Atlas of nascent RNA transcripts reveals enhancer to gene linkages

Rutendo F. Sigauke[1], Lynn Sanford[1], Zachary L. Maas[1,2], Taylor Jones[1], Jacob T. Stanley[1], Hope A. Townsend[1,3], Mary A. Allen[1], Robin D. Dowell[1,2,3*]

[1*]BioFrontiers Institute, University of Colorado Boulder, 3415 Colorado Ave., UCB 596, Boulder, 80309, CO, USA.
[2]Computer Science, University of Colorado Boulder, 1111 Engineering Drive, UCB 430, Boulder, 80309, CO, USA.
[3]Molecular, Cellular and Developmental Biology, University of Colorado Boulder, 1945 Colorado Ave, UCB 347, Boulder, 80309, CO, USA.

*Corresponding author(s). E-mail(s): robin.dowell@colorado.edu;
Contributing authors: rutendo.sigauke@colorado.edu;
lynn.sanford@colorado.edu; zachary.maas@colorado.edu;
taylor.jones-1@colorado.edu; jacob.stanley@colorado.edu;
hope.townsend@colorado.edu; mary.a.allen@colorado.edu;

Abstract

Gene transcription is controlled and modulated by regulatory regions, including enhancers and promoters. These regions are abundant in unstable, non-coding bidirectional transcription. Using nascent RNA transcription data across hundreds of human samples, we identified over 800,000 regions containing bidirectional transcription. We then identify highly correlated transcription between bidirectional and gene regions. The identified correlated pairs, a bidirectional region and a gene, are enriched for disease associated SNPs and often supported by independent 3D data. We present these resources as an SQL database which serves as a resource for future studies into gene regulation, enhancer associated RNAs, and transcription factors.

Keywords: nascent RNA sequencing, bidirectional transcription, enhancer RNA

Introduction

Transcription is a regulated process that is critical for cellular identity, differentiation, and response to the environment. Nascent transcription assays provide insight into transcription by measuring RNAs prior to their maturation into messenger RNA. In particular, run-on assays such as global run-on sequencing (GRO-seq) and precision run-on sequencing (PRO-seq) measure RNA as it is being produced by incorporating a marked nucleotide and selectively precipitating the labeled RNA. In this manner, the activity of cellular polymerases can be precisely measured.

Mammalian transcription initiation is predominantly bidirectional, with two oppositely oriented distinct transcription start sites in close proximity[1, 2]. These bidirectional signatures are observed at not only protein-coding genes but also transcribed regulatory regions (TREs)[2–4]. At genes, the upstream antisense RNA (uaRNA) is also referred to as a promoter upstream transcript (PROMPTs)[5–7]. At regulatory regions, the two transcripts are often referred to as enhancer associated RNAs (eRNAs)[7–10]. In every case, these non-gene transcripts are lowly transcribed, unstable, and not annotated. Hence numerous methods have been developed to identify sites of bidirectional transcription directly from run-on data[4, 11–14].

The function of these transcripts remains hotly debated, but regardless of their function, they have been shown to serve as excellent markers of regulation within the local genomic context. Not all instances of transcription factor binding result in altered gene regulation, however signatures of RNA polymerase activity near transcription factor binding sites effectively reflects the active subset of binding events[15–18]. In support of this notion, changes in bidirectional transcription activity (locations and levels) can be utilized to infer changes in transcription factor activity between two conditions[8, 18–24].

Such activity at regulatory regions can be essential for understanding changes in transcription at associated target genes. However, while some TREs have been definitively linked to target genes, the majority of these linkages are unknown. Enhancer regions are more distal to genes than promoters, which complicates the process of correctly assigning them to their targets. Yet transcription levels at regulatory regions clearly correlate with transcription levels at their associated target gene[18, 25]. In fact, correlated transcription has recently been used in small collections of data to link enhancers to their target genes[26, 27]. This suggests that patterns of correlation maintained over a large collection of nascent transcription data would identify enhancer to target gene pairings, which could then be used to decipher regulatory variation and refine our understanding of gene transcription regulation.

We sought to collect a repository of published run-on sequencing data from which we could catalog and characterize sites of bidirectional transcription. In total we collected thousands of samples from the sequencing archives from which we annotated hundreds of thousands of sites of bidirectional transcription. The majority of these sites did not reside at promoters and were either cell type or tissue specific. Finally, we link sites of bidirectional transcription to target genes, showing that many highly correlated pairs are supported by known enhancer-gene linkages. Our repository will serve as valuable resource for future studies into transcriptional regulation.

Results

A repository of run-on nascent RNA data

We began by assembling a large repository of previously published nascent transcription data sets (Figure 1). To this end, nascent RNA sequencing experiments were manually curated from the Gene Expression Omnibus (GEO) [28, 29] and the NIH Sequence Read Archive (SRA) [30]. We excluded metabolic labeling techniques as recovery of transcribed regulatory elements is highly sensitive to the length of incubation with the marked nucleotide. Metadata details such as organism, cell type, protocol used, library preparation, treatment type/conditions, and replicate information were collected for all samples from their associated database information and/or publication (See Supplementary Table 1). This metadata was collected into a MySQL database (hereafter DBNascent) where all treatment condition times were annotated in reference to the time of cell harvest. Raw fastq files were processed through standardized Nextflow pipelines (Figure 1A) that include mapping, quality control, and identifying regions of bidirectional transcription. In all cases, technical replicate fastq files were combined for downstream analysis.

In total, 3,638 raw samples from the NIH Sequence Read Archive (SRA) were combined into 2,880 biological samples across 20 organisms, collected from 287 projects, which consisted of either journal articles or Gene Expression Omnibus (GEO) datasets (Figure 1B). The samples were subjected to extensive quality control (QC), from which we developed a QC ranking metric based on both read depth and complexity (Figure 1C-D). We used this metric extensively as a filtering mechanism, and most downstream analyses using high quality samples with a QC score of 1-3, unless specified otherwise. As run-on assays necessarily depend on a pull down step involving antibodies, we also sought to assess the extent of nascent RNA enrichment. To this end, we developed an additional score to identify samples that exhibited patterns of nuclear run-on (NRO) sequencing, which could then be used as another potential filtering metric (Supplementary Figure 1).

Of the 2,880 samples in DBNascent, the vast majority (2,387) were derived from either human or mouse cells (Figure 1B), and these were exclusively used for downstream analysis, e.g. identifying bidirectional regions. Samples were distributed across 19 and 10 tissues from human and mouse, respectively. In both organisms, samples were mainly collected from cell lines or cultured primary cells (Supplementary Figure 2). Additionally, a principle component analysis on high quality human samples indicates that samples cluster predominantly by tissue of origin rather than quality score, indicating that differences in the data reflect underlying biological signal more than technical variation (Supplementary Figure 3).

Bidirectional regions in DBNascent overlap cis-regulatory elements

Nuclear run-on assays, such as GRO-seq and PRO-seq, give readout of transcription from all cellular RNA polymerases. Consequently, they recover signal at both coding and non-coding regions, much of which is not annotated. Two methods for identifying

transcribed regulatory regions are Tfit and dREG[4, 11, 12]. Tfit uses a mathematical model of RNA polymerase II to identify sites of polymerase loading and initiation, the majority of which are bidirectional. In contrast, dREG uses an unsupervised support vector machine approach to identify transcribed regulatory elements (TREs), most of which show bidirectional transcription. The two approaches are thus quite distinct and complementary, but both seek to identify sites of bidirectional transcription directly from the data.

As the two methods have distinct strengths and weaknesses, we combined the results of both methods to identify sites of bidirectional transcription (see Methods for complete details). For each of the 1,638 human and 750 mouse samples analyzed, there were on average $\sim$ 25,000 bidirectional regions identified by Tfit and $\sim$ 18,500 by dREG (Supplementary Figure 4A). Bidirectional calls were then combined using a modified version of *muMerge* (version 1.1.0) (see Methods Section) [19]. The merging strategy was performed in a hierarchical manner, merging across experiments first, then across cell type and finally between the bidirectional calling methods (Figure 2A). Since the resolution of Tfit calls at RNA polymerase initiation (typically the center region of bidirectional transcription) is better than dREG [14], the coordinates of Tfit calls were used when the two callers overlap (Supplementary Figure 4B-C). Called regions were filtered to retain high quality regions, based on the data's QC score (Supplementary Figure 5).

Genome wide, 847,521 unique bidirectional calls were obtained across all human data sets and 680,735 in mouse. Bidirectional regions, as expected, are generally much shorter than genes (Supplementary Figure 6) and are similarly distributed across the genome (Supplementary Figures 7 and 8). The majority of bidirectional regions overlap non-coding regions, while a smaller percentage are in exons (Figure 2B and Supplementary Figure 9A). Characterizing the number of human genes with bidirectional regions, we find that about 80% have a bidirectional call in their promoter region, and the transcripts without a bidirectional call at their TSS were not transcribed (Supplementary Tables 2 and 3). Outside the promoter region, bidirectional regions are found uniformly across annotated transcripts in both mouse and human (Supplementary Figure 10. Bidirectional regions within the gene are mostly intronic, with many overlapping the boundaries with an exon (Figure 2C and Supplementary Figure 9B).

To assess the quality of our called regions, we next compared our bidirectional calls to annotated cis-regulatory elements from ENCODE, EnhancerAtlas and FANTOM5, as these resources annotate regulatory regions using a variety of techniques [3, 25, 31–37] (Figure 2D and Supplementary Figure 11). ENCODE offers a large characterization of cis-regulatory elements based on histone, DNase and CTCF signal [33, 34]. While the FANTOM5 project identifies sites of transcription initiation primarily using CAGE (Cap Analysis of Gene Expression) data [32]. Lastly, EnhancerAtlas aims to combine assorted genomic data, including ENCODE and FANTOM5 as well as nascent RNA sequencing data [25, 31, 37]. Overall, about 40% to 60% of the cis-regulatory elements in these data resources are found in DBNascent (Supplementary Figures 11A and B). Interestingly, 29,106 human and 21,999 mouse bidirectional regions are contained in all three databases (Figure 2D and Supplementary Figure 11C). In general, we found a greater overlap between bidirectional regions and EnhancerAtlas regions. However,

upon closer examination we noticed that EnhancerAtlas regions tend to be wider compared to all the other database regions therefore yielding greater overlaps (Figure 2E and Supplementary Figure 12). Notably, EnhancerAtlas includes nascent RNA data and RNA polmerase II ChIP seq in its construction, which may contribute to both the observed region length and the overlap with our called bidirectional regions. In conclusion, we recover a large fraction of the previously annotated cis-regulatory elements, despite having data from far fewer tissues than was used in these databases.

Regulatory regions have also been identified based on large scale genome-wide association studies. In particular, the GTEx consortium examined genome variation for its ability to influence expression levels [38]. As sites of bidirectional transcription are often genetic enhancers, we next considered to what extent our bidirectional calls overlap with GTEx identified variation. While only a small number of GTEx variation resides within our bidirectional regions (Supplemental Figure 11A), we found that bidirectional regions showed a higher odds for containing significant expression quantitative trait loci (eQTL) variants compared to non-significant variants (Supplementary Figure 13) [38]. This further supports previous work showing an enrichment of eQTLs in enhancer regions [4].

Tissue specificity of transcription

We next sought to determine how transcription levels varied across different types of transcribed regions. Within a representative high-quality dataset [39], promoter bidirectional regions were most likely to be highly transcribed, followed by both coding and annotated noncoding genes (Figure 3A). Collectively, the exonic, intronic, and intergenic bidirectional regions (non-promoter bidirectional regions), which tend to be enhancers, are much more lowly expressed. This pattern held true across all 741 human samples, where coding genes and promoter bidirectional regions were more highly transcribed with less variability across samples than non-promoter bidirectional regions or noncoding genes (Figure 3B).

