2019-06-07

BarkBase: Epigenomic Annotation of Canine Genomes

Kate Megquier
Broad Institute of MIT and Harvard

Et al.

genes

MDPI

Article

BarkBase: Epigenomic Annotation of Canine Genomes

Kate Megquier [1], Diane P. Genereux [1], Jessica Hekman [1], Ross Swofford [1], Jason Turner-Maier [1], Jeremy Johnson [1], Jacob Alonso [1], Xue Li [1,2], Kathleen Morrill [1,2], Lynne J. Anguish [3], Michele Koltookian [1], Brittney Logan [2], Claire R. Sharp [4], Lluis Ferrer [5], Kerstin Lindblad-Toh [1,6], Vicki N. Meyers-Wallen [7], Andrew Hoffman [8,9] and Elinor K. Karlsson [1,2,10,*]

[1] Vertebrate Genomics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA; kmegq@broadinstitute.org (K.M.); genereux@broadinstitute.org (D.P.G.); jphekman@broadinstitute.org (J.H.); swofford@broadinstitute.org (R.S.); jturner@broadinstitute.org (J.T.-M.); jjohnson@broadinstitute.org (J.J.); jalonso@broadinstitute.org (J.A.); xue.li2@umassmed.edu (X.L.); Kathleen.Morrill@umassmed.edu (K.M.); perloski@broadinstitute.org (M.K.); kersli@broadinstitute.org (K.L.-T.)

[2] Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA 01655, USA; brittney.logan@umassmed.edu

[3] Baker Institute for Animal Health, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853, USA; lja2@cornell.edu

[4] School of Veterinary and Life Sciences, College of Veterinary Medicine, Murdoch University, Perth, Murdoch, WA 6150, Australia; C.Sharp@murdoch.edu.au

[5] Departament de Medicina i Cirurgia Animals Veterinary School, Universitat Autonoma de Barcelona, 08193 Barcelona, Spain; Lluis.Ferrer@tufts.edu

[6] Science for Life Laboratory, Department of Medical Biochemistry & Microbiology, Uppsala University, 751 23 Uppsala, Sweden

[7] Baker Institute for Animal Health and Department of Biomedical Sciences, College of Veterinary Medicine, Cornell University, Ithaca, NY 14850, USA; meyerswallen@gmail.com

[8] School of Veterinary Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA; hoffm018@upenn.edu

[9] Cummings School of Veterinary Medicine, Tufts University, Grafton, MA 01536, USA

[10] Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, MA 01655, USA

* Correspondence: elinor.karlsson@umassmed.edu

Received: 4 May 2019; Accepted: 3 June 2019; Published: 7 June 2019

Abstract: Dogs are an unparalleled natural model for investigating the genetics of health and disease, particularly for complex diseases like cancer. Comprehensive genomic annotation of regulatory elements active in healthy canine tissues is crucial both for identifying candidate causal variants and for designing functional studies needed to translate genetic associations into disease insight. Currently, canine geneticists rely primarily on annotations of the human or mouse genome that have been remapped to dog, an approach that misses dog-specific features. Here, we describe BarkBase, a canine epigenomic resource available at barkbase.org. BarkBase hosts data for 27 adult tissue types, with biological replicates, and for one sample of up to five tissues sampled at each of four carefully staged embryonic time points. RNA sequencing is complemented with whole genome sequencing and with assay for transposase-accessible chromatin using sequencing (ATAC-seq), which identifies open chromatin regions. By including replicates, we can more confidently discern tissue-specific transcripts and assess differential gene expression between tissues and timepoints. By offering data in easy-to-use file formats, through a visual browser modeled on similar genomic resources for human, BarkBase introduces a powerful new resource to support comparative studies in dogs and humans.

Keywords: dog; expression; genome; annotation; ATAC-seq; RNA-seq; epigenomic; canine; comparative

1. Introduction

The domestic dog is an increasingly important model for a wide variety of human diseases, including cancer, immune-mediated disorders, and psychiatric diseases, as well as for healthy aging [1–7]. The pet dog population is unusual in that it is comprised in a large part of ancestry from dog breeds, which are genetically isolated populations that are just a few hundred years old and have limited genetic diversity [8,9]. Purebred dogs often suffer from a high risk of diseases that closely model complex human diseases, and the population structure of breeds can make genetic mapping approaches far more powerful [8,10]. With a very large population size (70 million in the United States alone), and a shared environment with human owners, dogs are, in many ways, an ideal model. However, as canine whole genome datasets expand [11,12], comprehensive epigenomic profiling in dogs lags behind, impeding the translation of genetic associations into functional understanding.

