BMC Genomics

DATABASE

Open Access

Atlas of nascent RNA transcripts reveals tissue-specific 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] and Robin D. Dowell[1,2,3*]

Abstract

Gene transcription is controlled and modulated by regulatory regions, including enhancers and promoters. These regions are abundant in non-coding bidirectional transcription that results in generally unstable RNA. Using nascent RNA transcription data across hundreds of human samples, we identified over 800,000 regions containing bidirectional transcription. We then identify tissue specific, 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 a database called DBNascent (https://nascent.colorado.edu/) 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 [1]. 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 [2, 3]. 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 [2, 4]. These bidirectional signatures are observed at not only protein-coding genes but also transcribed regulatory regions (TREs) [2, 5, 6]. At genes, the upstream antisense RNA (uaRNA) is also referred to as a promoter upstream transcript (PROMPTs) [7, 8]. At regulatory regions, the two transcripts are often referred to as enhancer-associated RNAs (eRNAs) [8–10]. In every case, these non-gene transcripts are lowly transcribed and unstable. Hence numerous methods have been developed to identify sites of bidirectional transcription directly from run-on data [6, 11–13]. Despite the availability of these methods, bidirectional transcription regions largely remain unannotated.

Regardless the function of the transcripts produced in these regions, they have been shown to serve as excellent markers of regulation within the local genomic context [14]. While transcription factor binding is measured by assays such as chromatin immunoprecipitation, not all instances of transcription factor binding result in altered

*Correspondence:
Robin D. Dowell
robin.dowell@colorado.edu
[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

gene regulation nearby. However signatures of RNA polymerase activity near transcription factor binding sites effectively reflect the active subset of binding events [15–17]. In support of this notion, changes in bidirectional transcription activity (locations and levels) can be used to infer changes in transcription factor activity between two conditions [9, 17–21]. Consequently, nascent run-on data also provides unique insights into transcription factor activity [22].

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. Additionally, we used correlation analysis [23, 24] to link enhancers to their target genes, finding that most genes are regulated by enhancers in a tissue-specific manner. Our resulting atlas of nascent sequencing datasets, identified sites of bidirectional transcription and gene-bidirectional links serves as valuable resources 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 datasets (Fig. 1). To this end, nascent run-on RNA sequencing experiments were manually curated from Gene Expression Omnibus (GEO) [25, 26] and the NIH Sequence Read Archive (SRA) [27]. We excluded metabolic labeling techniques because 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

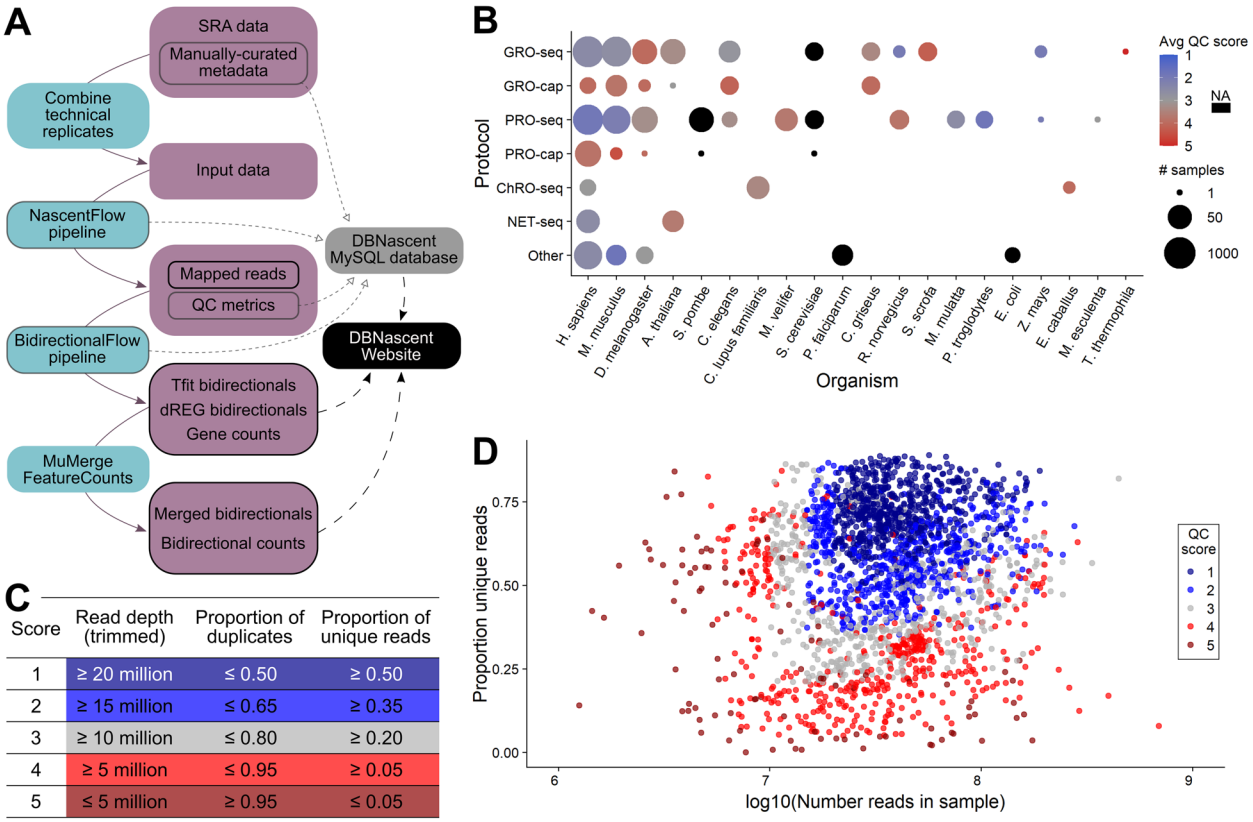

Fig. 1 Overview of DBNascent. **A** Data were derived from Sequence Read Archive fastq files and manually curated metadata. Technical replicate fastq files were combined, then data were processed to obtain metrics on quality, bidirectional regions, and read counts. Metadata, quality control metrics, 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

all samples from their associated database information and/or publication (see Supplementary Table 1). This metadata was collected in a MySQL database (hereafter DBNascent) where all treatment condition times were annotated in reference to the time of cell harvest. The raw fastq files were processed through standardized Nextflow pipelines (Fig. 1A) that include steps to map, quality control, and identify 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 in 20 organisms, collected from 287 projects, which consisted of journal articles or Gene Expression Omnibus (GEO) datasets (Fig. 1B). The samples were subjected to extensive quality control (QC), from which we developed a QC classification metric based on both read depth and complexity (Fig. 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 otherwise specified. 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 Fig. S1).

Of the 2,880 samples in DBNascent, the vast majority (2,387) were derived from human or mouse cells (Fig. 1B), and these were exclusively used for downstream analysis, e.g. identifying bidirectional regions. The samples were distributed across 19 and 10 tissues from human and mouse, respectively. In both organisms, samples were collected mainly from cell lines or cultured primary cells (Supplementary Fig. S2). Furthermore, a principal 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 the underlying biological signal more than technical variation (Supplementary Fig. S3).

Bidirectional regions in DBNascent overlap cis-regulatory elements

Nuclear run-on assays, such as GRO-seq and PRO-seq, give a readout of transcription from all cellular RNA polymerases. Consequently, they recover signal at both coding and noncoding regions, much of which is not annotated. Two methods for identifying transcribed regulatory regions are Tfit and dREG [6, 11, 12]. Tfit uses a mathematical model of RNA polymerase II (PolII) 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 Supplementary Methods for complete details). For each of the 1,638 human and 750 mouse samples analyzed, on average $\sim$ 25,000 bidirectional regions were identified by Tfit and $\sim$ 18,500 by dREG (Supplementary Fig. S4A). Bidirectional calls were then combined using a modified version of *muMerge* (version 1.1.0) (see Materials and Methods Section) [18]. The merging strategy was performed hierarchically, merging between experiments first, then cell types, and finally the bidirectional calling methods (Fig. 2A). Since the resolution of Tfit calls at the initiation of RNA polymerase (typically the center region of bidirectional transcription) is better than dREG [13], the coordinates of Tfit calls were used when the two callers overlap (Supplementary Fig. S4B-C). The selected regions were then filtered to retain high quality regions, based on the data's QC score (Supplementary Fig. S5).