Due to the consistent trends of the magnitude of bidirectional transcription within given region classes, we investigated the tissue specificity for these classes. We limited this investigation to samples with a QC score of 1-3 that were derived from unique tissues at least 5 samples in the database. As the number of samples in each tissue varied widely, we chose to assess tissue specificity with the SPECS score [40], which can accommodate uneven sample size across groups. The SPECS score ranges from 0 (indicating depletion) to 1 (indicating enrichment), with a ubiquitously transcribed gene scoring around 0.5. Considering all genes and bidirectional regions, the distribution of SPECS scores showed a larger proportion of bidirectional regions having lower SPECS scores, indicating they are more likely to be show higher expression in a limited set of tissues and low (or no) expression across all others (Supplementary Figure 14). For a given tissue, both genes and bidirectional regions had similar trends of high SPEC scores, with umbilical cord, prostate, and uterine samples containing the highest numbers of tissue specific genes and bidirectional regions (Supplementary Figure 15).

We next assessed the change in transcription between the most specific tissue (highest SPECS score) and next highest scoring tissue (Figure 3C). The resulting fold

change should be large for each transcript which is transcribed primarily in a single tissue. Interestingly, we observed a skew towards higher fold changes for non-promoter bidirectional regions as compared to genes. Promoter bidirectional calls showed a pattern indistinguishable from coding genes, whereas noncoding genes seemed to show more tissue specificity than coding genes, in line with previous work [41]. Within non-promoter bidirectional regions, those overlapping with exons were less tissue specific than intergenic or intronic bidirectional calls, likely due to some exonic bidirectional regions toward the 5' end of genes having spillover transcription signal from promoter regions.

The SPECS score analysis suggests that non-promoter bidirectional calls, primarily associated with enhancers, are the most tissue specific transcripts. To further evaluate this claim, for each region type we quantified 1) the number of tissues in which it was transcribed and 2) the variation of that transcription level. (Supplementary Figure 16, Supplementary Video 1). In all region classes, ubiquitously transcribed regions (transcribed in all 13 tissues) showed much less variation in transcription levels than tissue-specific regions (only present in one tissue). We further investigated the proportion of regions transcribed across the tissues analyzed (Figure 3D). By this measure, coding regions and promoter bidirectional regions are most likely to be ubiquitously transcribed, whereas intronic and intergenic bidirectional regions are most likely to show tissue specific transcription, consistent with previous reports[26]. Thus intronic and intergenic bidirectional regions associated with enhancers are transcribed and active in a small range of tissues compared to coding and noncoding genes.

Correlation analysis to identify putative bidirectional and gene pairs

Various methods have been developed to link enhancers to target genes with genomic sequencing data [25, 42–44]. Initially, the closest gene approach was primarily used for assigning enhancers (or ChIP sites) to target genes. However, the closest gene is not always accurate, as indicated by 3D information [45]. Prior work on nascent transcription showed that enhancers and their known target genes – as determined by 3D data – have correlated transcription levels[7, 26, 27, 46]. Therefore we sought to determine whether correlation across the collection was sufficient to identify enhancer to target gene linkages.

To this end, we calculated pairwise gene and bidirectional correlations within each chromosome and identified highly correlated pairs in a tissue specific manner for human samples (Supplementary Figure 17) (See Methods). In a collection of bidirectional regions and genes we found 1,094,246 unique pairs where the absolute Pearson correlation coefficient (PCC) was greater than 0.6, and the adjusted p-value was less than 0.01. Most pairs are on chromosome 1 (Supplementary Figure 18A), consistent with its large size (∼248 Mb) and high gene density.

While not a constraint of the approach, we found that most bidirectional regions within the pair were close to the gene TSS (Supplementary Figure 18B). Across these pairs, the median number of bidirectional regions assigned to a gene was 42, indicating that many bidirectional regions may factor into tuning the transcription level of a given gene. However, within the context of a single tissue we observe fewer bidirectional

regions linked to a gene (median = 2-8; Supplemental Figure 19). This estimate is on par with other estimates of number of enhancers linked to a gene[47–54]. In the other direction, the median number of genes assigned to a bidirectional transcript across all tissues was four, implying that a single bidirectional has only a few potential gene targets (Supplementary Figures 18C and D). As before, this decreases in a tissue specific context (median = 1-3; Supplementary Figure 20).

When assessing the number of tissues that support a pair, approximately 35% of pairs were supported by two or more tissues (Figure 4A and Supplementary Figure 21A). In total, 82.7% of genes are linked to a bidirectional transcript, while only 21.15% of the bidirectional regions have links (Supplementary Figures 22A and B). Interestingly, there was no relationship between SPECS scores and the number of tissues supporting a gene and bidirectional link (Supplementary Figures 23A and B).

We next sought to determine whether our correlated pairs were enriched for biologically meaningful pairs. To this end, we took a randomization strategy. We reasoned most biologically meaningful correlation would break down if the data were selected randomly from all bidirectional regions not on the current chromosome. Thus we shuffled the data associated with each bidirectional position, sampling the vector of transcription profiles from all other chromosomes. We then compared assigned pairs from the within chromosome comparisons to cross chromosome comparisons, finding that within chromosome comparisons had far more assigned pairs (Figure 4B), suggesting our assigned pairs contain real signal beyond random correlations. Notably, this randomization is imperfect, as we would retain some real correlation signal when the randomly selected bidirectional and current bidirectional shared an upstream regulator. Thus our randomization likely underestimates the proportion of real biological induced correlation relative to spurious correlations.

Next, we sought to evaluate our recovered pairs by comparison to collections of known enhancer to gene linkages. First, we examined the overlap of nascent derived pairs to experimentally validated enhancer and gene pairs from K562 cells, and observed a significant recovery of known interactions (Figure 4C) [55]. Next, we considered GTEx identified pairs and found that over 80,000 nascent derived pairs overlapped with GTEx pairs (∼15.28% of eQTLs pairs for variants that overlapped intergenic bidirectional regions compared to only 0.72% for random pairs), with most of these pairs near the gene TSS regions (Figure 4D). In all cases, we recover both previously identified pairs and new regions of high correlation (Supplementary Figure 24). For example, HCG18 is a long non-coding RNA that shows a high number of interactions (96 HCG18 and bidirectional pairs within 1MB of the TSS), and 65.62% of these pairs also overlap eQTLs from GTEx (Figure 4E and F)[38]. Overall, comparisons to known enhancer and gene pairs suggest that most of our identified pairs are supported by orthogonal methods.

As a final validation of our enhancer to gene linkages, we next sought to rank the bidirectional regions based on their ability to predict the transcription level of the associated target gene. To this end, we used a tree-based variable selection method[56], where the bidirectional regions were ranked based on their ability to predict the transcription levels of the paired gene. Overall, pairs found in GTEx were highly ranked across tissues (over 70% of pairs overlapping GTEx were in the top 10 of the ranks in

at least one tissue) (Figure 4G and Supplementary Figure 25A). However, depending on the tissue a pair is found in, some top-ranking pairs do not overlap GTEx pairs since not all tissues in DBNascent are also in GTEx (Supplementary Figure 25B). These rankings add a separate score for each bidirectional linked to a gene in each tissue, giving us another layer of confidence in our defined pairs.

Finally, we reasoned that one use of the enhancer to gene pairs is to link intergenic disease-associated variation, typically single nucleotide variations (SNPs) to the relevant gene target. As a test of this scenario, we examined leukemia-associated SNPs from the European GWAS catalog[57]. This collection contains 7,805 distinct rsIDs (3,245 SNPs in genes and 4,560 intergenic SNPs) associated with leukemia from 35 distinct publications. When the 4,560 intergenic SNPs are assigned to the closest gene, Enricher fails to identify leukemia as a relevant term in any of the categories not built on closest gene.

In contrast to closest gene, we sought to evaluate the intergenic SNPs with our enhancer to gene pairs. Of the 4,560 intergenic SNPs, 391 were in bidirectional regions used in our correlation analysis. Of the 11 tissues, blood had the highest number of SNPs overlapping enhancer – gene pairs (Supplementary Table 4), consistent with leukemia being cancer of the blood. In blood, 126 intergenic SNPs overlapped 111 bidirectional regions, which were linked with 259 genes. Moreover, in blood – but not other tissues – Enricher showed an enrichment for leukemia-related genes in multiple categories. Finally, forty-three of the genes linked by our method also had at least one intergenic leukemia SNP, further supporting the quality of the enhancer to gene pairs. Thus, the enhancer to gene linkages inferred here were able to identify leukemia as the relevant disease, even when the closest gene approach could not.

Discussion

Here we present DBNascent, an atlas that catalogs published nascent RNA sequencing data, with an emphasis on run-on assays. Previous work indicates that the detection of enhancer RNAs can vary by run-on protocols[58], thus we merge data across a large collection of high quality data from multiple protocols to identify sites of bidirectional transcription genome-wide across experiments, cell types, and tissues. As expected, sites of bidirectional transcription were randomly distributed across the genome, lowly transcribed, and highly variable. While previous work reports that enhancer associated transcripts are cell type specific[26, 41], our work further extends this conclusion showing they are also more specific than both protein coding genes and long non-coding RNAs. Finally, we assign cis-regulatory regions to likely target genes using a correlation based framework, identifying many possible enhancer to gene linkages.

Several previous papers showed correlations between enhancer and target gene transcription levels[18, 25–27]. Here we leveraged this fact to assign cis-regulatory regions to likely target genes using a correlation based framework. We identified correlated bidirectional and gene transcripts across human tissues. The correlated pairs we identified overlap experimentally validated enhancer—gene pairs as well as eQTLs from GTEx, supporting the use of these data to investigate regulatory region assignment. Given that the method we use to identify pairs relies on correlations of

transcription levels, spurious correlations are a real concern. To curb the false positive rate, we added constraints for assigning bidirectional regions to a gene. Namely, allowed correlations that were supported by the majority of samples, pairs that were within a 1 Mb window, and had a false discovery rate of less than 0.01 on the correlation p-values. These filter steps likely enrich for true correlations but may do so at some cost with respect to less frequent but real interactions.

Additionally, we estimated the relative enrichment of true correlations relative to spurious ones by using a randomization strategy. However, it is well worth noting that there may be real correlations in the random data, as we do not control for upstream regulators (e.g. transcription factors). Despite this, we obtain more high quality correlations within the biological data and many of these pairings are good predictors of gene activity. It is well known that many disease associated variants occur in noncoding regions of the genome, often in enhancers that are associated with regions of bidirectional transcription [4, 59, 60]. Further, we demonstrate that our enhancer to gene linkages enrich for disease relevant genes better than the closest gene strategy that is commonly used to assess disease associated SNPs. Thus we recover novel biologically relevant pairings in a tissue specific manner.

Finally, it is also worth noting that the correlation linkages identified here could be used more generally to infer gene regulatory networks (GRNs). Correlation based network inference methods [61] for GRNs are, in theory, an excellent starting point for these analysis. However, our experience indicates that the increase in data set size that arises from including many potential regulatory regions makes the practical utilization of these tools challenging. Further work on building networks from these data would offer a great condition-specific resource for further experimental validation.