In human genomics, available large scale epigenomic resources such as ENCODE [13], GTEx [14], and the National Institute of Health (NIH) Roadmap [15] catalog the functional elements active in diverse, healthy tissues. Such resources have proven exceptionally powerful for investigating the non-coding regulatory variants that make up the vast majority of risk factors identified in complex-trait genomewide association studies (GWAS), for finding which candidate variants are most likely to perturb cell function, and for distinguishing which cell and tissue types are most likely to be involved in the disease process [16]. By integrating multiple different types of epigenomic data for each cell type or tissue, active, noncoding regulatory elements map far more specifically [17,18]. Developing similar resources for a natural model organism like a dog would support comparative studies of the genome regulatory function and the conservation of disease mechanisms, further elucidating human disease biology [19].

Currently, there is no canine equivalent to these human epigenomic resources. Canine genes and non-coding RNAs are mapped in an Ensembl gene annotation from July 2012 and a University of California Santa Cruz (UCSC) Genome Browser track hub from 2014, both based on the same canine RNA-seq dataset [20–24]. This dataset includes a range of diverse adult canine tissues (blood, brain, heart, kidney, liver, lung, ovary, skeletal muscle, skin, and testis). The Ensembl annotation includes 19,856 coding genes and 11,898 non-coding genes, and the UCSC track has 22,798 protein coding genes (20,657 of high confidence) and 7224 putative long noncoding RNAs [25,26].

The RNA-seq dataset used to build these existing annotations has several significant limitations. First, it includes only adult tissues, potentially missing critical genes or isoforms active only during development. Second, it contains just one sample for each tissue and so provides no biological replicates to validate the data, to examine how gene expression varies among individuals, or to determine which genes are consistently up or down regulated in individual tissue types. Finally, RNA-seq alone does not capture untranscribed regulatory elements.

For untranscribed regulatory elements, the canine research community relies on genome annotations mapped onto the dog genome from other species, including sequence conservation scores [27,28]. Yet, while most protein coding genes are expected to be conserved across mammals, certain types of regulatory elements, such as enhancers, turn over rapidly [29,30]. The lack of epigenomic data for dog tissues is particularly concerning as dogs are a natural model for complex, polygenic diseases associated primarily with regulatory, rather than coding, variation [7,11,16]. Even with the limited epigenomic resources available today, canine studies have yielded new insight into human diseases, including cancers [4,31–37], behavioral traits and disorders [2,38,39], autoimmune diseases [3,40], and others [41–46].

Here, we introduce BarkBase (http://www.BarkBase.org), a public resource that builds on the paradigm established by ENCODE and NIH Roadmap. At present, BarkBase contains RNA sequencing data from 27 diverse tissues collected from five adult dogs, paired with 30× whole genome sequence data. BarkBase also includes RNA sequencing data for embryonic tissues from five carefully staged gestational time points. We are currently generating and uploading assay for transposase-accessible chromatin using sequencing (ATAC-seq) data for each of the adult tissue samples, and anticipate that this will be completed by the end of 2019.