Throughout the genome, 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 Fig. S6) and are similarly distributed throughout the genome (Supplementary Figs. S7 and S8). Most bidirectional regions overlap noncoding regions, while a smaller percentage overlap exons (Fig. 2B, Supplementary Figs. S9 A and S10). In exonic regions, most bidirectional regions overlap with the 3′UTR and beginning of the CDS region (Supplementary Fig. S10). Identifying the number of human genes with TSS bidirectional signals, we find that about 80% have a bidirectional call in their promoter region, and genes without a bidirectional call at their TSS were not transcribed (Supplementary Tables 2 and 3). Outside the promoter region, bidirectional regions are uniformly found across annotated transcripts in both mouse and human (Supplementary Fig. S11). Bidirectional regions within the gene are mostly intronic, with some overlapping the boundaries with an annotated exon (Fig. 2C and Supplementary Figs. S9B and S12). Similarly, while most genes and their isoforms did not have bidirectional regions overlapping any annotated exon, about 30% of isoforms had bidirectional regions overlapping their CDS and/or 3′UTR (Supplementary Fig. S13). Few isoforms had bidirectional regions within the 5′UTR, even when considering the gene TSS bidirectionals.

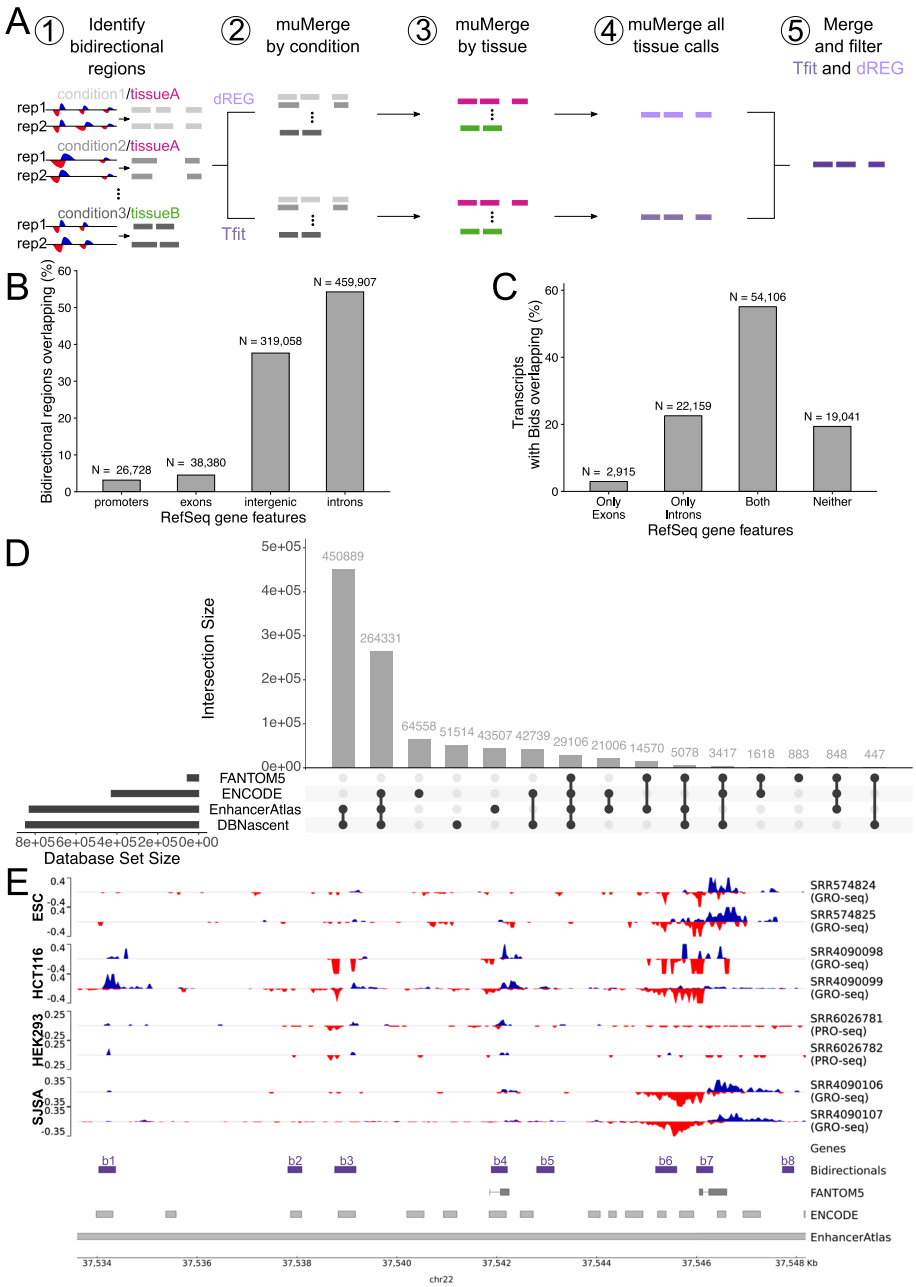

Fig. 2 Identification and characterization of bidirectional regions. **A** Schematic showing how bidirectional regions were identified and merged to give a consensus annotation set. Briefly, (1) bidirectional regions were inferred from nuclear run-on coverage data in each sample. (2) In a given experiment, regions were combined using muMerge [18] based on treatment conditions for both Tfit [11] and dREG [6, 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) finally 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 [28–31], FANTOM5 [5, 32–34] and EnhancerAtlas [35]). **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

To assess the quality of our called regions, we next compared our bidirectional calls to annotated candidate cis-regulatory elements (cCRE) from ENCODE, EnhancerAtlas and FANTOM5, as these resources annotate regulatory regions using a variety of techniques [5, 28–35] (Fig. 2D and Supplementary Fig. S14). ENCODE offers a characterization of candidate cis-regulatory elements based on histone marks, DNase and CTCF signal [28, 29]. While the FANTOM5 project identifies sites of transcription initiation primarily using CAGE (Cap Analysis of Gene Expression) data [34]. Lastly, EnhancerAtlas aims to combine assorted genomic data, including ENCODE and FANTOM5 as well as nascent RNA sequencing data [32, 33, 35]. Overall, about 40% to 60% of the cis-regulatory elements in these data resources are found in DBNascent (Supplementary Fig. S14A and B). Interestingly, 29,106 human and 21,999 mouse bidirectional regions are contained in all four databases (Fig. 2D and Supplementary Fig. S14C). 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 much wider compared to the other database regions, therefore yielding greater overlaps (Fig. 2E and Supplementary Fig. S15). Notably, EnhancerAtlas includes nascent RNA data and RNA polymerase II ChIP-sequencing data in its construction, which may contribute to both the observed long region length and the overlap with our called bidirectional regions. The overlap of bidirectional regions and regulatory categories as defined by ENCODE histone marks showed a significant overlap for most cCRE (p-value < 0.01, by both Fisher's Exact Test and empirical shuffling test) (Supplementary Figs. S16 and S17). However, human CTCF and DNase-H3 K4 me3 cCRE overlaps with bidirectional regions were not significant. 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 on the basis of large scale genome-wide association studies. In particular, the GTEx consortium examined genome variation for its ability to influence expression levels [36]. As sites of bidirectional transcription are often genetic enhancers, we next considered to what extent our bidirectional calls overlap with GTEx identified variation. Although only a small number of GTEx variation resides within our bidirectional regions (Supplementary Fig. S14A), we found that bidirectional regions showed a higher odds for containing significant expression quantitative trait loci (eQTL) variants compared to non-significant variants (Supplementary Fig. S18) [36]. This further supports previous work showing an enrichment of eQTLs in enhancer regions [6].

