypothesis that genetic variants (eVariant) regulate gene expression of the target gene (eGene) via physical contact (Wang et al. 2021; Kong and Jung 2020). The set of eQTLs was obtained from GTEx (Methods). For each eVariant, the interaction frequency with all genes falling in unique contact map bins at 1KB resolution was used to predict the eGene (Figure 2G). This is termed “contact-CRE’ where CRE stands for cis-regulatory elements and is expressed as an AUC with possible values ranging between 0-1, with 1 and 0 meaning that the eVariant and target eGenes have the highest and lowest interaction frequency respectively when compared to all eVariant and non-eGenes interactions. Similar to the previous benchmarking test we find that performance is linearly dependent on the log of sequencing depth and meta-analysis provides additional coverage; meta-3C network has higher performance when compared to any of the individual networks (Figure 2H). This emphasizes the significance of dense contact networks in identifying regulatory interactions.

We further evaluated the ability of meta-3C networks to predict target genes for variants by comparing their performance with a linear genomic distance-based predictor, the current standard approach. The distance between the variant and gene transcription start site (TSS) remains almost the only metric widely used to annotate target genes for variants (Stacey et al. 2019). For each eVariant the inverse of linear distance (1/TSS) with all genes is ranked and then used to predict the eGene (Figure 2G). This is termed “proximity-CRE’ and is expressed as an AUC with possible values ranging between 0-1, with 1 and 0 meaning that eGenes are the closest and farthest from the eVariant respectively. We compared contact-CRE of individual 3C networks, and meta-3C networks at various linear distance thresholds (Figure 2I). We reassuringly find that meta-3C network outperforms individual networks at all distance thresholds. Furthermore, the performance for both contact-CRE and proximity-CRE decreases in the absence of short-range contacts. This is in agreement with our previous observation where we find that contact coexpression decreases in the absence of short-range contacts. Interestingly, we find that for contact-CRE outperforming proximity-CRE is easier at larger distance thresholds suggesting that indeed physical contact is one of the critical factors for long-range regulatory contacts.