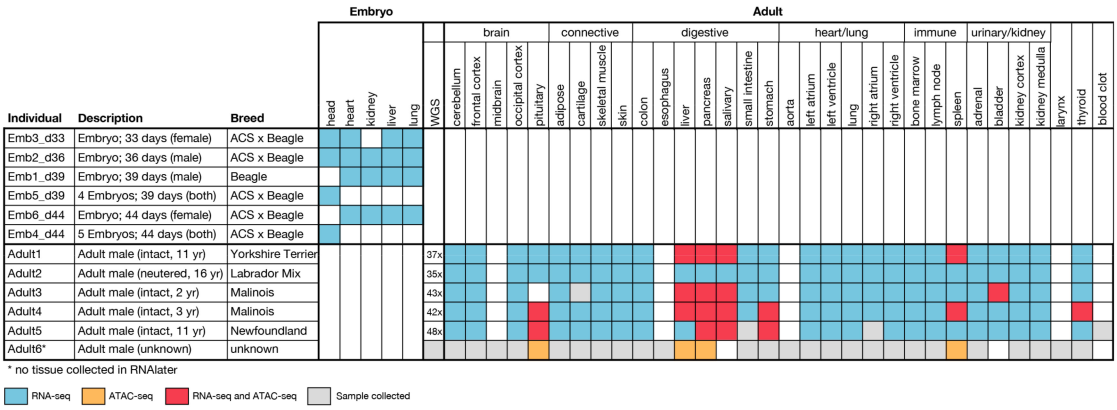
Our initial analyses of BarkBase data expand the catalog of dog genes, including noncoding genes, and identify novel regulatory elements from ATAC-seq peaks, confirming the value of BarkBase as a tool for canine genomics, and enhancing the value of the dog as a powerful natural model for the study of human disease.

2. Materials and Methods

2.1. Sample Collection

2.1.1. Adult Tissue

All research in this study was conducted according to NIH guidelines for the Care of Vertebrate Animals used in Testing, Research, and Training. Six dogs were enrolled through the Deceased Dog Donation program at Cummings School of Veterinary Medicine at Tufts University, with approval from the Institutional Animal Care and Use Committee (IACUC). All dogs were euthanized for medical reasons, and donated by owners after death. Inclusion was based primarily on availability, with the only exclusion criterion being a cancer diagnosis. By chance rather than study design, all dogs were male. They spanned a range of sizes, ages, and breeds, and included one mixed breed dog with a high proportion of Labrador Retriever ancestry (Figure 1, Table S1). We collected biopsies from up to 31 tissues (Figure 2), including heart (left atrium, left ventricle, right atrium, and right ventricle), brain (cerebellum, frontal cortex, occipital cortex, and pituitary), kidney (cortex and medulla), adipose tissue, adrenal gland, bladder, bone marrow, cartilage, colon, liver, lung, small intestine, lymph node, pancreas, salivary gland, skeletal muscle, skin, spleen, stomach, and thyroid gland. Sample collection was completed within two hours of death. Each sample was divided in two, with one half stored in RNAlater and the other flash frozen in liquid nitrogen, then stored at −80 °C until processed.



			Embryo						Adult																															
									brain					connective				digestive							heart/lung						immune			urinary/kidney						
Individual	Description	Breed	head	heart	kidney	liver	lung	WGS	cerebellum	frontal cortex	midbrain	occipital cortex	pituitary	adipose	cartilage	skeletal muscle	skin	colon	esophagus	liver	pancreas	salivary	small intestine	stomach	aorta	left atrium	left ventricle	lung	right atrium	right ventricle	bone marrow	lymph node	spleen	adrenal	bladder	kidney cortex	kidney medulla	larynx	thyroid	blood clot
Emb3_d33	Embryo; 33 days (female)	ACS x Beagle	🟦	🟦		🟦	🟦																																		
Emb2_d36	Embryo; 36 days (male)	ACS x Beagle	🟦	🟦	🟦	🟦	🟦																																		
Emb1_d39	Embryo; 39 days (male)	Beagle		🟦	🟦	🟦	🟦																																		
Emb5_d39	4 Embryos; 39 days (both)	ACS x Beagle	🟦																																						
Emb6_d44	Embryo; 44 days (female)	ACS x Beagle		🟦	🟦	🟦	🟦																																		
Emb4_d44	5 Embryos; 44 days (both)	ACS x Beagle	🟦																																						
Adult1	Adult male (intact, 11 yr)	Yorkshire Terrier						37x	🟦	🟦		🟦	🟦	🟦	🟦	🟦	🟦	🟦		🟥	🟥	🟥	🟦	🟦		🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥	🟦	🟦	🟦	🟦		🟦	
Adult2	Adult male (neutered, 16 yr)	Labrador Mix						35x	🟦	🟦		🟦	🟦	🟦	🟦	🟦	🟦	🟦		🟦	🟦	🟦	🟦	🟦		🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦		🟦	
Adult3	Adult male (intact, 2 yr)	Malinois						43x	🟦	🟦		🟦		🟦	▨	🟦	🟦	🟦		🟥	🟥	🟥	🟦	🟦		🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥	🟦	🟦		🟦	
Adult4	Adult male (intact, 3 yr)	Malinois						42x	🟦	🟦		🟦	🟥	🟦	🟦	🟦	🟦	🟦		🟥	🟥	🟥	🟦	🟥		🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥	🟦	🟦	🟦	🟦		🟥	
Adult5	Adult male (intact, 11 yr)	Newfoundland						48x	🟦	🟦		🟦	🟥	🟦	🟦	🟦	🟦	🟦		🟦	🟥	🟥	▨	🟥		🟦	🟦	🟦	▨	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦		🟦	▨
Adult6*	Adult male (unknown)	unknown						▨	▨	▨	▨	▨	🟧	▨	▨	▨	▨	▨	▨	🟧	🟧		▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	🟧	▨		▨	▨	▨	▨	