Next, we used the 447-way alignment from Zoonomia [37–39] to ask if human and mouse bidirectional regions were conserved. We first asked what fraction of bidirectional regions can be lifted over between mouse and human using these alignments, finding that only 53.6% of human bidirectional regions lifted over to the mouse genome. We next used these lift over regions to ask whether bidirectional regions were transcribed across the species. To this end, we first examined bidirectional regions identified across matched cell types, identifying three cell types, embryonic stem cells (ESC), hematopoietic progenitor cells (HPC) and $CD4^+$ T cells [12, 20, 40–42]. Within these matched tissues we compared the reciprocal lift over bidirectional regions and found that over 64% of regions were both identifiable and transcribed (> 1 TPM) across the species (Supplementary Fig. S19A). We next expanded this analysis to the entire database, considering all bidirectional regions we find only 38.4% of mouse regions could be lifted over and were transcribed (sum > 1 TPM across all samples) in any human and mouse sample (30.8% of human regions in mouse). We then asked how these corresponding regions compared in transcription levels, finding the average transcription levels were more similar than the maximum level (Supplementary Fig. S19B and C). Consistent with prior work [12, 38, 43], we find that the majority of conserved regions were promoter associated with the intergenic bidirectional regions being the lowest conserved (Supplementary Fig. S19D).

Tissue specificity of transcription

The drop in conservation observed across the larger database, where many cell types exist uniquely in only one species, led to speculation that the pattern could be driven by tissues specific bidirectional regions. Therefore, we next sought to determine how transcription levels varied across cell types and tissues in different types of transcribed regions. To this end, we first examined a single representative high-quality dataset [44]. In this sample, promoter bidirectional regions were, in general, the most highly transcribed, followed by both coding and annotated noncoding genes (Fig. 3A). Collectively, the exonic, intronic, and intergenic bidirectional regions (non-promoter bidirectional regions), which tend to be enhancers, are more lowly transcribed. This pattern held 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 (Fig. 3B).

We next 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 with 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 [45], which can accommodate uneven sample sizes across groups. The SPECS score ranges from 0 (indicating depletion) to 1 (indicating enrichment), with a ubiquitously transcribed gene scoring around 0.5. Taking into account all genes and bidirectional regions, the distribution of SPECS scores showed a larger proportion of bidirectional regions having lower SPECS scores, indicating that they are more likely to show higher transcription in a limited set of tissues and low (or no) transcription across all others (Supplementary Fig. S20). 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 number of tissue-specific genes and bidirectional regions (Supplementary Fig. S21).

Subsequently, we evaluated the transcription change between the most specific tissue (highest SPECS score) and next highest scoring tissue (Fig. 3C and Supplementary Fig. S22). The resulting fold change should be large for each transcript that is transcribed primarily in a single tissue. We observed a skew towards higher fold changes for non-promoter bidirectional regions as compared to genes. Bidirectional calls at promoters showed a pattern indistinguishable from coding genes, whereas annotated noncoding genes seemed to show more tissue specificity than coding genes, in line with previous work [46]. 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. These findings were also reproduced with the expression specificity (ESS) scores approach [47] where we evaluated the median transcription levels in each tissue across the samples (Supplementary Fig. S23).

The SPECS score and ESS analyses suggest that non-promoter bidirectional calls, primarily those 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 Fig. S24, Supplementary Video S25). 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 (Fig. 3D). By this measure, the coding regions and the bidirectional regions of the promoter are most likely to be ubiquitously transcribed, while the bidirectional intronic and intergenic regions are more likely to show tissue-specific transcription, consistent with previous reports [23]. Thus, bidirectional regions, both intronic and intergenic, are transcribed and active in a small range of tissues compared to both the coding and noncoding genes they regulate.

Tissue specific correlation analysis identifies putative bidirectional and gene pairs

To better understand how tissue-specific transcribed regulatory elements could lead to ubiquitous transcription patterns at the genes they target, we next sought to link enhancers to their target genes [32, 48–50]. Prior work indicates that nascent transcription levels between enhancers and their known target genes – as determined by 3D data – have correlated transcription levels [8, 23, 24, 51], providing strong functional activity information for candidate linkages. Thus, we wondered whether strong correlations between genes and regions of bidirectional transcription within our run-on database could identify candidate tissue-specific enhancer to target gene linkages.

To this end, we calculated pairwise correlations between genes and bidirectional regions within each chromosome and identified significantly correlated pairs in a tissue-specific manner for human samples (Fig. 4A) (see Supplementary Methods). Importantly, for regions of bidirectional transcription that reside within introns, we only utilized counts from the antisense strand to avoid confounding signals with the gene. We then selected the eleven tissues with more than ten samples for the subsequent correlation analysis (Supplementary Fig. S26). Across this collection, we found 6,700,460 unique pairs

(See figure on next page.)

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 [45]. The full plot can be viewed in S22. **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