1 Figures

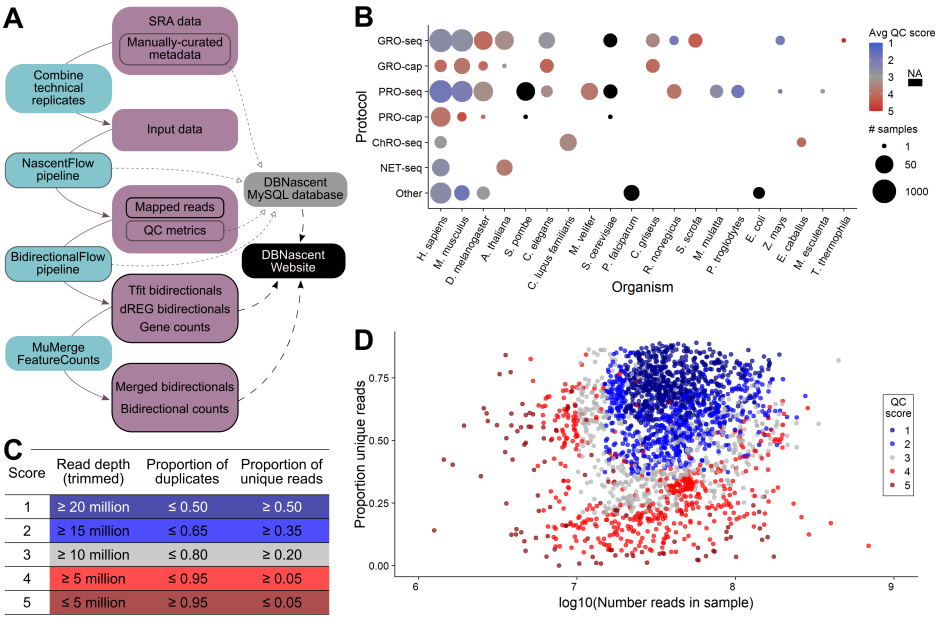

Fig. 1 Overview of DBNascent. (A) Data were derived from Sequence Read Archive fastq files and manually curated metadata. Technical replicate fastq files were combined, then data were processed to obtain metrics on quality, bidirectional regions, and read counts. Metadata, quality control metrics, and software version information from the pipeline were accumulated into a MySQL database. The DBNascent website (nascent.colorado.edu) draws from the MySQL database as well as processed analysis files for visualization and region-specific read counts. (B) Samples in DBNascent were derived from twenty different organisms and multiple different protocols. All species with genomes less than 25 Mb were not described well by the calculated QC score and thus are represented as black (NA). (C) Thresholds for calculation of the QC score, tuned for mammalian samples. (D) Complexity (y-axis) versus read depth (x-axis) of human and mouse samples. Two very low read depth samples have been omitted for the sake of visualization.

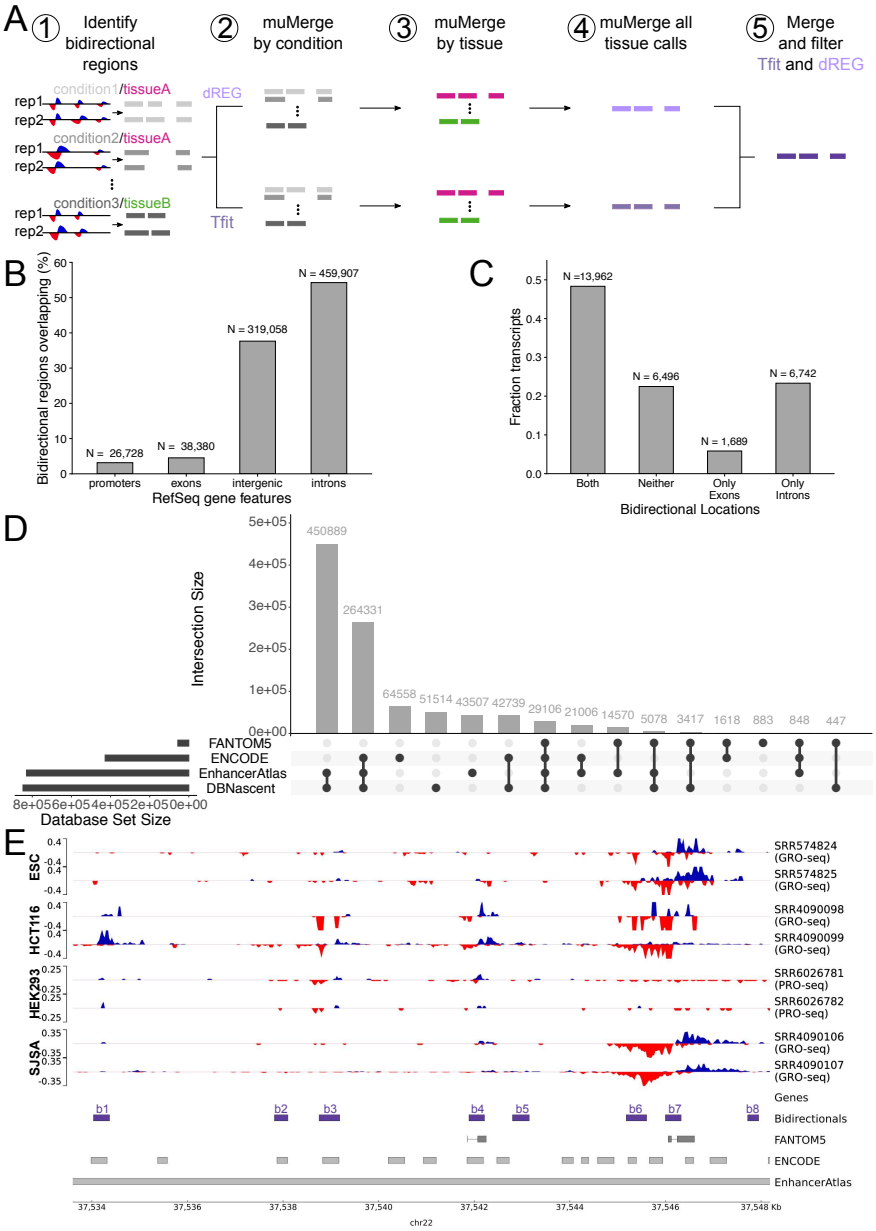

Fig. 2 Identification and characterization of bidirectional regions. (A) Schematic showing how bidirectional regions were identified and merged to give a consensus annotation set. Briefly, (1) First, bidirectional regions were inferred from nuclear run-on coverage data in each sample. (2) In a given experiment, regions were combined using muMerge[19] based on treatment conditions for both Tfit[11] and dREG[4, 12] called bidirectional regions. (3) Tfit and dREG calls were combined by muMerge based on cell/tissue type. (4) Master call lists were then obtained by muMerge and (5) combined and filtered. (B) Overlap between bidirectional regions and RefSeq hg38 gene features (exons, introns, promoters and intergenic regions). (C) Fraction of RefSeq annotated genes with bidirectional regions overlapping their introns and/or exons. (D) Overlap between bidirectional regions (DBNascent) and cis-regulatory elements from other databases (ENCODE[33–36], FANTOM5[3, 25, 31, 32] and EnhancerAtlas[37]). (E) An example region (chr22:37,533,600-37,548,188) showing mapped read coverage for two replicates each of four cell lines (ESCs, HCT116, HEK293, and SJSAs) along with bidirectional region calls (purple) and regulatory regions identified by FANTOM5, ENCODE and EnhancerAtlas. Importantly, each inferred bidirectional region has a distinct cell type and tissue specific transcription profile. b1: HCT116 only, b3, b4: HCT116 and SJSA, b7: ESC, HCT116 and SJSA, b2, b5, b6 and b8: not in these four cell lines; blue: positive strand, red: negative strand.

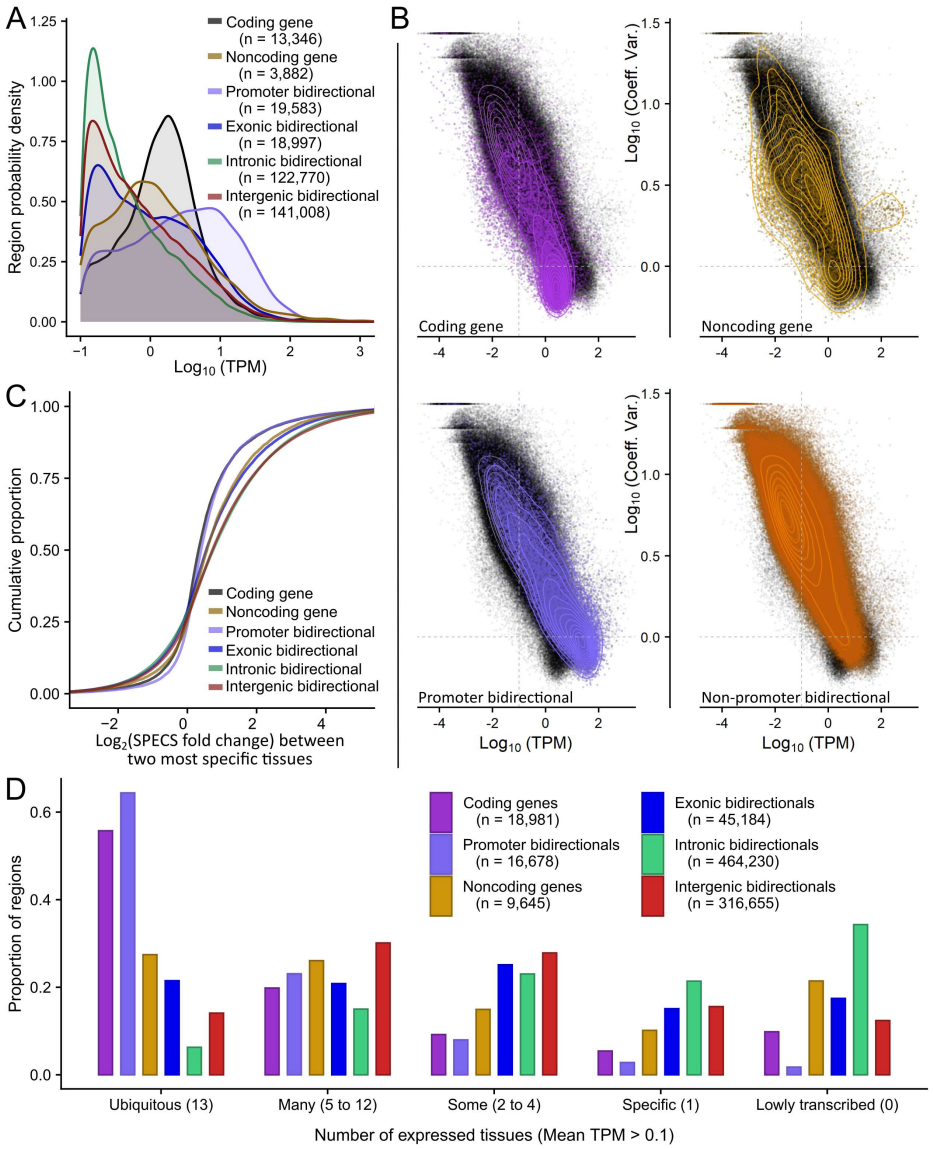

Fig. 3 Variation in transcription levels and tissue specificity across annotation types. (A) Distribution of average TPMs (x-axis) for different classes of regions across replicate high-quality MCF7 datasets (SRR5227979 and SRR5227980). Number of regions (n) with average TPM > 0.1. (B) Across high quality human samples, the coefficient of variation (y-axis) of each region class compared to transcription (x-axis, log(TPM)). Black points and gray density contours display all regions in all plots, overlaid by region-specific colored points and density contours. 'Non-promoter bidirectionals' includes intronic, exonic, and intergenic regions. (C) Cumulative distribution of fold changes between the tissue with the highest SPECS score and the tissue with the second highest SPECS score for each region class. This strategy is adapted from Everaert et al. 2020[40]. (D) Number of tissues (x-axis) in which a region is transcribed, by class of region. 'Lowly transcribed' refers to regions that failed to reach the TPM threshold of 0.1 in any tissue.