* no tissue collected in RNAlater

🟦 RNA-seq 🟧 ATAC-seq 🟥 RNA-seq and ATAC-seq ▨ Sample collected

Figure 1. BarkBase sample collection and data production. Samples were collected from a total of six embryos, and six adult dogs. BarkBase currently contains RNA-seq data from up to five tissues in d33, d36, d39, and d44 embryos, and from up to 27 tissues sampled from each of five adult dogs diverse in age and in breed ancestry. ATAC-seq data are currently available for eight tissues from a subset of individuals. Additional data sets will be posted as they become available.

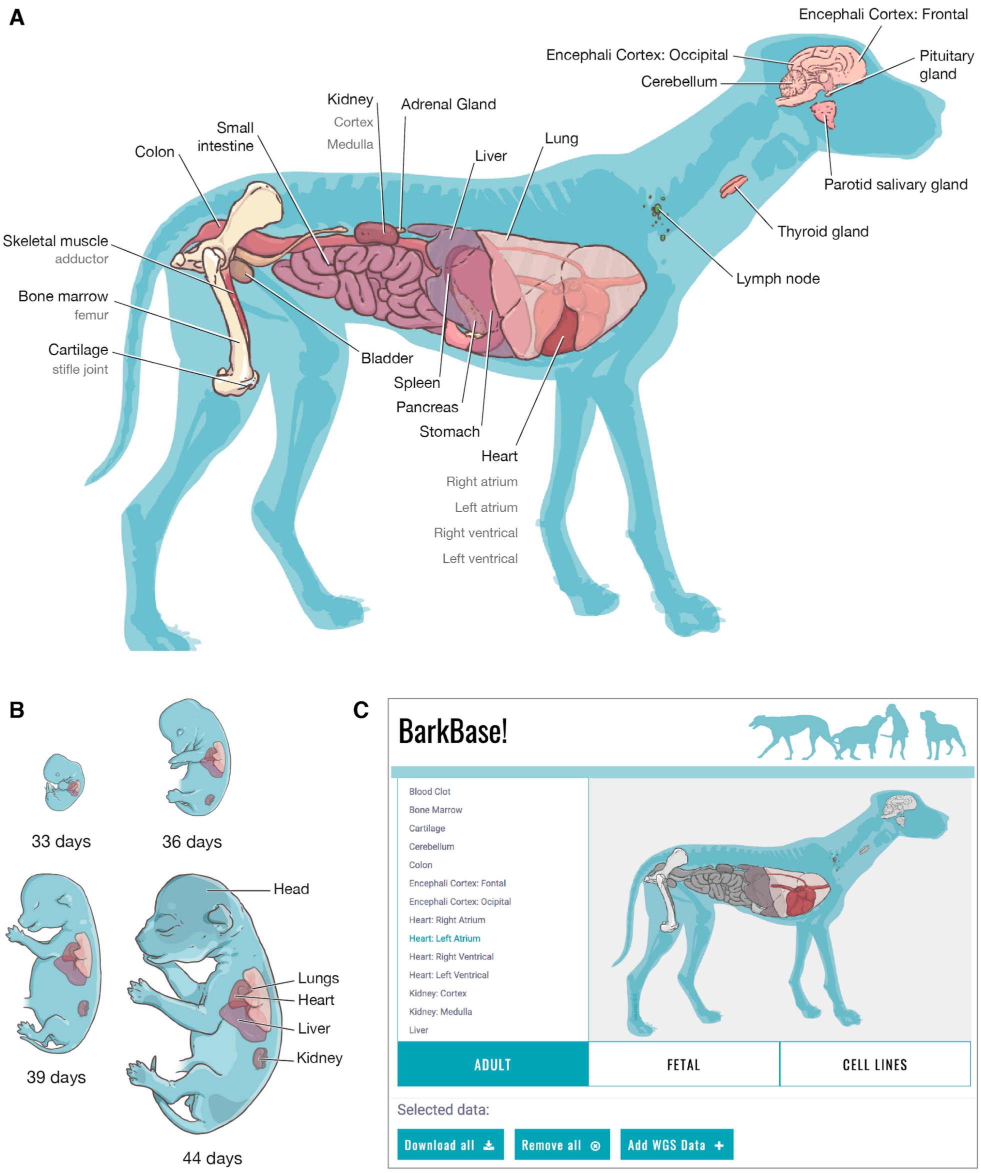


Figure 2. The BarkBase web portal. The BarkBase web portal enables download of whole genome sequence (WGS) data, RNA-seq data, and assay for transposase-accessible chromatin using sequencing (ATAC-seq) data for (**A**) up to 27 tissues from each of the five adult dogs; and (**B**) up to five tissues from canine embryos collected at each of the four staged gestational timepoints. Reads preprocessed and aligned to CanFam3.1 are available at BarkBase.org. From the BarkBase interface (**C**), users can readily select specific tissues and samples. Raw read data from RNA-seq and ATAC-seq is available through the Sequence Read Archive (SRA) (Table S1).

2.1.2. Embryonic Tissue

All animal care and experimental protocols related to embryonic tissue collection were approved by the Institutional Animal Care and Use Committee at Cornell University (1989–0068). Embryos were collected using hysterotomy, separated through microdissection and individually stored in RNAse-free phosphate buffered saline (PBS, 4C), as described in Meyers-Wallen et al [47]. Each embryo was developmentally staged based on photographs of external morphology [48]. Sex was ascertained by PCR targeting the *SRY* locus. For two samples (heads), tissues from several individuals of the same litter were stored and processed together to ensure sufficient material (Figure 1).