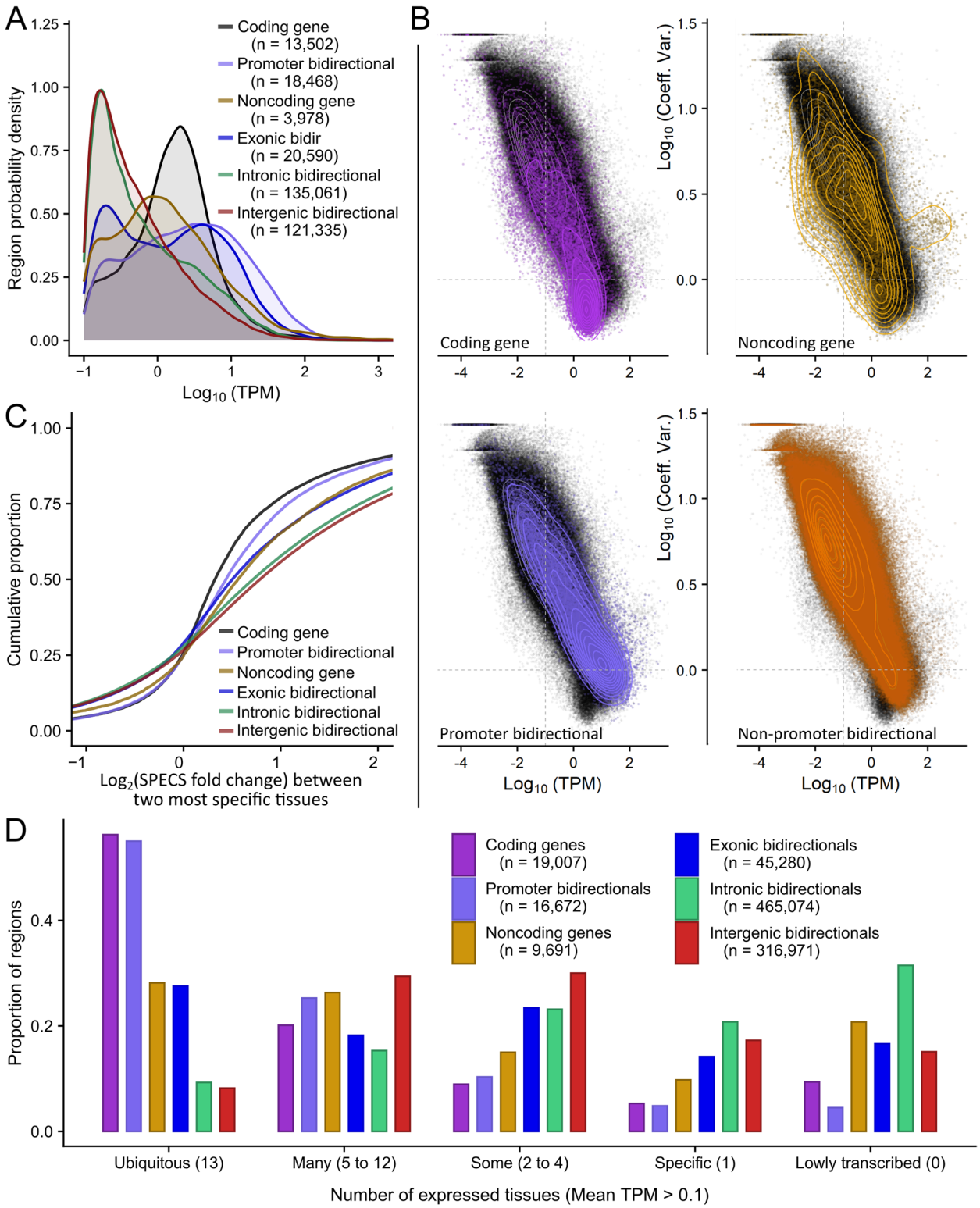

Fig. 3 (See legend on previous page.)

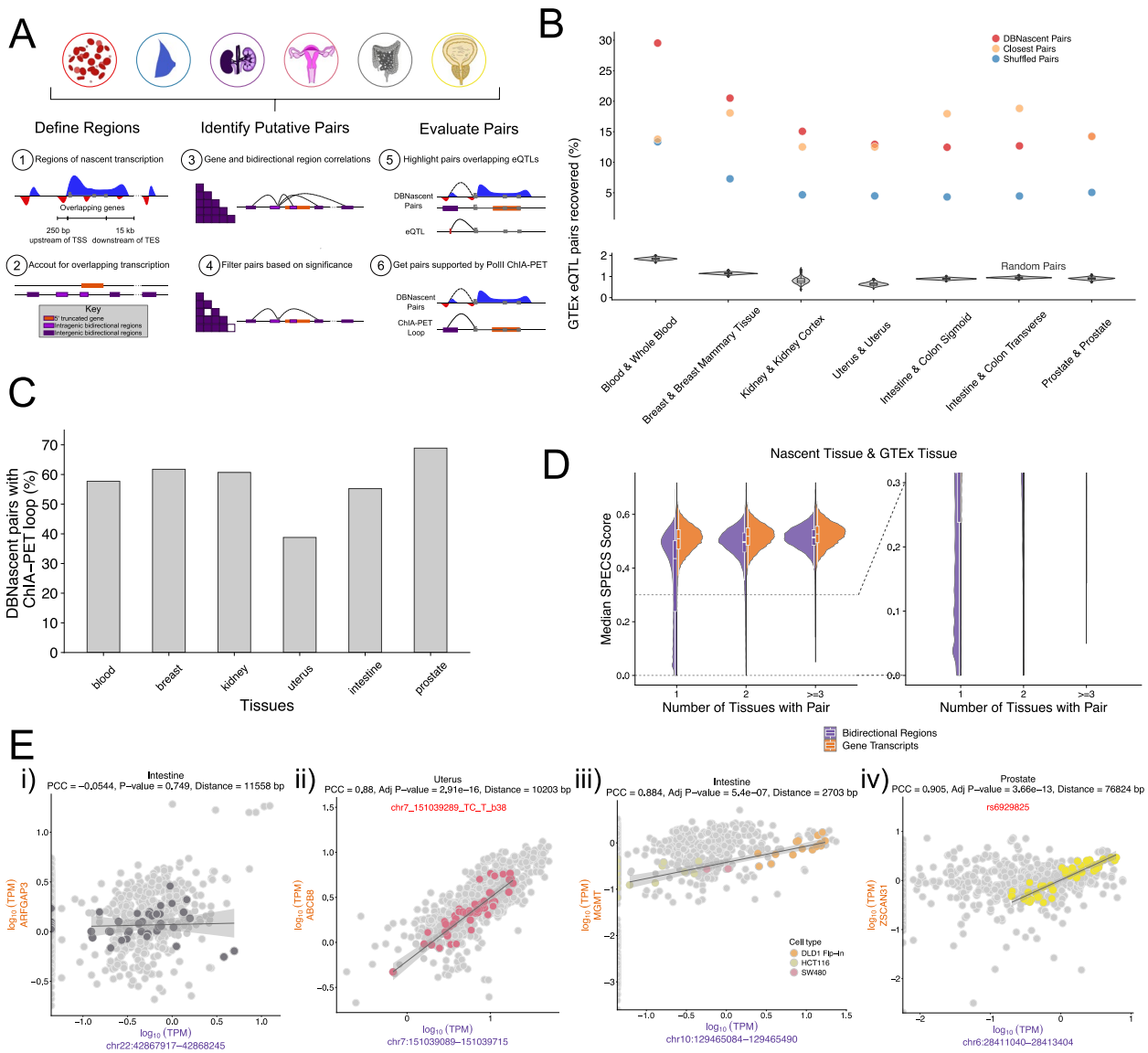

Fig. 4 DBNascent gene and bidirectional region pairs across tissues. **A** Schematic of methods used to identify gene and bidirectional DBNascent pairs. Briefly, (1) regions of bidirectional transcription are identified and (2) quantified. Normalized counts are then (3) correlated (within 1MB of the gene TSS) and pairs are (4) filtered based on FDR. Finally, pairs are evaluated based on recovery of (5) GTEx eQTLs and (6) RNA polII ChIA-PET loops. **B** Percent of GTEx eQTLs recovered by DBNascent pairs (red), closest gene assignment (orange), shuffled transcript pairs (blue) and random pairs (gray violin plots). The overlaps were performed on matched tissues (i.e. DBNascent Pair tissues and GTEx eQTL tissues). **C** DBNascent pairs supported by RNA PolII ChIA-PET loops across tissues. **D** Distribution of median SPECS Scores for genes (orange) and bidirectional regions (purple) split based on the number of tissues a pair is identified. The left panel zooms in to the lower median SPECS scores (0 - 0.3), showing lower median SPECS Score for bidirectional regions for pairs found in one tissue. **E** Example scatter plots of the raw data for interacting pairs: (i) an uncorrelated pair ARFGAP3 and chr22:42867917–42868245 in intestine tissues, (ii) ubiquitously correlated pair ABCB8 and chr7:151039089–151039715 (iii) cell-type specific interaction between MGMT and chr10:129465084–129465490, and (iv) a tissue specific interaction in prostate samples between ZSCAN31 and chr6:28411040–28413404. Samples in each tissue/cell-type are highlighted in color (Intestine, Uterus, Intestine cell-types, and Prostate respectively)

where the levels of nascent transcription at the gene correlated with levels at a bidirectional region within 1 Mbp (Benjamini Hochberg false discovery rate (FDR), p-adj < 0.01 and supported by 3 or more samples). In total, 96.3% of genes are linked to a bidirectional region, while 69.5% of the bidirectional regions have links. While not a constraint of the approach, we found that most bidirectional regions within the correlated pair were physically close to the gene TSS (Supplementary Fig. S27).