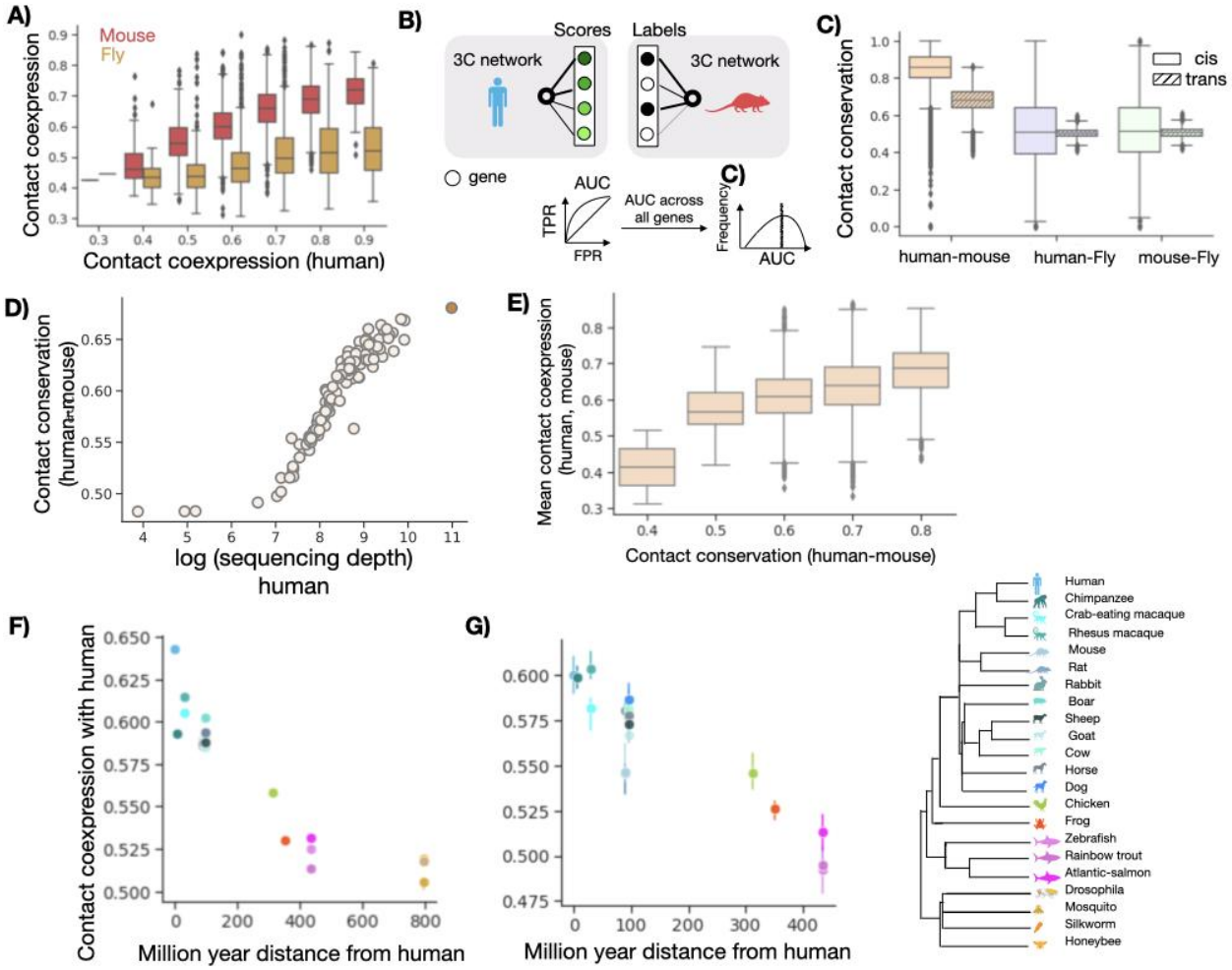

Fig 3: Chromatin contacts are conserved across species in both cis and trans. A) Contact coexpression in trans for 1:1 orthologs in Human-Mouse and Human-Fly. B) Contact conservation schematic. For each gene ranked edges in Human 3C network are used to predict the top 10% of Mouse 3C network edges. This task is repeated for each gene and in both directions and we report the average AUC. C) The distribution of contact conservation score using meta-3C network for various pairs of species. D) Human-Mouse contact conservation for various projects and meta-3C network as a function of sequencing depth. E) Avg contact coexpression score for each gene in Human and Mouse as a function of contact conservation. F) The median performance across all genes when the Human meta-3C network is used to predict coexpression across other species. G) Same as F) but only using the same set of ortholog genes across species. The error bars represent a 68% confidence interval.

Trans-chromosomal chromatin contacts show evolutionary conservation

Having established that meta-3C networks are well powered to capture meaningful contacts, we now use them to study the conservation of genomic contacts between species. Since chromatin contacts regulate gene expression, it is reasonable to expect some conservation of these contacts across species even in the context of large scale genomic alteration and, in the reverse, divergence in contacts across species can help explain regulatory evolution (Eres et al. 2019; Krefting et al. 2018). We evaluated the conservation of contacts across species in three different ways; we compare the contact coexpression scores for ortholog genes in each species pair, we use the 3C network of one species to predict either 3C network (“contact conservation”) or coexpression network in another species.

Before directly comparing the contact map across species we first compared the contact coexpression scores for 1:1 orthologous genes across species. We find a strong linear relationship between Human and Mouse scores and a somewhat weaker relationship between Human and Fly scores in both cis (Figure S3A) and trans (Figure 3A). This suggests that if a gene is spatially co-regulated in one species, it is likely to be spatially co-regulated across other species.

We next characterized the degree to which gene contacts are conserved by directly comparing the meta-3C network across species (Figure 3B). Each gene's shared neighborhood is defined by ranking all edges in the chromatin contact network and then using it to predict the gene's top 10% of edges in another species. We call this "contact conservation" and again, treat it as a prediction task with 1 meaning perfect contact conservation, 0.5 consistent with random reordering of neighborhoods, and 0 meaning that contacting partners have reversed. For trans conservation score, only the trans gene pairs in both species are used, similarly for cis analysis. As expected, we find that the contact conservation is higher for Human-Mouse (AUC > 0.8) when compared with Human-Fly (AUC 0.5) or Mouse-Fly (AUC 0.5) (Figure 3C) in both cis and trans. We also find that genes with high contact conservation in Human-Mouse are likely to have high contact coexpression.

We also re-validated the power of the meta-3C network: we compared the 'contact conservation' scores for individual and meta-3C networks at various resolutions. We reassuringly find that the meta-3C network outperforms individual projects at high resolutions in both cis and trans (Figure 3D, S3B, S3C). This again suggests that the meta-3C network is efficient in capturing chromatin contacts when compared to individual networks. Although the conservation of cis chromatin structure across species is not surprising and is evident in the presence of syntenic regions between species, the conservation of trans-chromatin contacts is noteworthy. It suggests that the trans-chromatin structure is likely selected for preservation to maintain function.

We further investigated the evolution of trans chromatin contacts in Human by comparing the degree to which the Human contacts can predict coexpression across several species. This method allowed us to extend our analysis to species for which the meta-3C network is not available. Each gene's neighborhood is defined by ranking all edges in the chromatin contact network of one species and then used to predict the gene's top 10% of coexpressed gene pairs in another species. We call this "contact coexpression conservation" and calculate the AUC as above. When contact coexpression conservation is plotted along with the phylogenetic distance across species, we find that the performance decreases with an increase in phylogenetic distance using both cis and trans meta-3C networks (Figure 3F). This suggests that the contacts diverge as the species pair becomes distant across evolution. The number of 1:1 orthologs also decreases with an increase in the phylogenetic distance and it seemed possible that our observation was dominated by the number of ortholog pairs between species. To eliminate this possibility, we redid our analysis but using only the same set of ortholog genes (429 genes) in each species and our result persisted (Figure 3G). Species more than 100 million years of distance (mya) from Humans have stronger divergence in contacts when compared to species within the Mammalia Class. The species included in Figure 3G were limited to the Chordata phylum to ensure a reasonable number of genes in the analysis.