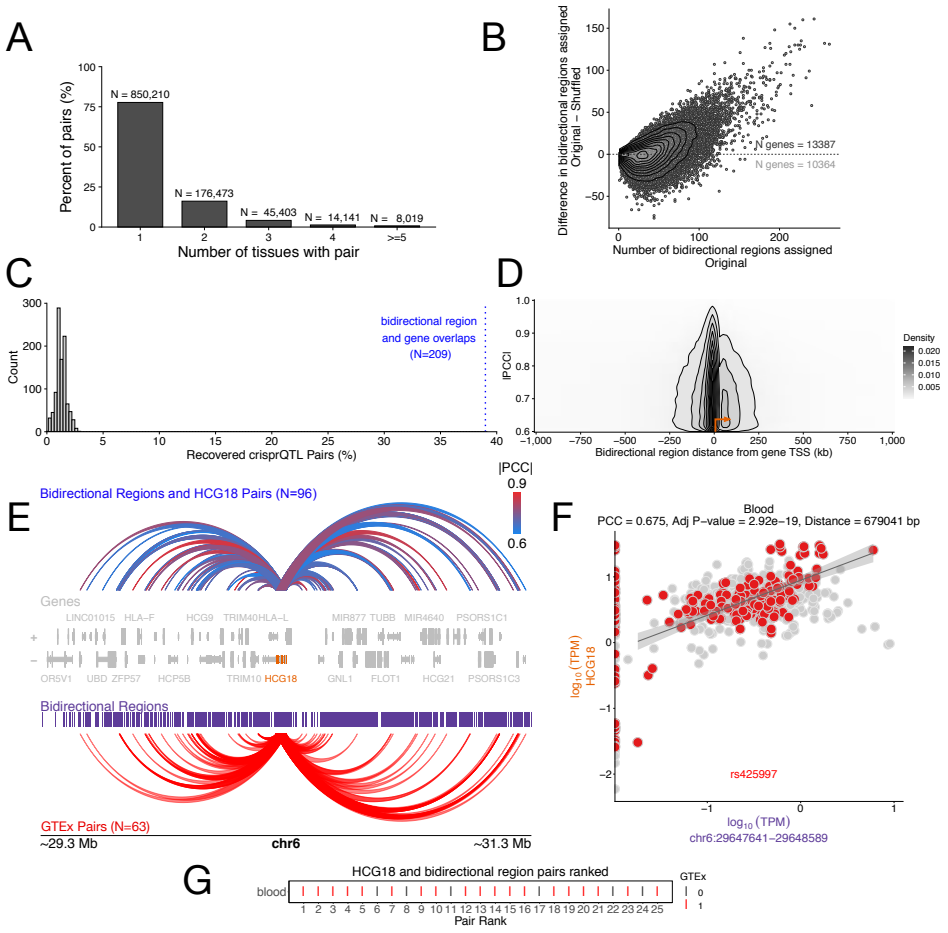

Fig. 4 Linking bidirectional regions to protein coding genes (A) Summary of pairs identified by tissue type (Number of unique pairs = 1,094,246). (B) Number of bidirectional regions assigned to each gene (original, x-axis) compared to the difference in number of regions assigned when shuffling (y-axis). Shuffled refers to genes correlated with bidirectional regions sampled from other chromosomes. (C) Comparison of recovered known enhancer to gene pairs[55] (blue) to pairs recovered in the shuffled approach (grey). (D) Typical distance between annotated gene and assigned bidirectional region (x-axis) and Pearson's correlation coefficients (y-axis) for pairs overlapping significant GTEx pairs (Number of overlapping pairs = 81,046). (E) Interactions for HCG18 (NR_024052.2) with top showing identified pairs colored by absolute Pearsons correlation coefficient and bottom showing GTEx pairs. (F) Scatter plot of the raw data for one interacting pair: HCG18 (y-axis) and the bidirectional region at chr6:29,647,641-29,648,589. Samples in blood are highlighted in red. (G) Relative ranking of linked bidirectional regions on their ability to predict HCG18 levels in blood. Interactions found in GTEx are colored red.

2 Methods and Materials

All the code and methods used in the meta-analysis of nascent RNA sequencing experiments can be found on this GitHub page https://github.com/Dowell-Lab/DBNascent_Analysis. The samples were processed on a compute cluster running CentOS Linux v7.

Mouse samples were mapped to the mm10 reference genome and human samples to the hg38 genome. Databases used for comparisons that were mapped to older reference genomes were lifted to the specified genomes above using liftOver [62]. NCBI RefSeq annotations were used for both human (hg38 release GCF_000001405.40-RS_2023_03) and mouse (mm10 release GCF_000001635.26_GRCm38.p6) data.

Nascent RNA sequencing experiments metadata collection

Nascent RNA sequencing experiments were manually curated from the Gene Expression Omnibus (GEO) [28, 29] and the Sequence Read Archive (SRA) [30]. All treatment condition times were annotated in reference to the time of cell harvest. Papers that had other high throughput experiments (including RNA-seq, ATAC-seq, ChIP-seq and 3D chromatin assays such as HiC) that were performed along with nascent RNA sequencing were noted. Two rounds of data curation were implemented where the first round was meant for data entry, and the second round for entry verification. In total, 3,638 raw samples were manually curated from 320 SRA projects (SRPs) and 287 papers. Types of data curated is described in Supplementary Table 1. Full list of papers is provided in the Supplemental References.

Preprocessing nascent RNA sequencing experiments

All SRR accessions were downloaded from the SRA and extracted with SRA Toolkit (versions 2.8.0 and 2.9.2). Replicate information was used, where available, to combine technical replicates by concatenating fastq files. New samples resulting from technical replicate combination within a given experiment were given SRZ designations with a number equivalent to the first numerical SRR contained within. In one case, technical replicates were combined using data from multiple papers ([63, 64]) as a result of further resequencing of previously published samples. Combined samples in this case were given SRM designations, with numeric conventions the same as the SRZs. In total, 2,880 samples were generated from the original 3,638 SRR entries after technical replicate concatenation. This collection of 2,880 samples was then the source of all downstream analysis.

Mapping reads to reference genomes

All samples in the database were trimmed, mapped to the corresponding reference genome, and assessed for sample quality using an in-house NextFlow pipeline (https://github.com/Dowell-Lab/Nascent-Flow), run with NextFlow (version 20.07.1) [65]. Briefly, within this pipeline fastq files were trimmed for adapter sequences and low quality bases using BBMap (version 38.05), then aligned to reference genomes with

HISAT2 (version 2.1.0) [66, 67]. Downstream mapped read files (CRAM files and IGV-compatible TDF files) were generated with Samtools (versions 1.8 and 1.10), Bedtools (versions 2.25.0 and 2.28.0), and IGVtools (version 2.3.75), and in some cases were done so through an additional NextFlow pipeline (https://github.com/Dowell-Lab/Downfile_pipeline) [68–70]. As data was processed over an extended time frame, some software was updated resulting in changes to software versions. As such, all versions used to process a specific sample are linked to that sample in the database.

Quality control and quality tiers

Samples were assessed for quality using metrics from the following software packages: FastQC (version 0.11.8), HISAT2 (version 2.1.0), Preseq (version 2.0.3), RSeQC (version 3.0.0), Picard tools (version 2.6.0), and BBMap[71–74]. As with mapping, specific software versions are linked to samples within the database.

Three specific metrics were used to classify samples into quality 'tiers' for filtering purposes: read depth after trimming, proportion of duplicates (as assessed using Picard tools), and complexity (as assessed using the modeled value for unique reads in 10 million output by Preseq). Thresholds were determined to classify samples into one of five tiers (see Figure 1), and analysis was performed on samples within tiers 1-3 unless specified otherwise.

A 'run-on score' was also assigned to human and mouse samples based on the following metrics: Exon/intron ratio, as calculated from the RSeQC output values for 'CDS Exons' and 'Introns', and when available from Tfit data, the total GC content of all bidirectional regions called by Tfit.

Identifying bidirectional transcripts

Regions of nascent transcription were identified using Tfit [11] and dREG [4, 20]. Identification of regions of transcription with dREG followed the recommended pipeline (per dREG github) where uniquely mapped reads were used to generate BigWig files. BigWig input files for dREG were generated by converting filtered BAM files using bedtools `bamToBed`, then BED files were converted to bedGraph format using `bedtools genomecov`, and finally the BigWig files were generated using `bedGraphToBigWig` (from https://www.encodeproject.org/software/bedgraphtobigwig/) [75]. Since dREG is compute-intensive, only high-quality data sets (QC < 4) were processed using dREG. Identification of bidirectional transcription with Tfit followed a pipeline where multimapped reads and reads with low map quality score were filtered as shown:

```
samtools view -@ 16 -h -q 1 \$\{SRR\}.bam | \
    grep -P '(NH:i:1|\textasciicircum@)' | \
    samtools view -h -b > \$\{SRR\}.filtered.bam.
```

Input bedGraph files were generated using `genomeCoverageBed` from bedtools. Tfit was run in a two step processes, first with the template matching module to identify sites of bidirectional transcription, then these regions were used as input to fit

the precise RNA polymerase behavior. The nextflow pipeline used for characterizing bidirectional transcription with both Tfit and dREG can be found on GitHub (https://github.com/Dowell-Lab/Bidirectional-Flow).

Merging regions of bidirectional transcription

Updated muMerge method

This project required the development of several new features for *muMerge*[19] in order to facilitate aspects of the analysis. Namely, the creation of a filter to remove regions occurring in only one sample ("singletons filter"; `-r, --remove_singletons`) and an option to save a record of which samples contribute to an individual mumerged region ("save sample IDs"; `-s, --save_sampids`). The singletons filter filters any output region supported by one one data input. The save sample IDs option adds a fourth column to the output file which contains a comma separated list of all the sample IDs that contributed at least one loci to the calculation of that given mumerged region. These sample IDs are reported in alphanumeric order. These two features have been included in `mumerge v1.1.0` (see https://pypi.org/project/mumerge/ for details).

muMerging samples across conditions

Regions (from replicates, conditions, and bidirectional calling methods) were merged using *muMerge* (described above). Since *muMerge* combines regions in a probabilistic way, replicate information and sample conditions were taken into account for the merging processes. Tfit and dREG bidirectional transcript calls were first mumerged separately by paper based on the experimental setup (that is by cell/tissue type, experimental condition and replicate information).

The *muMerge* bed files by experiment/paper were combined based on the cell/tissue types for Tfit and dREG, where the same cell/tissue types were treated as "replicates" and the different cell/tissue types were the "conditions". All samples were mumerged and calls were filtered based on the paper QC and the GC content of the 300bp region around the center of the call (Supplemental Figure 5). Bidirectional calls from papers with GC content ¿ 0.49, average paper QC score ¡ 3 were kept (Supplemental Figure 5). In summary, samples from 61 mouse papers and 101 human papers were used for the final species-specific *muMerge* steps. Furthermore, the dREG and Tfit *muMerge* files were combined such that Tfit calls were used for overlapping regions creating a master *muMerge* file for both species (shown schematically in Figure 2A).

Lastly, the dREG and Tfit master *muMerge* file was filtered. Regions that were greater than 150bp and less than 2.5kb were kept in the human bidirectional calls (less than 2kb in mouse regions). The minimum read length of 150bp was selected as this is the maximum read length for libraries in the nascent RNA samples in the database, and the upper limit was selected so that the largest regions from dREG and Tfit matched. In order to remove false positive bidirectional calls, bidirectional calls that were in regions of converging gene transcripts (where converging genes here are defined as sense and antisense transcripts that collide) within 1kb were also removed.

Some general observations on the two methods were made. First, we found that dREG broke bidirectional regions identified by Tfit into smaller chunks. Despite this,

Tfit calls slightly more regions per sample compared to dREG (Supplemental Figure 4A). Second, both methods struggled to call bidirectional regions within introns when the gene was transcribed robustly, though generally dREG called more in these regions (Supplemental Figure 26). Finally, when samples have poor quality, both dREG and Tfit tend to call a higher percent of gene TSS regions as these are more highly transcribed and therefore have the most robust signal (Supplemental Figure 27). More generally, we see a higher number of TSS regions called by dREG comapred to Tfit. As no gold standard exists for transcribed regulatory elements, it is unclear which of the two methods is more accurate. There are regions that are recovered by both methods, and regions unique to each bidirectional transcript caller (Supplemental Figures 4B and C). So in this paper we combined calls from both methods, keeping track of their origin, i.e. bidirectional caller.