Next, we sought to evaluate our recovered correlated pairs by comparison to collections of known enhancer to gene linkages. Given that GTEx eQTLs are collected across several tissues, this offers a resource to evaluate correlated pairs in a tissue-specific manner. We found that a total of 104,388 DBNascent pairs overlapped with GTEx pairs, using matched tissues between the two methods. Depending on the tissue, the range of overlapping eQTL pairs was 6% to 29.5%, while randomized pairs were never more than 2% (Fig. 4B and Supplementary Fig. S28). In most cases, the matched bidirectional region was near the gene TSS (Supplementary Fig. S29). Additionally, our DBNascent pairs recover more known pairs than the closest gene strategy (Fig. 4B), similar to other gene to enhancer studies [52–55]. Interestingly, we found better recovery of eQTLs in matched tissues, suggesting we are indeed capturing tissue-specific interactions with DBNascent pairs (Wilcoxon rank sum test W = 1058, *p*-value = 2.513e- 06) (Supplementary Fig. S30A). However, we also noted that the tissues with fewer samples had lower recovery of GTEx pairs (Supplementary Fig. S28), suggesting that the method may be sensitive to the quantity of data utilized in constructing the correlations. Additionally, we examined the overlap of nascent derived pairs to experimentally validated enhancer and gene linkages obtained in K562 cells [56]. We observed a significant recovery of known interactions (Supplementary Fig. S31). In total, we recovered 79% of crisprQTL pairs with DBNascent pairs compared to 2.5% with randomized pairs. In general, comparisons with known enhancer and gene links suggest that most of our identified pairs are supported by orthogonal methods.

In all tissues, we recover both previously identified linked pairs and new regions of high correlation (Supplementary Figs. S32 and S33). A total of 38.8% of annotated genes have links that overlap GTEx eQTLs (Supplementary Fig. S33). For example, HCP5B is a long noncoding RNA that shows a high number of interactions (184 HCP5B and bidirectional pairs within 1 Mbp of the TSS), and 45.65% of these pairs also overlap eQTLs from GTEx (Supplementary Fig. S34) [36]. The extent to which newly identified high-correlating pairs arise from biological signals or spurious correlations is unclear. Therefore, we next sought to estimate the extent of spurious correlations.

To this end, we used a randomization strategy, reasoning that most biologically meaningful correlation would break down if the data were randomly selected 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. Overall, 82.5% of genes have more bidirectional regions assigned in DBNascent pairs compared to the shuffled chromosome pairs (Supplementary Fig. S35). We then compared links between the shuffled chromosome pairs and DBNascent pair, finding that DBNascent pairs had far more assigned pairs supported by eQTLs (1.5% to 13.3 % of eQTls recovered compared to 6% to 29.5% with DBNascent pairs) (Fig. 4B and Supplementary Fig. S28), further supporting that our assigned pairs contain signal. 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. Hence our randomization likely underestimates the proportion of real biological induced correlation relative to spurious correlations.

To reduce false positive pairs and spurious correlations, we next took advantage of GTEx eQTLs and compared the recovery of eQTLs at varying adjusted *p*-value cutoffs on the significance of correlations. We found that the most optimal adjusted *p*-value cut-off across all the tissues assessed was < 0.001, according to the recovery rate (Supplementary Fig. S36). Using this stringent false discovery rate reduced the number of unique pairs to 4,853,276 across the eleven tissues. Within this set, the number of gene transcripts and bidirectional regions in at least one pair were 27,564 (95.4% of all genes) and 469,317 (55.4% of annotated bidirectional regions), respectively. Consequently, even at the stringent false discovery rate, we recover correlated interactions for the majority of gene transcripts and bidirectional regions.

Importantly, the current state of the art methods for linking enhancers to their target genes incorporate both functional activity information and three-dimensional contact data [53–55]. Consequently, we next combined the stringent pairs list with available pairs from PolII ChIA-PET data across all tissues. Most of the DBNascent pairs (55.7% of all pairs) are supported by PolII ChIA-PET (Fig. 4C). Notably, overlaps with ChIA-PET loops were better within the same tissue than between tissues, highlighting tissue-specific interactions within DBNascent pairs (Wilcoxon rank sum test W = 11029, *p*-value = 0.0056) (Supplementary Fig. S30B). As noted previously, tissues with low sample numbers exhibited the poorest overlap with ChIA-PET data (Fig. 4B, C and Supplementary Fig. S37), further suggesting that sample size (number of nascent samples) limits the quality of the correlation-based analysis.

Consequently, we summarize the 3D supported DBNascent pairs using only the six tissues with the highest sample counts, where all data suggest the highest quality signal. Across these pairs, the median number of bidirectional regions assigned to a gene was 83 (Supplementary Fig. S38A). However, within the context of a single tissue, many fewer bidirectional regions are linked to any given

gene (median = 4–40; Supplementary Fig. S39). The tissue specific estimate is on par with other estimates of the number of enhancers linked to a gene [57–62]. In the other direction, the median number of genes assigned to a bidirectional transcript across all tissues was three, implying that a single bidirectional has only a few potential, often tissue-specific gene targets (Supplementary Fig. S38B; Supplementary Fig. S40).

Collectively, these results indicate that a gene's transcription is tuned by a small number of bidirectional regions, many of which are tissue-specific (Fig. 4D and Supplementary Fig. S41). A more detailed examination of individual pairs identifies numerous interesting patterns (Fig. 4E and Supplementary Fig. S32). In some cases, the bidirectional region and gene are correlated across all tissues (see Fig. 4E, ii), whereas in others the correlation appears either cell-type or tissue specific (Fig. 4E, iii-iv). When assessing the number of tissues that support a pair, only approximately 40% of pairs were supported by two or more tissues (Supplementary Fig. S42). This suggests that a gene is controlled by distinct TREs in different cell types. Consistent with this finding, when p53 was stimulated across a variety of tissues, a previous paper [63] found that the genes responding to p53 across the tissues were largely consistent but that the transcribed regulatory elements (in our case, bidirectional signals) were much more cell-type specific.

Discussion

Here we present DBNascent, an atlas that catalogs previously published nascent sequencing data, with an emphasis on run-on data. Additionally, within each dataset we identify sites of bidirectional transcription, which occur at both the 5′ end of protein coding genes (the TSS) and at transcribed regulatory elements. Our collection of uniformly processed nascent sequencing data allowed the identification of a large collection of sites of bidirectional transcription. Sites of bidirectional transcription were randomly distributed across the genome, lowly transcribed, variably conserved and often cell type specific. Many of the bidirectional regions were intronic, where quantification of transcription levels was estimated using only the antisense strand transcript to avoid confounding signal from the overlapping gene. The resulting set of identified bidirectional regions represents a large collection of transcribed regulatory elements that have utility in various subsequent studies. In fact, they have already proven useful in both studies of transcription regulation [22, 64] and technical studies such as comparisons of run-on protocols [65] and spike-in controls [66].

We leverage our large collection of non-promoter bidirectional regions to confirm and expand on the tissue specificity of enhancers. Long noncoding RNAs have long been known to be more cell type specific than protein coding genes [67]. More recently, it has been reported that enhancer associated RNAs [55] are also more cell-type specific than protein coding genes. Here we compare all three types of transcripts and find that enhancer-associated transcripts are more tissue specific than annotated lncRNAs. Given that many bidirectional regions are enhancers that regulate genes (both protein coding and lncRNAs), these results suggest that a gene is likely regulated by distinct sets of enhancers in each tissue.

We sought to utilize correlation between enhancer and target gene transcription levels [17, 23, 24, 32] to assign cis-regulatory regions to likely target genes. However, the high tissue specificity of many bidirectional regions complicated the correlation calculation. Since bidirectional regions are generally lowly transcribed and tissue-specific, they are less likely to be captured in low coverage datasets. To avoid being overly influenced by a lack of data, correlations were calculated only on samples with transcription for both the gene and bidirectional region. The result is fewer data points, particularly in tissues with fewer samples. We further found that fewer samples led to more spurious correlation, as seen in the lower recovery rate observed in comparisons between DBNascent pairs and GTEx eQTLs. Despite these issues, the gene-bidirectional links identified in well-represented tissues were often confirmed by orthogonal data, supporting the use of these data to investigate regulatory region assignment.