2.2. Whole Genome Sequencing

2.2.1. Sequencing and Variant Calling

Genomic DNA was prepared utilizing custom indices from Integrated DNA Techologies (Coralville, Iowa, USA) designed by the Broad Institute Genomics Platform (Cambridge, MA, USA) and a HyperPrep library construction kit from Kapa Biosciences (Wilmington, MA, USA). Sequenced was done on a NovaSeq 6000 (Illumina, San Diego, CA, USA), using two color chemistry at 2 × 150 bp reads to 30× coverage, de-multiplexed, aggregated, and aligned to the CanFam3.1 reference via BWA-MEM [49]. Read data was processed according to GATK best practices, specifically base quality score recalibration, variant identification using HaplotypeCaller in GVCF mode, and joint calling, using the GATK 6/24/2016 nightly build against a background of 872 additional dogs and other canids (Table S2) [50,51]. Variant filtration was accomplished by hard filtering. For SNPs, the filtration parameters were: QD < 2.0 or FS > 60.0 or MQ < 40.0 or MQRankSum < -12.5 or ReadPosRankSum < -8.0. For insertions and deletions, the filtration parameters were: QD < 2.0 or FS > 200.0 or InbreedingCoeff < -0.8 or SOR > 10.0 or ReadPosRankSum < -20.0.

2.2.2. Relatedness

Relatedness of adult dogs was measured through pairwise identity by descent (IBD) estimation in Plink 2.0 [52] using the "genome" flag on 1,632,289 autosomal variants remaining after pruning variants in linkage disequilibrium. Pruning was done with the "indep" flag (window size of 50 variants, a window shift of five variants and variance inflation factor threshold of 2) [52,53].