The processing pipeline was performed with R (version 3.6.0), using the package data.table (version 1.14.2), and genome arithmetics were performed using bedtools [76, 77]. The pipeline used can be found on GitHub here: https://github.com/Dowell-Lab/bidirectionals_merged.

Summary of transcription

Bidirectional transcripts overlapping genomic features

Annotated bidirectional transcripts were overlapped with RefSeq genome feature using bedtools. The bidirectional coordinates were overlapped with intron, promoter (1kb upstream of a gene TSS), exon and intergenic regions (regions not annotated in the reference annotations). Across all features, the minimum fractions overlap required per region was 0.5 as shown:

```
bedtools intersect -a \${bidirectionals} -b \
    \${genome\_feature} -wa -u -f 0.5 > \
    \${feature}.mumerge\_overlap50perc.bed.
```

The percent of overlap in each feature category was calculated as a fraction of the total bidirectional transcripts (847,521 in human and 680,735 in mouse).

Gene summary statistics methods

All gene statistic analyses were performed using R (version 4.3.0), the R package data.table (version 1.14.8), and bedtools (version 2.30.0).

To identify TSS bidirectional regions we used bedtools intersect to find overlaps of the final bidirectional calls with a 600bp window around the TSS (300bp both directions). For gene transcripts with multiple bidirectional regions overlapping, the bidirectional whose center (μ) was closest to the TSS is considered the TSS bidirectional. The notebook corresponding to this analysis can be found on GitHub here: https://github.com/Dowell-Lab/bidirectionals_merged/notebooks/.

To further interrogate the number of intronic and exonic bidirectional regions at a gene-centric level, we used bedtools to intersect a 100bp window around μ with introns and exons, requiring that the bidirectional have majority coverage in the exon or intron. The following code was used:

```
bedtools intersect \
    -wo -f 0.51    \
    -a ${genome}_bid_100bpMU.bed \
    -b ${genome}_refseq_exons.bed ${genome}_refseq_introns.bed \
    -names exon intron >    \
    overlaps_${genome}_bid_exons_introns.bed
```

To get percentile based coordinates of bidirectional regions within genes, all positions (midpoint, start, and end of bidirectional and gene) were transformed in relation to the gene itself, with the TSS marking 0 and PAS marking the length of the gene. These coordinates were then multiplied by a size factor calculated as shown below to standardize the coordinates as percentiles where i refers to the transcript isoform.

$$\text{Size Factor}_\text{i} = \frac{100}{length_i} \tag{1}$$

where

$$\text{length}_\text{i} = abs(PAS_i - TSS_i) + 1 \tag{2}$$

The notebook corresponding to this percentile analysis can be found on GitHub here: https://github.com/Dowell-Lab/DBNascent_Analysis/analysis.

Overlapping bidirectional transcripts with candidate cis-regulatory elements

Region overlaps

Regions of bidirectional transcription were overlapped with candidate cis-regulatory element (cCRE) databases (ENCODE, EnhancerAtlas, FANTOM5 and eQTL data) using bedtools [3, 32, 33, 37]. The reprocessed version 9 of FANTOM5 data for mm10 and hg38 were downloaded from the FANTOM website https://fantom.gsc.riken.jp/5/datafiles/reprocessed/. EnhancerAtlas 2.0 data was download from the database website http://www.enhanceratlas.org/data/download/species_enh_bed.tar.gz and using liftOver, coordinates were converted to the mm10 or hg38 for mouse or human respectively. ENCODE candidate cis-regulatory elements were downloaded from the UCSC genome browser for both human http://hgdownload.soe.ucsc.edu/gbdb/hg38/encode3/ccre/encodeCcreCombined.bb and mouse https://hgdownload.soe.ucsc.edu/gbdb/mm10/encode3/ccre/encodeCcreCombined.bb. Finally, significant eQTLs from GTEx version 8 (GTEx_Analysis_v8_eQTL.tar) were downloaded from the GTEx portal (https://gtexportal.org/home/). Mouse eQTLs were downloaded from the from Gonzales et al. paper [78].

Regions of overlap were calculated using the minimum overlap of 1bp as shown:

```
bedtools intersect \
    -a ${cCRE_database} \
    -b ${bidirectionals} -wa -u \
    > ${cCRE}_bidirectional_overlap.bed
```

The percent overlap was calculated with respect to the database size.

Odds ratio with GTEx

Bidirectional transcription regions were overlapped with disease associated variants from GTEx (version 8) and odds ratio calculated [38]. All eQTLs including non-significant instances (GTEx_Analysis_v8_eQTL_all_associations) were downloaded from the google cloud location stated in the GTEx portal. The odds ratio was calculated by counting the number of GTEx eQTL variants that overlapped bidirectional transcripts for both significant and non-significant variants and getting the fraction of the variant overlapping bidirectional transcripts versus not overlapping the transcripts (Supplement Figure 13).

$$\text{Odds Ratio}_\text{t} = \frac{sb_t/sn_t}{nb_t/nn_t} \tag{3}$$

Where t is a given GTEx tissue, sb_t are GTEx eQTL variants that fall in bidirectional transcripts, sn_t are significant variants that fall outside of bidirectional transcripts, nb_t are non-significant variants that fall within bidirectional regions and nn_t are non-significant GTEx variants that do not fall in bidirectional regions. The odds ratio calculation was performed using the library epitools (version 0.5-10.1) in R [79].

Calculating base content

Base composition for bidirectional transcript calls from dREG and Tfit was defined as the ratio of GC content in the center 300bp (typically contains the RNA loading position) relative to larger bounding 3kb region. This was performed by extracting sequences using `bedtools getfasta`, and counting the frequency of bases (A/T/C/G) within the small (300bp) and large (3000bp) regions. This counting was performed in python 3.6 [80].

Counting reads

Reads were counted using featureCounts from RSubread (version 2.12.3) [81]. For gene transcripts, sense strand reads over gene bodies were counted as these show more consistency across nuclear run-on protocols [58]. Gene bodies were defined for genes over 300bp long as full gene lengths from the TSS to the TES, truncated at the 5′ end by 30% of the gene length up to a maximum of 750bp truncation. Genes $<$ 300bp were not truncated. For bidirectional regions, reads on both strands were counted across the entire called region using the coordinates start+1 to end-1. In both cases, multimapping reads were ignored. Multi-feature overlap was allowed for counting across genes but not bidirectional regions.

Normalizing read counts

Normalization of counts was done using transcripts per million (TPM) normalization as shown below [82, 83]:

$$\text{TPM}_\text{i} = \frac{r_i/l_i}{\sum_1^j r_j/l_j} \times 10^6 \tag{4}$$

where r_i are the mapped reads for transcript i (for all genes and bidirectional transcripts), l_i is the transcript length and $\sum_j r_j/l_i$ sums all j length normalized transcripts. The ratio is multiplied by a scaling factor of 10^6. The counts for genes was normalized over full length of the longest transcript and the 5′ end of that transcripts were truncated. To avoid double counting reads, for bidirectional transcripts that overlap genes, the opposite sense read counts were used in the normalization step. In summary, the total number of transcripts included 5′ end truncated genes, intergenic bidirectional transcripts and intragenic bidirectional regions where counts on the opposite strand of gene were used for these bidirectional transcripts.

Calculating summary statistics

The summary statistics described below were calculated using R. For all samples, the average and median transcription values were calculated based on the normalized counts.

Coefficient of variation

Across-sample coefficient of variation (CV) for human samples were calculated as follows:

$$\mathrm{CV_i} = \frac{\sigma_i}{\mu_i} \tag{5}$$

where for transcript i the standard deviation (σ_i) for normalized counts is divided by the average normalized counts (μ_i).

Principle component analysis

The principle component analysis (PCA) was performed using human GRO-seq and PRO-seq samples with QC 1 and 2 and with a GC content greater than 0.49. This resulted in 751 samples for analysis. RefSeq hg38 genes with high CV (greater than 3) were used for the PCA. Since bidirectional transcripts are lowly transcribed and have a higher coefficient of variation, more stringent filter was used to filter these transcripts (CV greater than 6 and average TPMs greater than 0.1). In both cases, normalized counts were log transformed, euclidean distances calculated using the R package distances (version 0.1.8), and PCA performed with the prcomp function in the stats package in R [84].

Tissue specificity

For analysis of variation and tissue specificity, genes were classified into 'coding' and 'noncoding' genes according to 'NM_' vs 'NR_' accession number prefixes. Bidirectional regions were classified into 'promoter', 'exonic', 'intronic', and 'intergenic' bidirectional regions according to bedtools overlaps with those annotations as described above, with exonic, intronic, and intergenic bidirectional regions also being lumped together as 'nonpromoter' bidirectional regions for some comparisons.

SPECS scores were calculated using a custom python-based implementation of the method described previously [40], using Python (version 3.6.3), Numpy (version

1.19.2), Pandas (version 1.0.2), and Scipy (version 1.4.1). This script can be found in the GitHub repository https://github.com/Dowell-Lab/DBNascent_Analysis/ [85–88].

Correlation and Co-transcription Analysis

Building the co-transcription bidirectional and gene pairs from nuclear run-on data was split into three steps: (1) finding pairs of highly correlated genes and bidirectional transcripts, (2) removing bidirectional transcripts that are in the downstream regions of genes and (3) filtering high confidence pairs (See Supplementary Figure 17).

Step 1: Pairwise correlation of gene and bidirectional transcripts

Pearson's correlation coefficients between genes and bidirectional regions were calculated using WGCNA (version 1.70-3)[61, 89] and all the filtering done in R. Comparison were among transcribed regions within a chromosome (i.e. no cross chromosome comparisons). The input to WGCNA was normalized counts for genes where the 5′ end was truncated (as described in the "Counting reads" section) along with bidirectional regions. These counts were log transformed as shown below.

$$\text{Transformed TPM}_\text{i} = \log_{10}(\text{TPM}_\text{i} + 1) \tag{6}$$

Additionally, transcripts with zero counts were excluded from the pairwise calculations. Given the samples with transcription, the Pearsons correlation coefficient (PCC) was calculated as follows:

$$\text{r}_{\text{x,y}} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}} \tag{7}$$

where x are the genes, y bidirectional transcripts and n are the number of samples transcribed for both gene and bidirectional regions. The p-value was calculated from the Student's t-distribution and the t statistic was calculated as:

$$\text{t} = \sqrt{n-2} \times r/\sqrt{1-r^2} \tag{8}$$

where n are the number of samples and r is the PCC. The output from the correlation calculations included the transcript identifiers, distance between the center of the bidirectional and the TSS and TES of a gene, the PCC, the p-value and adjusted p-values [90]. Additionally, the number of samples used in the correlation calculation as well as the tissue a pair is identified in are reported.

The code used to calculate pairwise correlations can be found on this GitHub repository https://github.com/Dowell-Lab/bidir_gene_pairs.

Step 2: Filtering bidirectional transcripts

It has been shown that gene transcription from nuclear run on assays shows transcription past the polyadenylation side (PAS) [2, 46]. Since bidirectional transcripts in these regions would contain reads from the gene, it is difficult to disentangle signal of the bidirectional transcript from the gene transcripts. This would result in high correlation between these bidirectional regions and the gene upstream. Therefore, pairs with bidirectional transcripts that were with 15kb downstream of the PAS were removed.