The gene-bidirectional links identified have several uses. The number of gene-bidirectional links identified further supports the idea that a gene can be ubiquitously transcribed and yet regulated by distinct enhancers in each tissue. Correlation-based strategies have long been important in the development and identification of gene regulatory networks [68]. Our links are an excellent starting point for building regulatory networks. More recently, correlation information has also been shown to improve curation of bona fide 3D interactions [55]. Finally, it is well known that many disease-associated variants occur in noncoding regions of the genome, often in regions associated with bidirectional transcription [6, 69, 70]. For example, bidirectional regions have recently been used to pinpoint functional genetic variants linked to asthma [71]. Hence the gene-bidirectional links provide one mechanism for identifying candidate gene targets for regulatory noncoding variation.

Materials and methods

Detailed methods can be found in the Supplementary Methods, which includes technical details, full citations for all data utilized, and version numbers for all software

utilized. Mouse samples were mapped to the mm10 reference genome and human samples to the hg38 genome.

Nascent RNA sequencing experiments metadata collection

Nascent RNA sequencing experiments were manually curated from the Gene Expression Omnibus (GEO) [25, 26] and the Sequence Read Archive (SRA) [27]. All treatment conditions were annotated with reference to the cell harvest time. Full details on the data curated, papers utilized, and accession numbers are provided in Supplementary Methods.

Data processing summary

After downloading 3,638 original SRR entries from the SRA, technical replicates were combined to result in 2,880 samples. All samples were subsequently trimmed, mapped to the corresponding reference genome, and assessed for sample quality using standardized NextFlow pipelines.

Three specific metrics were used to classify samples into quality 'tiers' for filtering purposes: read depth after trimming, proportion of duplicates, and complexity. Thresholds were determined to classify samples into one of five tiers (see Fig. 1C-D), and analysis was performed on samples within tiers 1–3, unless specified otherwise. Additionally, a minimum expected GC content for called bidirectionals was required, as prior work established that regions of bidirectional transcription have a GC bias centered at sites of transcription initiation [17, 22]. Details on each step of the data pre-processing are provided in the Supplementary Methods.

Identifying bidirectional transcripts summary

Regions of nascent transcription were identified using Tfit [11] and dREG [6, 19]. Identification of regions of transcription with dREG followed the recommended pipeline (per dREG GitHub) where uniquely mapped reads were used to generate BigWig input files. 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 removed.

Regions (from replicates, conditions, and bidirectional calling methods) were merged using a modified version of *muMerge* (now v1.1.0) (see Supplementary Methods for full details) [18, 72]. A few called regions were subsequently removed because of unusual size, often in regions of convergent transcription of nearby genes. The full identification, merging and filtering pipeline can be found on GitHub [73]. Additional details on every step are provided in the Supplementary Methods.

Summary of transcription

All summary statistics were performed using R, the R package data.table, and bedtools [74, 75].

Annotated bidirectional transcripts were overlapped with RefSeq genome features (e.g. exons, introns, UTRs, CDS) using `bedtools intersect`. Across all features, the minimum fraction of overlap required per region was 0.5. 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

Gene transcription start sites (TSS) bidirectional regions were identified by overlapping bidirectional calls with a 1kb window around the TSS (± 500bp) using `bedtools intersect`. To further assess how many genes had bidirectionals within specific gene features (e.g. exons, introns, UTRs, CDS), we performed the same overlapping with RefSeq genome feature components. The percentage of transcripts (either noted as genes or isoforms) with bidirectionals overlapping only their introns, exons, neither, or both were subsequently calculated. Metagene plots of bidirectional location were determined based on normalized gene transcript lengths. More stringent filters were applied when assigning intronic and exonic labels, see the Supplementary Methods for full details.

Bidirectional transcript regions overlapping cCREs

Regions of bidirectional transcription were overlapped with candidate cis-regulatory elements (cCREs) databases (ENCODE, EnhancerAtlas, FANTOM5 and eQTL data) using `bedtools intersect` [5, 28, 34, 35]. Regions of overlap were calculated using the minimum overlap of 1bp. The percent overlap was calculated with respect to the database size. See Supplementary Methods for further details. Evaluation of the overlap with ENCODE was performed with bedtools using the Fisher's Exact Test and a permutation approach [75–77].

Conservation of regions of bidirectional transcription

Human bidirectional regions from DBNascent were lifted over to the mouse genome using the alignment from Zoonomia with `halLiftover` [37–39]. Lifted over regions that were within 1kb were merged to a contiguous region. We then identified the closest called bidirectional region in mouse.

The mouse and human reciprocal `halLiftover` was performed on embryonic stem cell (SRZ7741177, SRR10669536), Hematopoietic progenitor cells (SRR930717, SRR5508393) and CD4+ T cells (SRR4012429, SRR4012393) [12, 20, 40–42]. Conservation of regions of bidirectional

transcription required > 1 TPM in the other species. Lifted regions that were within 1kb were merged and regions greater than 100bp were kept. The fraction of conserved regions was calculated relative to the regions with transcription.

Odds ratio with GTEx eQTLs

Bidirectional transcription regions were overlapped with disease associated variants from GTEx (version 8) and odds ratio calculated [36]. The odds ratio calculation was performed using the library epitools in R [78].

Counting reads

Reads were counted using `featureCounts` from RSubread [79]. Gene reads were counted over gene bodies (elongation region), excluding the initiation region. Bidirectional regions were counted on both strands over the region defined by muMerge [18]. Multimapping reads were ignored in all cases. Multi-feature overlap was allowed for counting across genes but not bidirectional regions. More details are provided in the Supplementary Methods.

Normalization of counts was done using transcripts per million (TPM) normalization [80, 81]. To avoid double counting reads, for bidirectional transcripts that overlap genes, only 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 the gene were used for intragenic bidirectional transcripts.

Calculating summary statistics

The summary statistics were calculated using R. For all samples, the average and median transcription values were calculated based on the normalized counts. Across-sample coefficient of variation (CV) for human samples were calculated for all transcripts.

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. Normalized counts were log transformed, euclidean distances calculated using the R package distances, and PCA performed with the `prcomp` function in the stats package in R [82].

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 [45]. The Expression specificity scores (ESSs) [47] were calculated from median TPM values per gene per tissue.

Correlation and co-transcription analysis summary

Building the co-transcription bidirectional and gene pairs from nuclear run-on data was split into three steps: (1) defining bidirectional and gene regions, (2) finding pairs of highly correlated genes and bidirectional transcripts, and (3) filtering high-confidence pairs (See Fig. 4A and Supplementary Methods for more details).

Evaluation of relative false positive rate

Here we reasoned that genes and bidirectionals correlated across distinct chromosomes would be primarily false positives, e.g. unlikely to be real pairs. Therefore, we assessed the relative false positive rate using a randomization strategy. Specifically, when considering a gene on a given chromosome, we keep the distance to same chromosome bidirectionals unaltered but sample a transcription levels vector from the complete set of bidirectionals not on the same chromosome. For each chromosome, the shuffling of bidirectional regions from the remaining chromosomes was done without replacement. See Supplementary Methods for complete details.

Overlap of pairs with eQTLs and crisprQTLs

Correlated pairs of gene and bidirectional were overlapped with pairs from crisprQTLs validated enhancer – gene pairs from Gasperini [56] and eQTLs from GTEx [36]. The gene and bidirectional pairs were randomly shuffled 1000 times within each chromosome, and the overlaps with the crisprQTLs/eQTLs assessed.

RNA PolII ChIA-PET supported pairs

Overlaps between the RNA PolII ChIA-PET loops and DBNascent pairs were performed using `bedtools pairtopair`. DBNascent pairs were defined as bidirectional regions and gene promoters (± 1000bp around the TSS). The PolII ChIA-PET data was collected from the 4D Nucleome program and ENCODE project [83–86]. See Supplementary Methods for a complete list of accession numbers utilized.