Data availability and Online Tool

In order to facilitate the broad adoption of meta-3c by the community, we have made data available via an online tool. Contact data can be obtained in two ways: a] Network download: Direct download of the desired resolution, species meta-3C contact matrix in cis or trans available at https://labshare.cshl.edu/shares/gillislab/resource/HiC in HiCMatrix format (https://github.com/deeptools/HiCMatrix). b] Gene vector download: Contact frequency with every genomic loci at the chosen resolution and for any desired gene found in the respective species (Fig 4a). The downloaded file is in bed file format which can be uploaded to UCSC genome browser for further analysis as desired (Fig4 b).

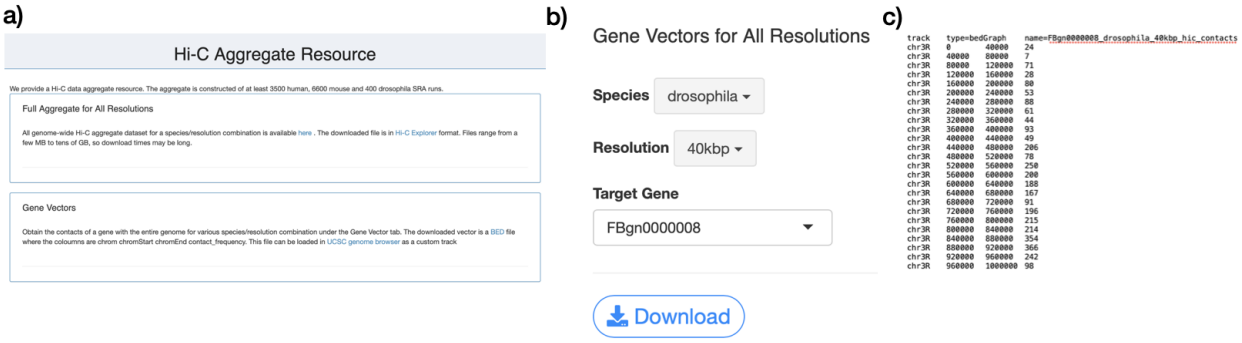

Fig4: Snapshot of the tool page (**a**) gene vector download page (**b**). **c)** Downloaded bed file snapshot where the first column is chromosome name, next two columns are genomic bin and the last column is raw contact frequency.

Discussion

In this work, we created a high-depth, genome-wide chromatin contact map using a meta-analytic approach, validated it, and further used it to reveal chromatin structure to function relationships. We find that for the three species analyzed in this study (Human, Mouse, and Fly), chromatin contacts strongly predicted the coexpression of genes. We also show that chromatin contacts are better than linear proximity for predicting eQTLs when high-resolution chromatin contact data is available. Our results persist even when only long-range chromatin contacts are analyzed. Additionally, we find that trans chromosomal contacts show evidence of conservation across species.

Meta-3C networks are an effective means for capturing otherwise hard to characterize long-range interactions providing potentially uniquely important practical applications. One important application for a wide area of genomics is their ability to prioritize distant target genes for variants. We expect these networks to be powerful training data for future machine learning attempts to predict chromosomal contacts, an important area of ongoing research (Fudenberg et al. 2020; Zhang et al. 2019; Zhang et al. 2018; Schwessinger et al. 2020). Additionally, meta-3C networks can be used with other cell-type-specific 'omics datasets such as ChIP-Seq to reveal cell-type-specific enhancer-promoter contacts. Previously, Nasser et al (Nasser et al. 2021; Fulco et al. 2019) used averaged Hi-C data across 10 cell-types in their ABC model to accurately make cell-type-specific enhancer–gene predictions. Thus, the continuing evolution of methods with improved specificity is likely to complement our better-powered but less condition-specific meta-analytic approach.

Within 3C analysis, and even outside of it, aggregation of data is well appreciated to be a useful strategy. Reproducible biological replicates within the same study are often combined to increase the density of 3C data thereby capturing more interactions (Won et al. 2016; Rao et al. 2014). Our approach can be thought of as the most extreme version of this idea, combining experiments as broadly as possible to capture statistical relationships that are common. This is most useful if the depth is a major limitation, as in trans contacts, as it comes with the cost of a loss of condition-specificity. Thus, the route forward for the field as a whole will doubtless involve both improved specificity, integration, and interpretive methods.

In summary, our study sheds new light on the functional role of long-range and trans chromosomal contacts and provides a critical resource for use by a wide range of genomics research.