2.2.3. Annotation

Variants were annotated with 18 functional categories using SnpEff [54] and CanFam3 (v3.1.86) from Ensembl [22] as the reference database.

2.3. RNA Sequencing, Normalization, and Analysis

2.3.1. Sequencing

RNA samples with an RNA quality score (RQS) >5.5 were poly-A selected, strand-specific cDNA synthesized, and an Illumina library constructed via the Illumina TruSeq protocol. Libraries were size selected (450–550 bp inserts) and sequenced on an Illumina HS2500 (2 × 101 bp reads) to >50 million reads/library. We achieved a median of 64 million reads, substantially exceeding the ENCODE standard of 30 million reads per sample [55].

2.3.2. Read Preprocessing

RNA-seq reads were analyzed by the workflow outlined by Pertea et al. [56]. Briefly, reads were aligned to the CanFam3.1 dog genome using HISAT2 [56,57] in dta mode. Transcripts were assembled from the aligned reads using StringTie [56,58], with a gene annotation from Ensembl as a reference (version 95). Transcripts from all samples were merged using StringTie to create a new reference file, which was then used to estimate transcript abundance in each sample. Raw read counts at both the gene and transcript level were generated from StringTie transcript abundances using the "prepDE.py" Python script provided in the StringTie online manual (https://ccb.jhu.edu/software/stringtie/dl/prepDE.py), using the −l flag to set read length to 101 base pairs. The raw read counts at the gene and transcript level are available at https://data.broadinstitute.org/barkbase/.

2.3.3. TMM Normalization to Calculate Counts Per Million (CPM)

The trimmed mean of M (TMM) normalization method [59] as implemented in the R package edgeR [60,61] was used to normalize raw RNA-seq read counts. TMM normalization helps avert underestimation of the abundance of lowly or moderately expressed genes in samples with very high expression from a subset of genes, thereby avoiding inflation of the number of genes inferred to be differentially expressed between samples. A TMM normalization factor was calculated for each sample, then applied to calculate counts per million (CPM) from raw read count. For most analyses, transcripts were kept if they were expressed at >0.16 counts per million (CPM) in two or more samples, which is equivalent to requiring approximately 10 reads in our median library size of 64 million [62]. For the comparison with human RNA-seq data, a more stringent cutoff of 1 CPM in two or more samples was implemented. Filtered, normalized CPM counts at both cutoffs are available at: https://data.broadinstitute.org/barkbase/.

2.3.4. Cumulative Abundance and Tissue-Specific Reads

CPM values were used to calculate overall fraction of the transcriptome contributed by each gene. Genes were sorted by CPM, the cumulative sum calculated, and the fractional contributions of the 1000 top-contributing genes plotted for each sample.

2.3.5. Hierarchical Clustering of Samples and Adult Dogs

Euclidean distances among CPM values for the samples from adult dogs and from embryos were calculated using the dist function in R. Overall distances among the adult dogs were calculated by concatenating CPM values for the 23 tissues for which data were available from all five individuals, and distances calculated as for the single-tissue samples. Hierarchical clustering was performed using the hclust function in R [63].

2.3.6. Differential Gene Expression (DGE) and Enrichment Analysis of Embryonic Data

The edgeR package [60,61] was used to identify genes differentially expressed between embryonic time points d36 and d44 in head, liver, lung, heart, and kidney. At present, we have access to only one carefully staged embryonic sample for each time point. To accommodate the lack of biological replicates, we instead calculated dispersion values using data from 3027 genes from among the 3119 that were previously identified as "housekeeping genes" in human RNA-seq data [64] and that were also expressed in dog. Significance scores calculated in edgeR and in parallel using a different inference package in R, DESeq2 [65], were strongly correlated ($p < 2.2 \times 10^{-16}$). We defined genes with false discovery rate (FDR) < 0.1 in the edgeR-adjusted results as differentially expressed between d36 and d44. Ingenuity Pathway Analysis (IPA) from Qiagen (Hilden, Germany) was used to identify enrichment of specific categories within "Diseases or Functions Annotation”.