Evaluation of bidirectional region and gene pairs

The evaluation of bidirectional transcript and gene pairs was performed based on the recovery of GTEx eQTLs and ChIA-PET loops. Evaluation was performed on four

pair assignment methods: pairs assigned using (1) correlations of transcripts (DBNascent Pairs), (2) assignment of the closest bidirectional region to a gene (Closest Pairs), (3) correlations of bidirectional transcripts and genes from other chromosomes (Shuffled Chromosome Pairs), and (4) randomized gene and bidirectional pairs (Random DBNascent Pairs) (Supplementary Fig. S45). Additionally, the recovery of eQTLs from GTEx and PolII ChIA-PET loops was compared for matching tissue overlaps to unmatched tissues (Supplementary Figs. S46 and S47). The medians between the recovery rates were compared using a Wilcoxon rank sum test in R.

Nascent database structure : back-end

The MySQL database backend for DBNascent was built using Python and SQLAlchemy. A front-end website for DBNascent [87] was built in Python using the packages Django and uWSGI and is served using nginx. The website is maintained by the IT group at the BioFrontiers Institute.

Abbreviations

GRO-seq	Global run-on sequencing
PRO-seq	Precision run-on sequencing
TRE	Transcribed regulatory region
uaRNA	Upstream antisense RNA
PROMPTs	Promoter upstream transcript
QC	Quality control
eRNAs	Enhancer-associated RNAs
NRO	Nuclear run-on
UTR	Untranslated region
TSS	Transcription start site
TES	Transcription end site
CDS	Coding DNA Sequence
cCRE	Candidate cis-regulatory element
CAGE	Cap analysis of gene expression
eQTL	Expression quantitative trait loci
ESC	Embryonic stem cells
HPC	Hematopoietic progenitor cell
TPM	Transcripts per million
ESS	Expression specificity score
FDR	False discovery rate
GEO	Gene Expression Omnibus
SRA	Sequence Read Archive
CV	Coefficient of variation
PCA	Principle component analysis

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11568-z.

Additional file 1. Contains Supplementary Methods, Figures and Tables.

Additional file 2. The movie showing transcription level and variation of transcriptional regions by level of tissue specificity.

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.

Authors' 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.

Funding

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

Data availability

Processed data is available on the DBNascent website (nascent.colorado.edu) and intermediate analysis files can be found on Zenodo (accession number 10.5281/zenodo.14519113) [87, 88].

Code availability

All the code and methods used in the meta-analysis of nascent RNA sequencing experiments can be found on GitHub (https://github.com/Dowell-Lab/DBNascent_Analysis) [89]. All scripts for database construction and maintenance, as well as a visual schema of the database, can be found on GitHub (https://github.com/Dowell-Lab/DBNascent-build) [90].

- Database backend: https://github.com/Dowell-Lab/DBNascent-build [90]
- Data preprocessing: https://github.com/Dowell-Lab/Nascent-Flow [91]
- Bidirectional calling and read counting: https://github.com/Dowell-Lab/Bidirectional-Flow [92]
- muMerge and combining bidirectional regions: https://github.com/Dowell-Lab/bidirectionals_merged [73]
- Correlation of bidirectional regions and genes: https://github.com/Dowell-Lab/bidir_gene_pairs [93]
- Downstream analyses and notebooks: https://github.com/Dowell-Lab/DBNascent_Analysis [89]

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 23 December 2024 Accepted: 3 April 2025
Published online: 25 April 2025

References

1. Wissink EM, Vihervaara A, Tippens ND, Lis JT. Nascent RNA analyses: tracking transcription and its regulation. Nat Rev Genet. 2019;20:705–23.
2. Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science. 2008;322:1845–8.
3. Kwak H, Fuda NJ, Core LJ, Lis JT. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science (New York, N.Y.). 2013;339:950–3.
4. Shiraki T, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci. 2003;100:15776–81.
5. Forrest ARR, et al. A promoter-level mammalian expression atlas. Nature. 2014;507:462–70.
6. Danko CG, et al. Identification of active transcriptional regulatory elements from GRO-seq data. Nat Meth. 2015;12:433–8.
7. Preker P, et al. RNA exosome depletion reveals transcription upstream of active human promoters. Science. 2008;322:1851–4.