Methods

3C Data Sources and processing pipeline

The 3C data for each species were obtained from SRA search (https://www.ncbi.nlm.nih.gov/sra/) with the field limitations of {"Organism": ["Homo sapiens", "Mus musculus", "Drosophila melanogaster"], "Strategy": "hi c"}. We found 3913, 8431, 502 samples for Human, Mouse, and Fly respectively. We also added 268, 17, and 25 samples manually that were labeled OTHER in SRA, but were deemed to be valid 3C data based on publication details. After manual additions, filtering out Runs without available restriction enzyme information, and excluding Runs that failed processing, we had 3621, 6733, 487 samples for Human, Mouse, and Fly. In total, we aggregated 119 Human Projects, 33 Mouse Projects, and 29 Fly Projects (Table S1, S2, and S3).

All samples were reprocessed from short read sequence data to reduce differential computational noise across experiments. Restriction enzymes were identified for each sample from the literature. SRA files were downloaded using prefetch, then converted to paired FASTQ files using fasterq-dump. FASTQ files were processed using the HiCUP tool (Wingett et al. 2015), with the alteration that short reads were aligned using the STAR aligner, instead of the default Bowtie2. HiCUP truncates the reads based on restriction site, aligns them, and filters artifactual and duplicated data. Reads were aligned to the hg38, mm10, and dm6 genomes. Output SAM files were converted to indexed and compressed Pairs files using the bam2pairs tool. Finally, pairwise chromosome-chromosome contact matrices were generated at single base-pair resolution.

Building CC-maps:

To obtain CC-maps, each chromosome is divided up into "bins" of a specific size. The number of base pairs in each bin represents the "resolution" of the matrix. The contact frequency for each bin pair is obtained by summing the reads falling in that bin.

CC-maps were generated at 8 resolutions (1KB, 5KB, 10KB, 25KB, 40KB, 100KB, 250KB, and 500KB) in cis for all species and trans for only Fly. For Human and Mouse, trans CC-map at 1KB and 5KB resolutions were not processed due to high memory requirements (more than 2TB). These files were written in HiCMatrix (https://github.com/deeptools/HiCMatrix) h5 format. For each species, we excluded sex chromosomes and considered only autosomes (Human: chr1 to chr22, Mouse:chr1 to chr19 and Fly: chr2L, chr2R, chr3L, chr3R, chr4).

The contact frequency of each genomic pair coordinate was summed across runs to generate project-level CC-maps. The sequencing depth of a project in cis and trans is obtained by summing all the contacts in cis and trans respectively.

The contact frequency was KR-normalized separately for the cis and trans networks to adjust for nonuniformities in coverage introduced due to experimental bias (Knight and Ruiz 2012) using hicCorrectMatrix tool of HiCexplorerV3.6 (Wolff et al. 2020).

All project level CC-maps within each species were further summated to create species level meta-3C maps.

To determine contact frequency between each gene pair we use the maximum of contact frequency between each bin in which genes reside. Gene TSS and TES were used to determine the bins in which the gene resides. List of genes, TSS, and TES were obtained as GTF files from ENSEMBEL (Spetember 2019).

A list of 1:1 orthologs for pair of species was obtained from OrthoDB (Kriventseva et al. 2019). Species diverge time were sourced from Timetree (Kumar et al. 2017).

Coexpression data

The coexpression network used in this study is a 'high confidence gene' aggregated coexpression network generated using the method previously described in CoCoCoNet (Lee et al. 2020). In brief, several bulk RNA-seq datasets were obtained from NCBI's SRA database (unique SRA Study IDs). Networks for each dataset are built by calculating the Spearman correlation between all pairs of genes, then ranking the

correlation coefficients for all gene-gene pairs, with NAs assigned the median rank. Each network is then rank standardized and normalized by dividing through by the maximum rank. Aggregate networks are then generated by averaging rank standardized networks from individual datasets.

eQTL data source and processing

A list of tissue-specific 'significant' variant gene pair associations and 'all' variant gene pair associations (including non-significant associations) across 54 tissues along with the distance between the variant and gene TSS (at bp resolution) were obtained from GTEx Portal v8 at https://gtexportal.org. Since the meta-3C network is not tissue-specific, we combined the data across tissues to generate a set of unique 'significant' and 'all' variant gene pair associations. To obtain a list of 'non-significant' gene pair associations, 'significant' variant gene pair associations were removed from 'all' variant gene pair associations data. All variants in the coding regions and up to 1KB of any gene TSS and TES were removed. For performance score 1KB cis CC-map is used and for each eVariant only genes in unique bins are tested.

Data Availability

The Meta 3C networks for Human, Mouse, and Fly are available for download from online tool (https://gillisweb.cshl.edu/HiC/) or direct download (https://labshare.cshl.edu/shares/gillislab/resource/HiC/)

Supplementary

Fig S1: B) C) D) and E) of Fig2 for Mouse
Fig S2: B) C) D) and E) of Fig2 for Fly
Fig S3: Fig3 B) in cis and crosss-species conservation at various resolutions

Table:
Table S1: Table for each SRA project, number of samples, contact counts, AUC in cis and trans (https://labshare.cshl.edu/shares/gillislab/resource/HiC/human_aggregates_summary.csv)
Table S2: Same as S1 but for Mouse: (https://labshare.cshl.edu/shares/gillislab/resource/HiC/mouse_aggregates_summary.csv)
Table S3: Same as S1 but for Fly (https://labshare.cshl.edu/shares/gillislab/resource/HiC/drosophila_aggregates_summary.csv)