Step 3: Filtering for high confidence pairs

To ensure robustness of the correlations, we filtered our tissues to those that include at least 15 samples, leaving 11 tissues to assess. In each tissue, significant bidirectional and gene pairs were defined as pairs with 1Mb (from the center of the bidirectional transcript to the gene TSS), where the absolute Pearsons correlation coefficient (PCC) was greater than 0.6, and the adjusted p-value was less than 0.01 (Supplementary Figure 17). Importantly, pairs had to be supported by over 10 samples, in order to curb spurious correlations.

Evaluation of relative false positive rate

Since it has been shown that most enhancers regulate gene expression in a distance-dependent manner and within topologically associated domains (TADs) [91, 92], we reasoned that linking genes with bidirectional transcripts from outside TADs would yield an estimation of false positive links. Thus, we reasoned that genes and bidirectionals on distinct chromosomes would be highly unlikely to be real pairs. Using this, we assessed the relative false positive rate using a randomization strategy. Specifically, when considering a gene on a given chromosome, we consider all bidirectionals from the same chromosome (e.g. relative distance to a bidirectional region is unaltered) but sampled a random transcription vector from the set of all bidirectionals not on the chromosome. For each chromosome, the shuffling of bidirectional regions from the remaining chromosomes was done without replacement. The shuffling process was done in R using the `sample` function in the base package. The correlation analysis and filtering of pairs was done with the same methods described above (See correlation methods). The code used for the shuffling method can be found on GitHub (https://github.com/Dowell-Lab/DBNascent_Analysis).

Overlap of pairs with eQTLs and crisprQTLs

Gene and bidirectional pairs derived from the co-expression method were overlapped with pairs from crisprQTLs validated enhancer – gene pairs from Gasperini and company [55]. The gene and bidirectional pairs were randomly shuffled 1000 times within each chromosome, and the overlaps with the crisprQTLs assessed. The percent overlap was calculated based on crisprQTLs that were present and transcribed in our dataset (536 out of the 664 tested crisprQTLs). The random pairs and true pair overlaps were compared and plotted as a histogram (See Figure 4C). For eQTLs from GTEx, the randomization was only done once [38]. The eQTLs/crisprQTLs were overlapped with bidirectional transcript coordinates using `bedtools intersect` and the genes were matched based on the gene name in R. The shuffling was done using the `sample` function from the base package in R.

Ranking bidirectional transcripts with GENIE3

To rank bidirectional transcripts assigned to each gene in each tissue, the normalized counts for bidirectional regions and their paired gene were used as input to the GENIE3 (version 1.22.0) library in R (version 4.3.1) [56]. The bidirectional transcripts were

labeled as regulators. Regression trees were learned using random forests, and the number of candidate regulators selected at each tree node was set to K=$\sqrt{(}p)$ where p is the number of bidirectional transcripts (i.e. regulators). The number of trees that were grown per ensemble was 1000. The output from this analysis returned a list of gene and bidirectional pairs in each tissue, the rank for each bidirectional transcript, and the weight of link.

GWAS SNPs connected with correlation pairs

We used custom python software with the pandas (version 2.1.3) library to link SNPs to genes. GWAS-detected leukemia SNPs were downloaded from the NHGRI-EBI GWAS Catalog (EFO_0000565) [57]. SNPs (rsIDs) were filtered to remove those without position and deduplicated (12,629 total rsIDs and 7,805 rsIDs after de-duplication). Bedtools was used to find the closest gene to each SNP and to determine which SNPs were intergenic. Bedtools was also used to determine which SNPs were within a bidirectional region in DBNascent. SNPs found in intergenic bidirectional regions were overlapped with DBNascent bidirectional region – gene pairs. When evaluating for enrichment, the gene within the pair or the closest gene was fed to Enrichr[93]. DBNascent-pair genes were loaded to Enrichr with the background of all genes identified in any pair within blood. The closest gene list showed enrichment for leukemia only in gene lists that utilized closest gene in their construction, such as PhenGenI Association, GWAS Catalog 2023, DisGeNET, and GeDiPNet[94–96]. While DBNascent-paired genes showed enrichment for leukemia in both the OMIM and OMIM enriched categories [97]. OMIM links genes to diseases via literature and OMIM-enriched adds protein-protein interactions to the literature.

The code used for this analysis can be found on GitHub here https://github.com/Dowell-Lab/snpconnecter.

Data visualization

Plot generation

Plots were generated using ggplot2 (versions 3.6.0 and 3.4.2), cowplot (version 1.1.1), ComplexUpset (version 1.3.3) and VennDiagram (version 1.7.3) packages in R [98–100].

Genome tracks

Genome tracks were generated with pyGenomeTracks (version 3.8) in python [101, 102]. Gene and bidirectional pairs were represented using the plotgardener (version 1.6.4) package in R with hg38 gene annotations from TxDb.Hsapiens.UCSC.hg38.knownGene (version 3.17.0) [103, 104].

Illustrations

Illustrations were generated using the open-source tool Inkscape (version 1.0.1).

Nascent Database Structure : Back-end

All scripts for database construction and maintenance, as well as a visual schema of the database, can be found at https://github.com/Dowell-Lab/DBNascent-build.

The MySQL database backend for DBNascent was built using Python and SQLAlchemy (version 1.4.31). All metadata collected was stored in text files divided by paper (for experiment-level metadata) and samples within a paper (for sample-level metadata), which were then read to parse into database tables. Broader metadata detailing organism and cell type/tissue information was also manually curated into tables, available in the repository. Quality control metrics output by the software packages FastQC, HiSAT2, Picard tools, Preseq, RSeQC, and Pileup were pulled directly from files output by those packages. Software version information and bidirectional summary statistics were pulled from pipeline outputs from our Nextflow pipelines (NascentFlow and BidirectionalFlow), detailed previously.

A front-end website for DBNascent (nascent.colorado.edu) was built in Python 3.6.8 using the packages Django (version 3.2.16) and uWSGI (version 2.0.21) and is served using nginx (version 1.20.1). The website is maintained by the IT group at the BioFrontiers Institute.

Data Availability

Processed data and intermediate files can be found on Zenodo (10.5281/zenodo.10223322) and on the DBNascent website (nascent.colorado.edu).

Supplementary information

1. Supplemental Figures
2. Video File 1

Funding

This work was funded by the National Science Foundation under grants ABI1759949 and the National Institutes of Health grant GM125871 and HL156475.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

RFS, MAA and RDD conceived and designed the analysis. RFS and LS collected data and managed metadata curation. RFS performed analysis with help from LS and ZLM to construct DBNascent, LS for cell type specificity analysis, TJ for GC analysis, MAA for linking SNPs to genes, and HAT in feature overlap characterization. JTS revised *muMerge*. RFS, LS, RDD wrote the paper. All authors revised manuscript.

Acknowledgements

We thank Joseph Cardiello, Samuel Hunter, Jesse Kurland, Kendra Meer, Marko Melnick, Daniel Ramirez, Antonio Salcido-Alcantar, Gilson Sanchez, Jessica Westfall, Qing Yang and Chi Zhang for contributions to meta-data curation. We thank Charles Danko for assistance and advice regarding running dREG. We are also grateful to the BioFrontiers IT department for their support in building the database.

Code availability

All the code and methods used in the meta-analysis of nascent RNA sequencing experiments can be found on this GitHub page https://github.com/Dowell-Lab/DBNascent_Analysis. All scripts for database construction and maintenance, as well as a visual schema of the database, can be found at https://github.com/Dowell-Lab/DBNascent-build. Processed data and intermediate files can be found on Zenodo (10.5281/zenodo.10223322).

- Database backend: https://github.com/Dowell-Lab/DBNascent-build
- Data preprocessing: https://github.com/Dowell-Lab/Nascent-Flow
- Bidirectional calling and read counting: https://github.com/Dowell-Lab/Bidirectional-Flow
- muMerge and combining bidirectional regions: https://github.com/Dowell-Lab/bidirectionals_merged
- Correlation of bidirectional regions and genes: https://github.com/Dowell-Lab/bidir_gene_pairs
- Downstream analysis of pairs: https://github.com/Dowell-Lab/DBNascent_Analysis
- Linking GWAS SNP to genes with DBNascent: https://github.com/Dowell-Lab/snpconnecter

References

[1] Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., Fukuda, S., Sasaki, D., Podhajska, A., Harbers, M., Kawai, J., Carninci, P., Hayashizaki, Y.: Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proceedings of the National Academy of Sciences **100**(26), 15776–15781 (2003)

[2] Core, L., Lis, J.: Transcription regulation through promoter-proximal pausing of RNA polymerase II. Science **319**, 1791 (2008)

[3] Forrest, A.R.R., Kawaji, H., Rehli, M., Kenneth Baillie, J., Hoon, M.J.L., Haberle, V., Lassmann, T., Kulakovskiy, I.V., Lizio, M., Itoh, M., Andersson, R., Mungall, C.J., Meehan, T.F., Schmeier, S., Bertin, N., Jørgensen, M., Dimont, E., Arner, E., Schmidl, C., Schaefer, U., Medvedeva, Y.A., Plessy, C., Vitezic, M., Severin, J., Semple, C.A., Ishizu, Y., Young, R.S., Francescatto, M.,

Alam, I., Albanese, D., Altschuler, G.M., Arakawa, T., Archer, J.A.C., Arner, P., Babina, M., Rennie, S., Balwierz, P.J., Beckhouse, A.G., Pradhan-Bhatt, S., Blake, J.A., Blumenthal, A., Bodega, B., Bonetti, A., Briggs, J., Brombacher, F., Maxwell Burroughs, A., Califano, A., Cannistraci, C.V., Carbajo, D., Chen, Y., Chierici, M., Ciani, Y., Clevers, H.C., Dalla, E., Davis, C.A., Detmar, M., Diehl, A.D., Dohi, T., Drabløs, F., Edge, A.S.B., Edinger, M., Ekwall, K., Endoh, M., Enomoto, H., Fagiolini, M., Fairbairn, L., Fang, H., Farach-Carson, M.C., Faulkner, G.J., Favorov, A.V., Fisher, M.E., Frith, M.C., Fujita, R., Fukuda, S., Furlanello, C., Furuno, M., Furusawa, J.-i., Geijtenbeek, T.B., Gibson, A.P., Gingeras, T., Goldowitz, D., Gough, J., Guhl, S., Guler, R., Gustincich, S., Ha, T.J., Hamaguchi, M., Hara, M., Harbers, M., Harshbarger, J., Hasegawa, A., Hasegawa, Y., Hashimoto, T., Herlyn, M., Hitchens, K.J., Ho Sui, S.J., Hofmann, O.M., Hoof, I., Hori, F., Huminiecki, L., Iida, K., Ikawa, T., Jankovic, B.R., Jia, H., Joshi, A., Jurman, G., Kaczkowski, B., Kai, C., Kaida, K., Kaiho, A., Kajiyama, K., Kanamori-Katayama, M., Kasianov, A.S., Kasukawa, T., Katayama, S., Kato, S., Kawaguchi, S., Kawamoto, H., Kawamura, Y.I., Kawashima, T., Kempfle, J.S., Kenna, T.J., Kere, J., Khachigian, L.M., Kitamura, T., Peter Klinken, S., Knox, A.J., Kojima, M., Kojima, S., Kondo, N., Koseki, H., Koyasu, S., Krampitz, S., Kubosaki, A., Kwon, A.T., Laros, J.F.J., Lee, W., Lennartsson, A., Li, K., Lilje, B., Lipovich, L., Mackay-sim, A., Manabe, R.-i., Mar, J.C., Marchand, B., Mathelier, A., Mejhert, N., Meynert, A., Mizuno, Y., Lima Morais, D.A., Morikawa, H., Morimoto, M., Moro, K., Motakis, E., Motohashi, H., Mummery, C.L., Murata, M., Nagao-Sato, S., Nakachi, Y., Nakahara, F., Nakamura, T., Nakamura, Y., Nakazato, K., Nimwegen, E., Ninomiya, N., Nishiyori, H., Noma, S., Nozaki, T., Ogishima, S., Ohkura, N., Ohmiya, H., Ohno, H., Ohshima, M., Okada-Hatakeyama, M., Okazaki, Y., Orlando, V., Ovchinnikov, D.A., Pain, A., Passier, R., Patrikakis, M., Persson, H., Piazza, S., Prendergast, J.G.D., Rackham, O.J.L., Ramilowski, J.A., Rashid, M., Ravasi, T., Rizzu, P., Roncador, M., Roy, S., Rye, M.B., Saijyo, E., Sajantila, A., Saka, A., Sakaguchi, S., Sakai, M., Sato, H., Satoh, H., Savvi, S., Saxena, A., Schneider, C., Schultes, E.A., Schulze-Tanzil, G.G., Schwegmann, A., Sengstag, T., Sheng, G., Shimoji, H., Shimoni, Y., Shin, J.W., Simon, C., Sugiyama, D., Sugiyama, T., Suzuki, M., Suzuki, N., Swoboda, R.K., Hoen, P.A.C., Tagami, M., Takahashi, N., Takai, J., Tanaka, H., Tatsukawa, H., Tatum, Z., Thompson, M., Toyoda, H., Toyoda, T., Valen, E., Wetering, M., Berg, L.M., Verardo, R., Vijayan, D., Vorontsov, I.E., Wasserman, W.W., Watanabe, S., Wells, C.A., Winteringham, L.N., Wolvetang, E., Wood, E.J., Yamaguchi, Y., Yamamoto, M., Yoneda, M., Yonekura, Y., Yoshida, S., Zabierowski, S.E., Zhang, P.G., Zhao, X., Zucchelli, S., Summers, K.M., Suzuki, H., Daub, C.O., Kawai, J., Heutink, P., Hide, W., Freeman, T.C., Lenhard, B., Bajic, V.B., Taylor, M.S., Makeev, V.J., Sandelin, A., Hume, D.A., Carninci, P., Hayashizaki, Y., Consortium, T.F., RIKEN PMI, (DGT), C.: A promoter-level mammalian expression atlas. Nature **507**(7493), 462–470 (2014)