8. Andersson R, Refsing Andersen P, Valen E, Core LJ, Bornholdt J, Boyd M, Heick Jensen T, Sandelin A. Nuclear stability and transcriptional directionality separate functionally distinct RNA species. Nat Commun. 2014;5(1):5336.
9. Kim T-k, et al. Widespread transcription at neuronal activity-regulated enhancers. Nature. 2010;465:182–7.
10. De Santa F, et al. A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol. 2010;8:e1000384.
11. Azofeifa JG, Dowell RD. A generative model for the behavior of RNA polymerase. Bioinformatics. 2016;33:227–34.
12. Danko CG, et al. Dynamic evolution of regulatory element ensembles in primate CD4+ T cells. Nat Ecol Evol. 2018;2:537–48.
13. Yao L, et al. A comparison of experimental assays and analytical methods for genome-wide identification of active enhancers. Nat Biotechnol. 2022;40:1056–65.
14. Cardiello JF, Sanchez GJ, Allen MA, Dowell RD. Lessons from eRNAs: understanding transcriptional regulation through the lens of nascent RNAs. Transcription. 2020;11:3–18.
15. Zhou X, O'Shea EK. Integrated approaches reveal determinants of genome-wide binding and function of the transcription factor Pho4. Mol Cell. 2011;42:826–36.
16. Savic D, et al. Promoter-distal RNA polymerase II binding discriminates active from inactive CCAAT/enhancer-binding protein beta binding sites. Genome Res. 2015;25:1791–800.
17. Azofeifa JG, et al. Enhancer RNA profiling predicts transcription factor activity. Genome Res. 2018;28:334–44.
18. Rubin JD, et al. Transcription factor enrichment analysis (TFEA): Quantifying the activity of hundreds of transcription factors from a single experiment. Nat Commun Biol. 2021;4:661.
19. Wang Z, Chu T, Choate LA, Danko CG. Identification of regulatory elements from nascent transcription using dREG. Genome Res. 2019;29:293–303.
20. Kaikkonen MU, et al. Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. Mol Cell. 2013;51:310–25.
21. Chu T, et al. Chromatin run-on and sequencing maps the transcriptional regulatory landscape of glioblastoma multiforme. Nat Genet. 2018;50:1553–64.
22. Jones T, Sigauke RF, Sanford L, et al. TF Profiler: a transcription factor inference method that broadly measures transcription factor activity and identifies mechanistically distinct networks. Genome Biol. 2025;26:92. https://doi.org/10.1186/s13059-025-03545-2.
23. Lidschreiber K, et al. Transcriptionally active enhancers in human cancer cells. Mol Syst Biol. 2021;17:e9873.
24. Lee SA, Kristjánsdóttir K, Kwak H. eRNA co-expression network uncovers TF dependency and convergent cooperativity. Sci Rep. 2023;13:19085.
25. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30:207–10.
26. Barrett T, et al. NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res. 2012;41:D991–5.
27. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database Collaboration. The sequence read archive. Nucleic Acids Res. 2010;39:D19–21.
28. ENCODE Project Consortium, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57.
29. Davis CA, et al. The encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. 2018;46:D794–801.
30. Luo Y, et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Res. 2020;48:D882–9.
31. Hitz BC, et al. The ENCODE uniform analysis pipelines. bioRxiv 2023.04.04.535623. https://doi.org/10.1101/2023.04.04.535623.
32. Andersson R, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014;507:455–61.
33. Lizio M, et al. Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol. 2015;16:1–14.
34. Abugessaisa I, et al. FANTOM enters 20th year: expansion of transcriptomic atlases and functional annotation of non-coding RNAs. Nucleic Acids Res. 2021;49:D892–8.
35. Gao T, Qian J. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. Nucleic Acids Res. 2020;48:D58–64.
36. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369:1318–30.
37. Zoonomia Consortium. A comparative genomics multitool for scientific discovery and conservation. Nature. 2020;587:240–5.
38. Kuderna LF, et al. Identification of constrained sequence elements across 239 primate genomes. Nature. 2024;625:735–42.
39. Armstrong J, et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature. 2020;587:246–51.
40. Liu X, et al. Dynamic change of transcription pausing through modulating NELF protein stability regulates granulocytic differentiation. Blood Adv. 2017;1:1358–67.
41. Aeby E, et al. Decapping enzyme 1A breaks X-chromosome symmetry by controlling Tsix elongation and RNA turnover. Nat Cell Biol. 2020;22:1116–29.
42. Smith JP, Dutta AB, Sathyan KM, Guertin MJ, Sheffield NC. PEPPRO: quality control and processing of nascent RNA profiling data. Genome Biol. 2021;22:1–17.
43. Villar D, et al. Enhancer evolution across 20 mammalian species. Cell. 2015;160:554–66.
44. Liu Y, et al. Transcriptional landscape of the human cell cycle. Proc Natl Acad Sci. 2017;114:3473–8.
45. 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. 2020;21:58.
46. Cabili MN, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 2011;25:1915–27.
47. Torres JM, et al. A multi-omic integrative scheme characterizes tissues of action at loci associated with type 2 diabetes. Am J Hum Genet. 2020;107:1011–28.
48. Hariprakash JM, Ferrari F. Computational biology solutions to identify enhancers-target gene pairs. Comput Struct Biotechnol J. 2019;17:821–31.
49. Xu H, Zhang S, Yi X, Plewczynski D, Li MJ. Exploring 3D chromatin contacts in gene regulation: the evolution of approaches for the identification of functional enhancer-promoter interaction. Comput Struct Biotechnol J. 2020;18:558–70.
50. Wang J, et al. Nascent RNA sequencing analysis provides insights into enhancer-mediated gene regulation. BMC Genomics. 2018;19:1–18.
51. Azofeifa JG, Allen MA, Lladser ME, Dowell RD. An annotation agnostic algorithm for detecting nascent RNA transcripts in GRO-seq. IEEE/ACM Trans Comput Biol Bioinforma. 2017;14:1070–81.
52. Qin T, et al. Comprehensive enhancer-target gene assignments improve gene set level interpretation of genome-wide regulatory data. Genome Biol. 2022;23:105.
53. Fulco CP, et al. Activity-by-contact model of enhancer-promoter regulation from thousands of crispr perturbations. Nat Genet. 2019;51:1664–9.
54. Nasser J, et al. Genome-wide enhancer maps link risk variants to disease genes. Nature. 2021;593:238–43.
55. Sheth MU, et al. Mapping enhancer-gene regulatory interactions from single-cell data. bioRxiv 2024.11.23.624931. https://doi.org/10.1101/2024.11.23.624931.
56. Gasperini M, et al. A genome-wide framework for mapping gene regulation via cellular genetic screens. Cell. 2019;176:377–90.
57. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature. 2012;489:109–13.
58. De Laat W, Duboule D. Topology of mammalian developmental enhancers and their regulatory landscapes. Nature. 2013;502:499–506.
59. Mills C, et al. PEREGRINE: a genome-wide prediction of enhancer to gene relationships supported by experimental evidence. PLoS ONE. 2020;15:e0243791.
60. Schmidt F, et al. Integrative analysis of epigenetics data identifies gene-specific regulatory elements. Nucleic Acids Res. 2021;49:10397–418.
61. Moody J, et al. A single-cell atlas of transcribed cis-regulatory elements in the human genome. bioRxiv 2023.11.13.566791. https://doi.org/10.1101/2023.11.13.566791.

62. Hafner A, Boettiger A. The spatial organization of transcriptional control. Nat Rev Genet. 2023;24:53–68.
63. Andrysik Z, et al. Identification of a core TP53 transcriptional program with highly distributed tumor suppressive activity. Genome Res. 2017;27:1645–57.
64. Maas Z, Sigauke R, Dowell R. Deconvolution of nascent sequencing data using transcriptional regulatory elements. Pac Symp Biocomput. 2023;2024:564–78.
65. Hunter S, Sigauke RF, Stanley JT, Allen MA, Dowell RD. Protocol variations in run-on transcription dataset preparation produce detectable signatures in sequencing libraries. BMC Genomics. 2022;23:1–18.
66. Maas ZL, Dowell RD. Internal and external normalization of nascent RNA sequencing run-on experiments. BMC Bioinformatics. 2024;25:19.
67. Mattioli K, et al. High-throughput functional analysis of lncRNA core promoters elucidates rules governing tissue specificity. Genome Res. 2019;29:344–55.
68. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:1–13.
69. Maurano MT, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337:1190–5.
70. Farh KK-H, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature. 2015;518:337–43.
71. Sasse SK, et al. Enhancer RNA transcription pinpoints functional genetic variants linked to asthma. Nat Commun. 2025;16:2750.
72. Stanley J. mumerge. 2022. https://pypi.org/project/mumerge/. v1.1.0, 2022-04-27
73. Sigauke R, Townsend H. Bidirectional mumerge pipeline. 2023. https://github.com/Dowell-Lab/bidirectionals_merged. v1.0.0
74. Dowle M, et al. Package 'data.table'. Extension of 'data. frame'. 2019;596. v1.16.2, 2024-10-9
75. Quinlan AR, Hall IM. Bedtools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
76. Fisher RA. Statistical methods for research workers. Edinburgh: Oliver and Boyd; 1934.
77. De S, Pedersen BS, Kechris K. The dilemma of choosing the ideal permutation strategy while estimating statistical significance of genome-wide enrichment. Brief Bioinform. 2014;15:919–28.
78. Aragon TJ, Fay MP, Wollschlaeger D, Omidpanah A, Omidpanah MA. Package 'epitools'. v0.5-10.1, 2017-10-26.
79. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923–30.
80. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:1–16.
81. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci. 2012;131:281–5.
82. Jolliffe IT. Principal component analysis. Technometrics. 2003;45:276.
83. Dekker J, et al. The 4D nucleome project. Nature. 2017;549:219–26.
84. Reiff SB, et al. The 4D Nucleome Data Portal as a resource for searching and visualizing curated nucleomics data. Nat Commun. 2022;13:2365.
85. Affymetrix ENCODE Transcriptome Project, Cold Spring Harbor Laboratory ENCODE Transcriptome Project. Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. Nature. 2009;457:1028–32.
86. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
87. Sanford L. DBNascent website. 2025. https://nascent.colorado.edu/.
88. Sigauke R, et al. DBNascent data repository data. 2025. https://doi.org/10.5281/zenodo.14519113.
89. Sigauke R, Sanford L, Townsend H. DBNascent Analysis. 2025. https://github.com/Dowell-Lab/DBNascent_Analysis. v1.0.0
90. Sanford L, Maas Z. DBNascent build. 2025. https://github.com/Dowell-Lab/DBNascent-build. v1.0.0
91. Sanford L, Tripodi I, Gruca M, Allen M, Dowell R. Nascent-flow. 2025. https://doi.org/10.5281/zenodo.14953945.
92. Sigauke R, Sanford L, Allen M, Dowell R. Bidirectional-flow. 2025. https://doi.org/10.5281/zenodo.14953943.
93. Sigauke R. DBNascent bidirectional gene pair correlations. 2023. https://github.com/Dowell-Lab/bidir_gene_pairs. v1.0.0

Publisher's Note