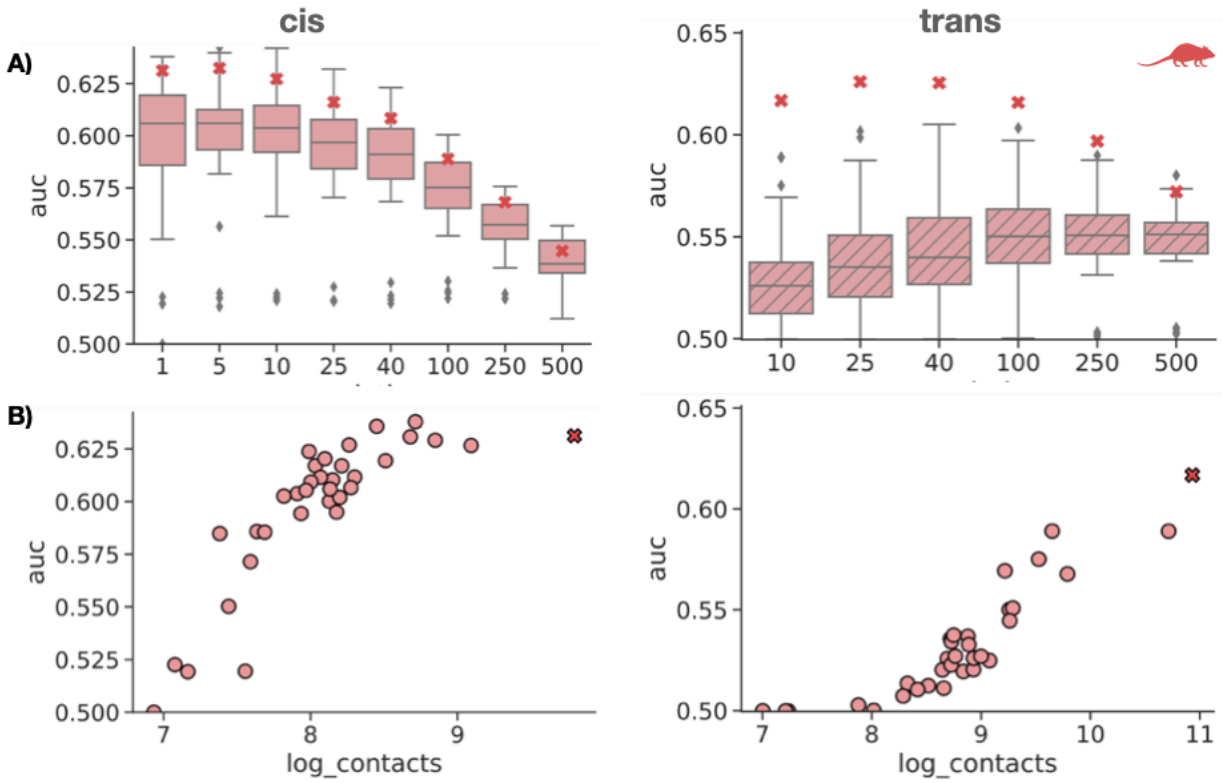

Fig S1: A) The boxplot shows the distribution of median AUC across all genes for each project at various resolutions in cis (left) and trans (right). B) The circles represent the median AUC across all genes for each project vs sequencing depth at 1KB resolution in cis (left) and 10KB resolution in trans (right).