[4] Danko, C.G., Hyland, S.L., Core, L.J., Martins, A.L., Waters, C.T., Lee, H.W.,

Cheung, V.G., Kraus, W.L., Lis, J.T., Siepel, A.: Identification of active transcriptional regulatory elements from GRO-seq data. Nat Meth **12**(5), 433–438 (2015)

[5] Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M.S., Mapendano, C.K., Schierup, M.H., Jensen, T.H.: RNA exosome depletion reveals transcription upstream of active human promoters. Science **322**(5909), 1851–1854 (2008)

[6] Preker, P., Almvig, K., Christensen, M.S., Valen, E., Mapendano, C.K., Sandelin, A., Jensen, T.H.: PROMoter uPstream Transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters. Nucleic acids research **39**(16), 7179–7193 (2011)

[7] Andersson, R., Refsing Andersen, P., Valen, E., Core, L.J., Bornholdt, J., Boyd, M., Heick Jensen, T., Sandelin, A.: Nuclear stability and transcriptional directionality separate functionally distinct RNA species. Nat Commun **5** (2014)

[8] Kim, T.-k., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., Markenscoff-Papadimitriou, E., Kuhl, D., Bito, H., Worley, P.F., Kreiman, G., Greenberg, M.E.: Widespread transcription at neuronal activity-regulated enhancers. Nature **465**(7295), 182–187 (2010)

[9] De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B.K., Muller, H., Ragoussis, J., Wei, C.-L., Natoli, G.: A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS biology **8**(5), 1000384 (2010)

[10] Sigova, A.A., Mullen, A.C., Molinie, B., Gupta, S., Orlando, D.A., Guenther, M.G., Almada, A.E., Lin, C., Sharp, P.A., Giallourakis, C.C., Young, R.A.: Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. Proceedings of the National Academy of Sciences **110**(8), 2876–2881 (2013)

[11] Azofeifa, J.G., Dowell, R.D.: A generative model for the behavior of RNA polymerase. Bioinformatics **33**(2), 227–234 (2016)

[12] Danko, C.G., Choate, L.A., Marks, B.A., Rice, E.J., Wang, Z., Chu, T., Martins, A.L., Dukler, N., Coonrod, S.A., Tait Wojno, E.D., Lis, J.T., Kraus, W.L., Siepel, A.: Dynamic evolution of regulatory element ensembles in primate CD4+ T cells. Nature Ecology & Evolution **2**(3), 537–548 (2018)

[13] Allison, K.A., Kaikkonen, M.U., Gaasterland, T., Glass, C.K.: Vespucci: a system for building annotated databases of nascent transcripts. Nucleic acids research **42**(4), 2433–2447 (2014)

[14] Yao, L., Liang, J., Ozer, A., Leung, A.K.-Y., Lis, J.T., Yu, H.: A comparison of experimental assays and analytical methods for genome-wide identification of active enhancers. Nature Biotechnology, 1–10 (2022)

[15] Zhou, X., O'Shea, E.K.: Integrated approaches reveal determinants of genome-wide binding and function of the transcription factor Pho4. Molecular cell **42**(6), 826–836 (2011)

[16] Cusanovich, D.A., Pavlovic, B., Pritchard, J.K., Gilad, Y.: The functional consequences of variation in transcription factor binding. PLoS Genet **10**(3), 1004226 (2014)

[17] Savic, D., Roberts, B.S., Carleton, J.B., Partridge, E.C., White, M.A., Cohen, B.A., Cooper, G.M., Gertz, J., Myers, R.M.: Promoter-distal RNA polymerase II binding discriminates active from inactive CCAAT/enhancer-binding protein beta binding sites. Genome research **25**(12), 1791–1800 (2015)

[18] Azofeifa, J.G., Allen, M.A., Hendrix, J.R., Read, T., Rubin, J.D., Dowell, R.D.: Enhancer RNA profiling predicts transcription factor activity. Genome Research (2018)

[19] Rubin, J.D., Stanley, J.T., Sigauke, R.F., Levandowski, C.B., Maas, Z.L., Westfall, J., Taatjes, D.J., Dowell, R.D.: Transcription factor enrichment analysis (TFEA): Quantifying the activity of hundreds of transcription factors from a single experiment. Nature Communications Biology (2021)

[20] Wang, Z., Chu, T., Choate, L.A., Danko, C.G.: Identification of regulatory elements from nascent transcription using dREG. Genome Research **29**(2), 293–303 (2019)

[21] Kaikkonen, M.U., Spann, N.J., Heinz, S., Romanoski, C.E., Allison, K.A., Stender, J.D., Chun, H.B., Tough, D.F., Prinjha, R.K., Benner, C., Glass, C.K.: Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. Molecular Cell **51**(3), 310–325 (2013)

[22] Kristjánsdóttir, K., Dziubek, A., Kang, H.M., Kwak, H.: Population-scale study of erna transcription reveals bipartite functional enhancer architecture. Nature Communications **11**(1), 5963 (2020)

[23] Bae, S., Kim, K., Kang, K., Kim, H., Lee, M., Oh, B., Kaneko, K., Ma, S., Choi, J.H., Kwak, H., *et al.*: RANKL-responsive epigenetic mechanism reprograms macrophages into bone-resorbing osteoclasts. Cellular & Molecular Immunology **20**(1), 94–109 (2023)

[24] Chu, T., Rice, E.J., Booth, G.T., Salamanca, H.H., Wang, Z., Core, L.J., Longo, S.L., Corona, R.J., Chin, L.S., Lis, J.T., *et al.*: Chromatin run-on and sequencing maps the transcriptional regulatory landscape of glioblastoma multiforme.

Nature genetics **50**(11), 1553–1564 (2018)

[25] Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., Bagger, F.O., Jorgensen, M., Andersen, P.R., Bertin, N., Rackham, O., Burroughs, A.M., Baillie, J.K., Ishizu, Y., Shimizu, Y., Furuhata, E., Maeda, S., Negishi, Y., Mungall, C.J., Meehan, T.F., Lassmann, T., Itoh, M., Kawaji, H., Kondo, N., Kawai, J., Lennartsson, A., Daub, C.O., Heutink, P., Hume, D.A., Jensen, T.H., Suzuki, H., Hayashizaki, Y., Muller, F., Consortium, T.F., Forrest, A.R.R., Carninci, P., Rehli, M., Sandelin, A.: An atlas of active enhancers across human cell types and tissues. Nature **507**(7493), 455–461 (2014)

[26] Lidschreiber, K., Jung, L.A., Emde, H., Dave, K., Taipale, J., Cramer, P., Lidschreiber, M.: Transcriptionally active enhancers in human cancer cells. Molecular systems biology **17**(1), 9873 (2021)

[27] Lee, S.A., Kristjánsdóttir, K., Kwak, H.: eRNA co-expression network uncovers TF dependency and convergent cooperativity. Scientific Reports **13**(1), 19085 (2023)

[28] Edgar, R., Domrachev, M., Lash, A.E.: Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research **30**(1), 207–210 (2002)

[29] Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., *et al.*: NCBI GEO: archive for functional genomics data sets—update. Nucleic acids research **41**(D1), 991–995 (2012)

[30] Leinonen, R., Sugawara, H., Shumway, M., Collaboration, I.N.S.D.: The sequence read archive. Nucleic acids research **39**(suppl_1), 19–21 (2010)

[31] Lizio, M., Harshbarger, J., Shimoji, H., Severin, J., Kasukawa, T., Sahin, S., Abugessaisa, I., Fukuda, S., Hori, F., Ishikawa-Kato, S., *et al.*: Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome biology **16**(1), 1–14 (2015)

[32] Abugessaisa, I., Ramilowski, J.A., Lizio, M., Severin, J., Hasegawa, A., Harshbarger, J., Kondo, A., Noguchi, S., Yip, C.W., Ooi, J.L.C., *et al.*: FANTOM enters 20th year: expansion of transcriptomic atlases and functional annotation of non-coding RNAs. Nucleic Acids Research **49**(D1), 892–898 (2021)

[33] Consortium, E.P., *et al.*: An integrated encyclopedia of DNA elements in the human genome. Nature **489**(7414), 57 (2012)

[34] Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I.,

Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., *et al.*: The encyclopedia of DNA elements (ENCODE): data portal update. Nucleic acids research **46**(D1), 794–801 (2018)

[35] Luo, Y., Hitz, B.C., Gabdank, I., Hilton, J.A., Kagda, M.S., Lam, B., Myers, Z., Sud, P., Jou, J., Lin, K., *et al.*: New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic acids research **48**(D1), 882–889 (2020)

[36] Hitz, B.C., Lee, J.-W., Jolanki, O., Kagda, M.S., Graham, K., Sud, P., Gabdank, I., Strattan, J.S., Sloan, C.A., Dreszer, T., et al.: The ENCODE uniform analysis pipelines. bioRxiv, 2023–04 (2023)

[37] Gao, T., Qian, J.: EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. Nucleic acids research **48**(D1), 58–64 (2020)

[38] Consortium, G.: The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science **369**(6509), 1318–1330 (2020)

[39] Liu, Y., Chen, S., Wang, S., Soares, F., Fischer, M., Meng, F., Du, Z., Lin, C., Meyer, C., DeCaprio, J.A., *et al.*: Transcriptional landscape of the human cell cycle. Proceedings of the National Academy of Sciences **114**(13), 3473–3478 (2017)

[40] Everaert, C., Volders, P.-J., Morlion, A., Thas, O., Mestdagh, P.: SPECS: a non-parametric method to identify tissue-specific molecular features for unbalanced sample groups. BMC Bioinformatics **21**(1), 58 (2020)

[41] Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., Rinn, J.L.: Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes & development **25**(18), 1915–1927 (2011)