2.3.7. Comparison of Gene Expression between Dog and Human Tissues

Human RNA-seq data from 53 tissues of 714 human donors was downloaded from the GTEx Portal (GTEx Analysis V7, https://gtexportal.org/home/datasets). The raw human read counts from GTEx were converted to CPM using the TMM normalization method described above. Samples within a species were grouped by tissue type and the expression summarized using the median CPM for each gene. The Ensembl reference annotation was previously provided to StringTie as part of the transcript merging process, during which StringTie annotated transcripts with matching Ensembl gene names. These gene names, as well as cross-species orthology data downloaded from Ensembl [66], were used to map between BarkBase genes and human genes in GTEx. All BarkBase genes not uniquely annotated with an Ensembl gene name were discarded, as were all genes that did not uniquely map between human and dog. For each tissue, genes with median CPM ≤ 1 in both species were discarded. The Spearman correlation between the two species was then calculated using scipy [67]. The correlation matrix comparing dog and human tissues includes all tissue types for which we had data in both species.

2.3.8. Comparison of Transcript Set to the Existing Ensembl Canine Reference Annotation

The GffCompare tool [68] was used to compare the transcripts annotated by StringTie to the Ensembl canine reference annotation. Transcripts were annotated as overlapping if they completely or partially overlapped a reference transcript (GffCompare class codes "=", "c", "k", "m", and "j"), excluding cases where the overlap was on the opposite strand, completely contained within an intron of the reference transcript, and in cases where the reference was completely contained within an intron of the StringTie transcript. A reference track containing non-dog RefSeq genes aligned to the canine reference ("Other RefSeq" track) was downloaded from the UCSC Genome Browser [24,69–71] in General Transfer Format (GTF) using the table browser. Human genes in this track were annotated using Ensembl BioMart [66]. Novel transcripts were identified as those not overlapping either annotation (class code "u" in the GffCompare output). Enrichment analysis was performed using the GOseq [72] Bioconductor [73] package. Canine genes were first mapped to 1:1 human orthologs using Ensembl BioMart, and the human gene names input into GoSeq. Enrichment was tested using the set of all canine Ensembl genes that map to 1:1 human orthologs as the background. Using this approach, multiple canine snoRNAs mapping to a single human gene name were collapsed for the enrichment analysis. The Wallenius approximation was used to correct for gene length. Multiple testing correction was performed using the Benjamini–Hochberg procedure, with significance set at an FDR of 0.05. The genomic coordinates covered by BarkBase and the Ensembl and Hoeppner et al. annotations were compared using the BEDTools [74] intersect tool using the −s flag for strand specific comparisons. The GTF files annotating each dataset were first converted to BED files. The BEDTools merge tools was used to create unique non-overlapping intervals within each file. Each merged BED file was then compared to each of the others using the BEDTools intersect tool, and the number of base pairs covered by the intersecting intervals summed.

2.3.9. Comparison of Unannotated Transcripts to the RefSeq Vertebrate Mammalian Proteins

Multi-exonic transcripts not matching either the Ensembl or Hoeppner et al. annotation were aligned using blastx 2.2.30 [75] to the RefSeq vertebrate mammalian protein database, release 93 [76]. Query transcripts were considered to significantly match a RefSeq protein if blastx returned a match with a bit score greater than 60.

2.4. ATAC-seq

2.4.1. Sample Preparation and Sequencing

To extract cellular nuclei, each tissue sample was homogenized in an EZ lysis buffer using Kimble Dounce All-Glass Tissue Grinders, starting with a small (pea-sized) piece of tissue. The sample was incubated on ice for five minutes, poured through a cell strainer, and then pelleted and resuspended in 150 µL of resuspension buffer (Nuclei Isolation Kit: Nuclei EZ Prep from Sigma-Aldrich, St. Louis, MO, USA). To determine the volume needed for 50,000 cells, 10 µL of resuspension buffer was mixed with 1 µL of Trypan Blue and cells counted on a hemocytometer. The needed volume of the cell preparation was centrifuged, the resuspension buffer removed, and 50 µL of the master mix (25 µL 2x TD buffer, 5 µL Tn5, and 20 µL distilled H_2O) added. The sample was mixed by vigorous pipetting and incubated at 37°C for 30 min while rocking. The Qiagen MinElute Reaction Cleanup Kit was used to elute the pure DNA into 10 µL of elution buffer. The sample was then transferred to a 96-well plate and amplified via PCR to create the final libraries. The libraries were quantified using the BioAnalyzer 2100 from Agilent Technologies (Santa Clara, CA, USA) and the Kapa Library QuantKit, and sequencing was performed on two lanes of the HS2500 Rapid Run 2 × 25. We averaged around 80 million autosomal reads per sample, exceeding the ENCODE minimum recommendation of 50 million reads [55].