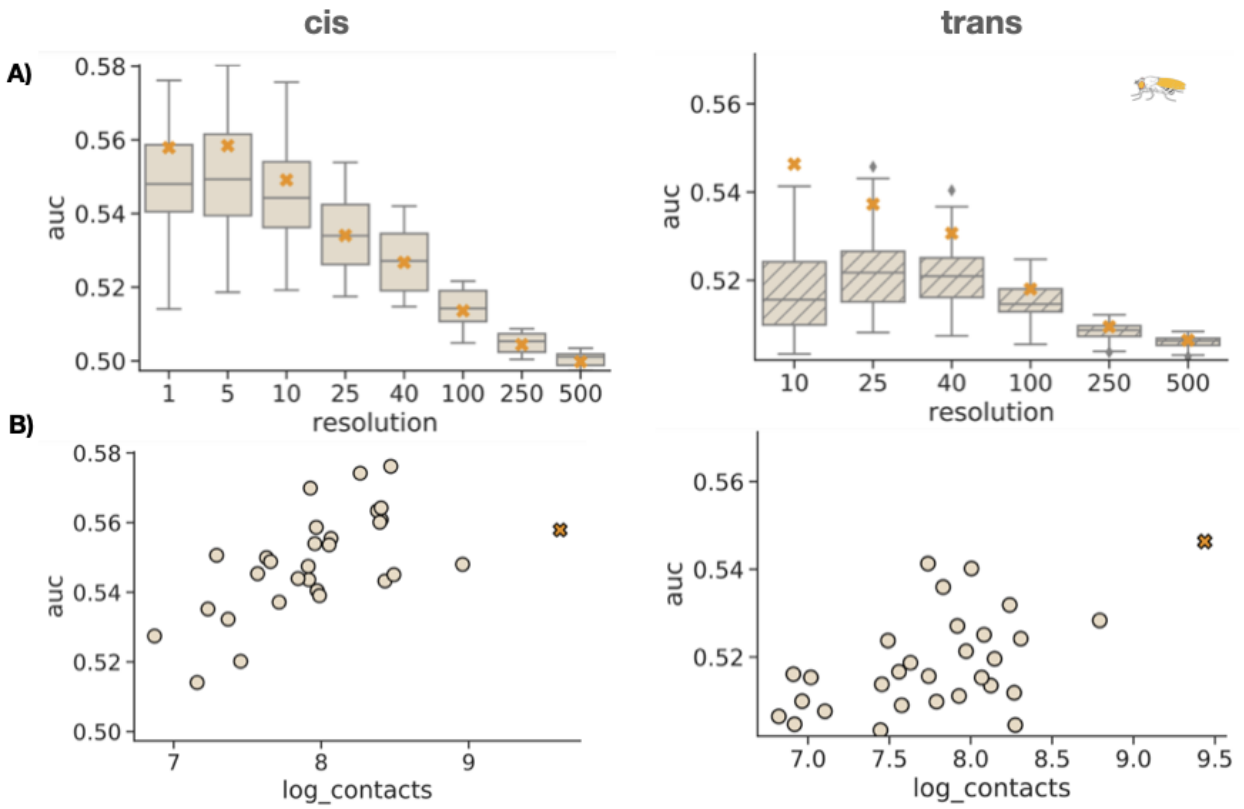

Fig S2: A) The boxplot shows the distribution of median AUC across all genes for each project at various resolutions in cis (left) and trans (right). B) The circles represent the median AUC across all genes for each project vs sequencing depth at 1KB resolution in cis (left) and 10KB resolution in trans (right).

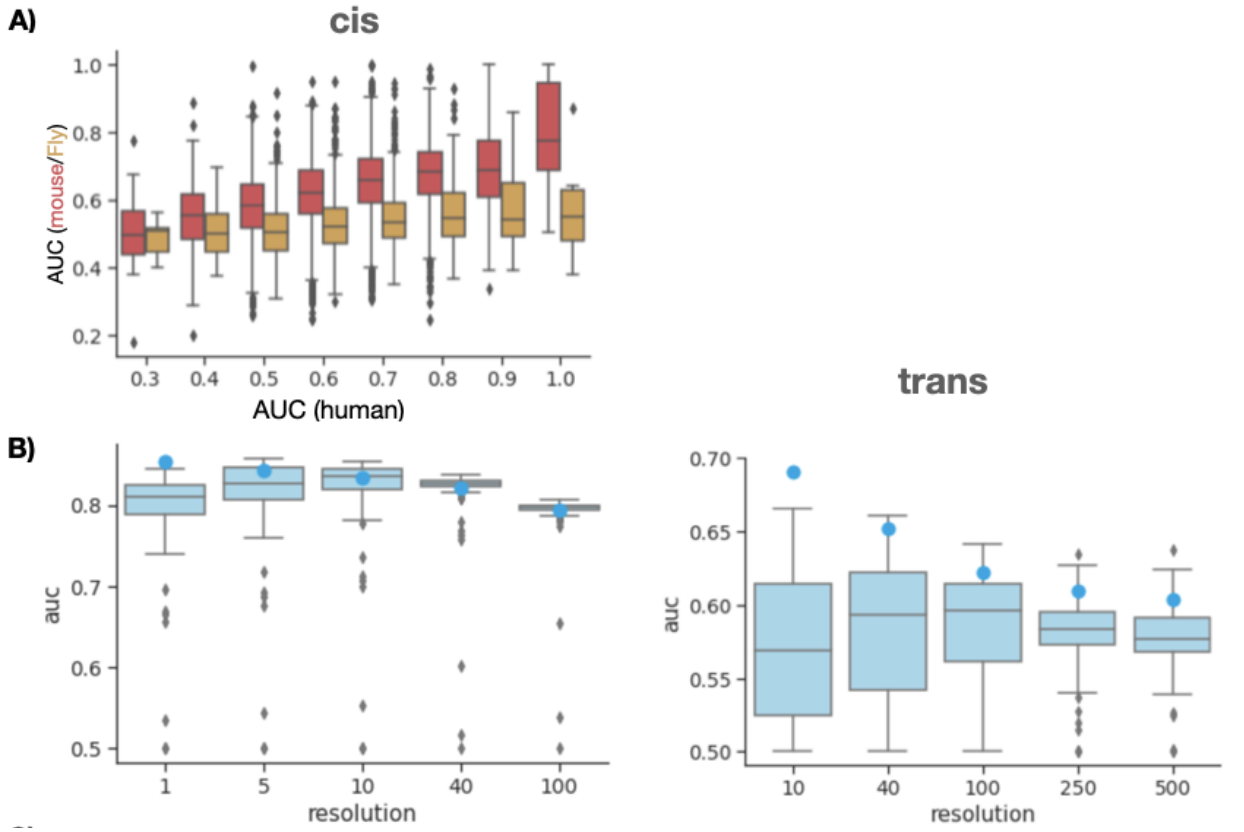

Fig S3: Chromatin contacts are conserved across species. A) Performance of gene contact to predict coexpression in Human vs other species (Mouse and Fly) in cis. B) The boxplot shows the distribution of median AUC across all genes for each project at various resolutions in cis (left) and trans (right).