[42] Hariprakash, J.M., Ferrari, F.: Computational biology solutions to identify enhancers-target gene pairs. Computational and structural biotechnology journal **17**, 821–831 (2019)

[43] Xu, H., Zhang, S., Yi, X., Plewczynski, D., Li, M.J.: Exploring 3D chromatin contacts in gene regulation: the evolution of approaches for the identification of functional enhancer-promoter interaction. Computational and structural biotechnology journal **18**, 558–570 (2020)

[44] Wang, J., Zhao, Y., Zhou, X., Hiebert, S.W., Liu, Q., Shyr, Y.: Nascent RNA sequencing analysis provides insights into enhancer-mediated gene regulation. BMC genomics **19**(1), 1–18 (2018)

[45] Lettice, L.A., Heaney, S.J., Purdie, L.A., Li, L., Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E., Graaff, E.: A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Human molecular genetics **12**(14), 1725–1735 (2003)

[46] Azofeifa, J.G., Allen, M.A., Lladser, M.E., Dowell, R.D.: An annotation agnostic algorithm for detecting nascent RNA transcripts in GRO-seq. IEEE/ACM Transactions on Computational Biology and Bioinformatics **14**(5), 1070–1081 (2017)

[47] Mills, C., Muruganujan, A., Ebert, D., Marconett, C.N., Lewinger, J.P., Thomas, P.D., Mi, H.: PEREGRINE: a genome-wide prediction of enhancer to gene relationships supported by experimental evidence. PloS one **15**(12), 0243791 (2020)

[48] Sanyal, A., Lajoie, B.R., Jain, G., Dekker, J.: The long-range interaction landscape of gene promoters. Nature **489**(7414), 109–113 (2012)

[49] De Laat, W., Duboule, D.: Topology of mammalian developmental enhancers and their regulatory landscapes. Nature **502**(7472), 499–506 (2013)

[50] Marinić, M., Aktas, T., Ruf, S., Spitz, F.: An integrated holo-enhancer unit defines tissue and gene specificity of the Fgf8 regulatory landscape. Developmental cell **24**(5), 530–542 (2013)

[51] Lorberbaum, D.S., Ramos, A.I., Peterson, K.A., Carpenter, B.S., Parker, D.S., De, S., Hillers, L.E., Blake, V.M., Nishi, Y., McFarlane, M.R., *et al.*: An ancient yet flexible cis-regulatory architecture allows localized Hedgehog tuning by patched/Ptch1. Elife **5**, 13550 (2016)

[52] Hafner, A., Boettiger, A.: The spatial organization of transcriptional control. Nature Reviews Genetics **24**(1), 53–68 (2023)

[53] Schmidt, F., Marx, A., Baumgarten, N., Hebel, M., Wegner, M., Kaulich, M., Leisegang, M.S., Brandes, R.P., Göke, J., Vreeken, J., *et al.*: Integrative analysis of epigenetics data identifies gene-specific regulatory elements. Nucleic acids research **49**(18), 10397–10418 (2021)

[54] Moody, J., Kouno, T., Kojima, M., Koya, I., Leon, J., Suzuki, A., Hasegawa, A., Akiyama, T., Akiyama, N., Amagai, M., et al.: A single-cell atlas of transcribed cis-regulatory elements in the human genome. bioRxiv, 2023–11 (2023)

[55] Gasperini, M., Hill, A.J., McFaline-Figueroa, J.L., Martin, B., Kim, S., Zhang, M.D., Jackson, D., Leith, A., Schreiber, J., Noble, W.S., *et al.*: A genome-wide framework for mapping gene regulation via cellular genetic screens. Cell **176**(1), 377–390 (2019)

[56] Huynh-Thu, V.A., Irrthum, A., Wehenkel, L., Geurts, P.: Inferring regulatory networks from expression data using tree-based methods. PloS one **5**(9), 12776 (2010)

[57] Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., Groza, T., Güneş, O., Hall, P., Hayhurst, J., *et al.*: The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. Nucleic acids research **51**(D1), 977–985 (2023)

[58] Hunter, S., Sigauke, R.F., Stanley, J.T., Allen, M.A., Dowell, R.D.: Protocol variations in run-on transcription dataset preparation produce detectable signatures in sequencing libraries. BMC genomics **23**(1), 1–18 (2022)

[59] Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., *et al.*: Systematic localization of common disease-associated variation in regulatory DNA. Science **337**(6099), 1190–1195 (2012)

[60] Farh, K.K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W.J., Beik, S., Shoresh, N., Whitton, H., Ryan, R.J.H., Shishkin, A.A., Hatan, M., Carrasco-Alfonso, M.J., Mayer, D., Luckey, C.J., Patsopoulos, N.A., De Jager, P.L., Kuchroo, V.K., Epstein, C.B., Daly, M.J., Hafler, D.A., Bernstein, B.E.: Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature **518**(7539), 337–343 (2015)

[61] Langfelder, P., Horvath, S.: WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics **9**(1), 1–13 (2008)

[62] Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., *et al.*: The UCSC genome browser database: update 2006. Nucleic acids research **34**(suppl_1), 590–598 (2006)

[63] Hah, N., Danko, C.G., Core, L., Waterfall, J.J., Siepel, A., Lis, J.T., Kraus, W.L.: A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. Cell **145**(4), 622–634 (2011)

[64] Hah, N., Murakami, S., Nagari, A., Danko, C.G., Kraus, W.L.: Enhancer transcripts mark active estrogen receptor binding sites. Genome Research **23**(8), 1210–1223 (2013)

[65] Di Tommaso, P., Chatzou, M., Floden, E.W., Barja, P.P., Palumbo, E., Notredame, C.: Nextflow enables reproducible computational workflows. Nature biotechnology **35**(4), 316–319 (2017)

[66] Institute, J.G.: BBMap. https://sourceforge.net/projects/bbmap/ (2015)

[67] Kim, D., Langmead, B., Salzberg, S.L.: HISAT: a fast spliced aligner with low memory requirements. Nature methods **12**(4), 357–360 (2015)

[68] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R.: The sequence alignment/map format and SAMtools. Bioinformatics **25**(16), 2078–2079 (2009)

[69] Quinlan, A.R., Hall, I.M.: Bedtools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**(6), 841–842 (2010)

[70] Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., Mesirov, J.P.: Integrative genomics viewer. Nature Biotechnology **29**, 24 (2011)

[71] Simons, A.: A quality control tool for high throughput sequence data. Available online at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ **10**, 1000 (2010)

[72] Daley, T., Deng, C., Li, T., Smith, A.: The preseq manual (2014)

[73] Wang, L., Wang, S., Li, W.: RSeQC: quality control of RNA-seq experiments. Bioinformatics **28**(16), 2184–2185 (2012)

[74] Institute, B.: Picard toolkit. http://broadinstitute.github.io/picard/ (2019)

[75] Kent, W.J., Zweig, A.S., Barber, G., Hinrichs, A.S., Karolchik, D.: BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics **26**(17), 2204–2207 (2010)

[76] R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2019). R Foundation for Statistical Computing

[77] Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., Stetsenko, P., Short, T., Lianoglou, S., Antonyan, E., Bonsch, M., Parsonage, H., et al.: Package 'data.table'. Extension of 'data. frame' **596** (2019)

[78] Gonzales, N.M., Seo, J., Hernandez Cordero, A.I., St Pierre, C.L., Gregory, J.S., Distler, M.G., Abney, M., Canzar, S., Lionikas, A., Palmer, A.A.: Genome wide association analysis in a mouse advanced intercross line. Nature communications **9**(1), 1–12 (2018)

[79] Aragon, T.J., Fay, M.P., Wollschlaeger, D., Omidpanah, A., Omidpanah, M.A.: Package 'epitools' (2017)

[80] Van Rossum, G., Drake, F.L.: Python 3 Reference Manual. CreateSpace, Scotts Valley, CA (2009)

[81] Liao, Y., Smyth, G.K., Shi, W.: featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics **30**(7), 923–930 (2014)

[82] Li, B., Dewey, C.N.: RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC bioinformatics **12**(1), 1–16 (2011)

[83] Wagner, G.P., Kin, K., Lynch, V.J.: Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory in biosciences **131**(4), 281–285 (2012)

[84] Jolliffe, I.T.: Principal component analysis. Technometrics **45**(3), 276 (2003)

[85] Harris, C.R., Millman, K.J., Van Der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., *et al.*: Array programming with NumPy. Nature **585**(7825), 357–362 (2020)

[86] Reback, J., McKinney, W., Van Den Bossche, J., Augspurger, T., Cloud, P., Klein, A., Hawkins, S., Roeschke, M., Tratner, J., She, C., et al.: pandas-dev/pandas: Pandas 1.0.2. Zenodo (2020)

[87] McKinney: Data Structures for Statistical Computing in Python. In: Walt, Millman (eds.) Proceedings of the 9th Python in Science Conference, pp. 56–61 (2010)

[88] Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., *et al.*: SciPy 1.0: fundamental algorithms for scientific computing in python. Nature methods **17**(3), 261–272 (2020)

[89] Zhang, B., Horvath, S.: A general framework for weighted gene co-expression network analysis. Statistical applications in genetics and molecular biology **4**(1) (2005)

[90] Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) **57**(1), 289–300 (1995)

[91] Yokoshi, M., Segawa, K., Fukaya, T.: Visualizing the role of boundary elements in enhancer-promoter communication. Molecular cell **78**(2), 224–235 (2020)

[92] Cavalheiro, G.R., Pollex, T., Furlong, E.E.: To loop or not to loop: what is the role of TADs in enhancer function and gene regulation? Current Opinion in Genetics & Development **67**, 119–129 (2021)

[93] Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang,

Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., *et al.*: Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic acids research **44**(W1), 90–97 (2016)

[94] Ramos, E.M., Hoffman, D., Junkins, H.A., Maglott, D., Phan, L., Sherry, S.T., Feolo, M., Hindorff, L.A.: Phenotype–Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. European Journal of Human Genetics **22**(1), 144–147 (2014)

[95] Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., Furlong, L.I.: DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic acids research, 943 (2016)

[96] Kundu, I., Sharma, M., Barai, R.S., Pokar, K., Idicula-Thomas, S.: GeDiPNet: Online resource of curated gene-disease associations for polypharmacological targets discovery. Genes & Diseases **10**(3), 647 (2023)

[97] Hamosh, A., Scott, A.F., Amberger, J., Valle, D., McKusick, V.A.: Online Mendelian inheritance in man (OMIM). Human mutation **15**(1), 57–61 (2000)

[98] Wickham, H.: ggplot2: Elegant Graphics for Data Analysis, (2016)

[99] Wilke, C.O.: cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. (2020)

[100] Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., Pfister, H.: UpSet: visualization of intersecting sets. IEEE transactions on visualization and computer graphics **20**(12), 1983–1992 (2014)

[101] Lopez-Delisle, L., Rabbani, L., Wolff, J., Bhardwaj, V., Backofen, R., Grüning, B., Ramírez, F., Manke, T.: pyGenomeTracks: reproducible plots for multivariate genomic datasets. Bioinformatics **37**(3), 422–423 (2021)

[102] Ramírez, F., Bhardwaj, V., Arrigoni, L., Lam, K.C., Grüning, B.A., Villaveces, J., Habermann, B., Akhtar, A., Manke, T.: High-resolution TADs reveal DNA sequences underlying genome organization in flies. Nature communications **9**(1), 189 (2018)

[103] Kramer, N.E., Davis, E.S., Wenger, C.D., Deoudes, E.M., Parker, S.M., Love, M.I., Phanstiel, D.H.: Plotgardener: cultivating precise multi-panel figures in R. Bioinformatics **38**(7), 2042–2045 (2022)

[104] Team, B.C., Maintainer, B.P.: Txdb.hsapiens.ucsc.hg38.knowngene (2019)