References

Ay, Ferhat, Timothy L. Bailey, and William Stafford Noble. 2014. "Statistical Confidence Estimation for Hi-C Data Reveals Regulatory Chromatin Contacts." *Genome Research* 24 (6): 999–1011.

Babaei, Sepideh, Ahmed Mahfouz, Marc Hulsman, Boudewijn P. F. Lelieveldt, Jeroen de Ridder, and Marcel Reinders. 2015. "Hi-C Chromatin Interaction Networks Predict Co-Expression in the Mouse Cortex." Edited by Guillaume Joseph Filion. *PLoS Computational Biology* 11 (5): e1004221.

Bulathsinghalage, Chanaka, and Lu Liu. 2020. "Network-Based Method for Regions with Statistically Frequent Interchromosomal Interactions at Single-Cell Resolution." *BMC Bioinformatics* 21 (Suppl 14): 369.

Dekker, Job, and Tom Misteli. 2015. "Long-Range Chromatin Interactions." *Cold Spring Harbor Perspectives in Biology* 7 (10): a019356.

Delaneau, O., M. Zazhytska, C. Borel, G. Giannuzzi, G. Rey, C. Howald, S. Kumar, et al. 2019. "Chromatin Three-Dimensional Interactions Mediate Genetic Effects on Gene Expression." *Science* 364 (6439). https://doi.org/10.1126/science.aat8266.

Diament, Alon, and Tamir Tuller. 2019. "Modeling Three-Dimensional Genomic Organization in Evolution and Pathogenesis." *Seminars in Cell & Developmental Biology* 90 (June): 78–93.

Dryden, Nicola H., Laura R. Broome, Frank Dudbridge, Nichola Johnson, Nick Orr, Stefan Schoenfelder, Takashi Nagano, et al. 2014. "Unbiased Analysis of Potential Targets of

Breast Cancer Susceptibility Loci by Capture Hi-C." *Genome Research* 24 (11): 1854–68.
Eres, Ittai E., Kaixuan Luo, Chiaowen Joyce Hsiao, Lauren E. Blake, and Yoav Gilad. 2019. "Reorganization of 3D Genome Structure May Contribute to Gene Regulatory Evolution in Primates." *PLoS Genetics* 15 (7): e1008278.
Fang, Rongxin, Miao Yu, Guoqiang Li, Sora Chee, Tristin Liu, Anthony D. Schmitt, and Bing Ren. 2016. "Mapping of Long-Range Chromatin Interactions by Proximity Ligation-Assisted ChIP-Seq." *Cell Research* 26 (12): 1345–48.
Fudenberg, Geoff, David R. Kelley, and Katherine S. Pollard. 2020. "Predicting 3D Genome Folding from DNA Sequence with Akita." *Nature Methods* 17 (11): 1111–17.
Fulco, Charles P., Joseph Nasser, Thouis R. Jones, Glen Munson, Drew T. Bergman, Vidya Subramanian, Sharon R. Grossman, et al. 2019. "Activity-by-Contact Model of Enhancer-Promoter Regulation from Thousands of CRISPR Perturbations." *Nature Genetics* 51 (12): 1664–69.
Knight, Philip A., and Daniel Ruiz. 2012. "A Fast Algorithm for Matrix Balancing." *IMA Journal of Numerical Analysis* 33 (3): 1029–47.
Kong, Nahyun, and Inkyung Jung. 2020. "Long-Range Chromatin Interactions in Pathogenic Gene Expression Control." *Transcription* 11 (5): 211–16.
Krefting, Jan, Miguel A. Andrade-Navarro, and Jonas Ibn-Salem. 2018. "Evolutionary Stability of Topologically Associating Domains Is Associated with Conserved Gene Regulation." *BMC Biology* 16 (1): 87.
Kriventseva, Evgenia V., Dmitry Kuznetsov, Fredrik Tegenfeldt, Mosè Manni, Renata Dias, Felipe A. Simão, and Evgeny M. Zdobnov. 2019. "OrthoDB v10: Sampling the Diversity of Animal, Plant, Fungal, Protist, Bacterial and Viral Genomes for Evolutionary and Functional Annotations of Orthologs." *Nucleic Acids Research* 47 (D1): D807–11.
Kumar, Sudhir, Glen Stecher, Michael Suleski, and S. Blair Hedges. 2017. "TimeTree: A Resource for Timelines, Timetrees, and Divergence Times." *Molecular Biology and Evolution* 34 (7): 1812–19.
Lee, John, Manthan Shah, Sara Ballouz, Megan Crow, and Jesse Gillis. 2020. "CoCoCoNet: Conserved and Comparative Co-Expression across a Diverse Set of Species." *Nucleic Acids Research* 48 (W1): W566–71.
Lieberman-Aiden, Erez, Nynke L. Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome." *Science* 326 (5950): 289–93.
Maass, Philipp G., A. Rasim Barutcu, and John L. Rinn. 2019. "Interchromosomal Interactions: A Genomic Love Story of Kissing Chromosomes." *The Journal of Cell Biology* 218 (1): 27–38.
Nasser, Joseph, Drew T. Bergman, Charles P. Fulco, Philine Guckelberger, Benjamin R. Doughty, Tejal A. Patwardhan, Thouis R. Jones, et al. 2021. "Genome-Wide Enhancer Maps Link Risk Variants to Disease Genes." *Nature* 593 (7858): 238–43.
Rao, Suhas S. P., Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, et al. 2014. "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping." *Cell* 159 (7): 1665–80.
Sarnataro, Sergio, Andrea M. Chiariello, Andrea Esposito, Antonella Prisco, and Mario Nicodemi. 2017. "Structure of the Human Chromosome Interaction Network." *PloS One* 12 (11): e0188201.
Schwessinger, Ron, Matthew Gosden, Damien Downes, Richard C. Brown, A. Marieke Oudelaar, Jelena Telenius, Yee Whye Teh, Gerton Lunter, and Jim R. Hughes. 2020. "DeepC: Predicting 3D Genome Folding Using Megabase-Scale Transfer Learning." *Nature Methods* 17 (11): 1118–24.
Stacey, David, Eric B. Fauman, Daniel Ziemek, Benjamin B. Sun, Eric L. Harshfield, Angela M.

Wood, Adam S. Butterworth, Karsten Suhre, and Dirk S. Paul. 2019. “ProGeM: A Framework for the Prioritization of Candidate Causal Genes at Molecular Quantitative Trait Loci.” *Nucleic Acids Research* 47 (1): e3.

Varrone, Marco, Luca Nanni, Giovanni Ciriello, and Stefano Ceri. 2020. “Exploring Chromatin Conformation and Gene Co-Expression through Graph Embedding.” *Bioinformatics* 36 (Supplement_2): i700–708.

Wang, Hao, Jiaxin Yang, Yu Zhang, and Jianrong Wang. 2021. “Discover Novel Disease-Associated Genes Based on Regulatory Networks of Long-Range Chromatin Interactions.” *Methods* 189 (May): 22–33.

Wingett, Steven, Philip Ewels, Mayra Furlan-Magaril, Takashi Nagano, Stefan Schoenfelder, Peter Fraser, and Simon Andrews. 2015. “HiCUP: Pipeline for Mapping and Processing Hi-C Data.” *F1000Research* 4 (November): 1310.

Wolff, Joachim, Leily Rabbani, Ralf Gilsbach, Gautier Richard, Thomas Manke, Rolf Backofen, and Björn A. Grüning. 2020. “Galaxy HiCExplorer 3: A Web Server for Reproducible Hi-C, Capture Hi-C and Single-Cell Hi-C Data Analysis, Quality Control and Visualization.” *Nucleic Acids Research* 48 (W1): W177–84.

Won, Hyejung, Luis de la Torre-Ubieta, Jason L. Stein, Neelroop N. Parikshak, Jerry Huang, Carli K. Opland, Michael J. Gandal, et al. 2016. “Chromosome Conformation Elucidates Regulatory Relationships in Developing Human Brain.” *Nature* 538 (7626): 523–27.

Xiong, Kyle, and Jian Ma. 2019. “Revealing Hi-C Subcompartments by Imputing Inter-Chromosomal Chromatin Interactions.” *Nature Communications* 10 (1): 5069.

Xu, Hang, Shijie Zhang, Xianfu Yi, Dariusz Plewczynski, and Mulin Jun Li. 2020. “Exploring 3D Chromatin Contacts in Gene Regulation: The Evolution of Approaches for the Identification of Functional Enhancer-Promoter Interaction.” *Computational and Structural Biotechnology Journal*. Elsevier B.V. https://doi.org/10.1016/j.csbj.2020.02.013.

Zhang, Shilu, Deborah Chasman, Sara Knaack, and Sushmita Roy. 2019. “In Silico Prediction of High-Resolution Hi-C Interaction Matrices.” *Nature Communications* 10 (1): 5449.

Zhang, Yan, Lin An, Jie Xu, Bo Zhang, W. Jim Zheng, Ming Hu, Jijun Tang, and Feng Yue. 2018. “Enhancing Hi-C Data Resolution with Deep Convolutional Neural Network HiCPlus.” *Nature Communications 2018 9:1* 9 (1): 1–9.