2.4.2. Processing Reads, Calling Peaks, and QC of Libraries

Reads were processed and peaks called following the ATAC-seq guidelines developed by John M. Gaspar [77]. Briefly, reads were aligned to CanFam3.1 using Bowtie2 in "very sensitive" mode. Mitochondrial reads were then removed using Samtools. PCR duplicates were removed using Picard tools, and non-uniquely mapping reads were removed using Samtools. Finally, bam files were converted to bed coverage interval files using the SAMtoBED.py script by John M. Gaspar. Quality metrics including peak counts, transcription start sites (TSS) scores, and mapping stats were derived using ATAQV [78]. As the existing gene annotation for dog is not as comprehensive as the human annotations, we could not directly apply the transcription start site enrichment score thresholds recommended by ENCODE [55]. Instead, we removed all samples falling below one standard deviation less than the mean, which left us with 37 samples with TSS scores of 1.3 or higher (mean of 2).

2.4.3. Assessing Overlap of Peaks

The BEDTools intersect tool [74] was used to compare overlaps of called peaks among and across tissues. Subsequent intervals were annotated and analyzed using the ChIPseeker R package [79], in conjunction with annotations from Ensembl [21,80]. This tool retrieves the nearest genes around an ATAC-seq peak.

2.4.4. Data Sharing

Data files are available for download at our ftp site (https://data.broadinstitute.org/barkbase) and analysis scripts at https://github.com/broadinstitute/barkbase_paper_analysis/.

3. Results

3.1. BarkBase Website

We collected samples of 33 adult tissue types, with 28 tissues sampled from at least five individuals (1,2 A). We also collected five tissues (head, heart, kidney, liver, and lung) across four embryonic time points (1,2 B). To date, we have completed RNA sequencing of 150 samples (27 adult tissues and five embryonic tissues) and have generated ATAC-seq data for 36 samples (15 adult tissues); 22 adult samples (nine tissues) currently have both data types. The embryonic dataset, which includes only RNA-seq data, is complete, with data for all tissues and timepoints except for one kidney sample, which failed to yield RNA of adequate quality for sequencing. Data generation (RNA-seq, ATAC-seq, and whole genome sequencing (WGS)) in the adult samples is ongoing.

To share this data, we developed a website inspired by the data access portal for the NIH Roadmap project [81] (Figure 2C). Like the Roadmap Visual Browser, BarkBase includes anatomical illustrations of the samples, both at each embryonic time point and in the adult dog, displaying the scope of the data set. By selecting specific tissues and samples, users can construct a customized dataset for download. BarkBase.org provides GTF formatted files for RNA-seq and ATAC-seq, and Variant Call Format (VCF) files for WGS. The total size of the data available through BarkBase.org is currently ~7 gigabytes, with more data being generated. The capacity to subset data prior to download can substantially reduce download times. The read data, aligned to CanFam3.1, is available through the Sequence Read Archive [82].

3.2. Whole Genome Sequencing

For each of the five adult dogs with nearly complete RNA-sequencing data, we have also generated high coverage whole genome sequencing of the germline DNA. For each dog, we sequenced DNA extracted from cerebellum tissue to 35–48× coverage (average coverage 41×). Across these five samples, we successfully called 56,244,173 of 57,498,383 possible variant sites (call rate = 97.82%), of which 4.74% were heterozygous and 3.1% were homozygous non-reference. Annotation of the 8,147,173 sites with non-reference allele frequency >0, using SnpEff [54], classified 53.5% (4,352,137) as intergenic, 32.25% (2,624,663) as intronic, 12.63% (1,028,340) as near genes, and 1.74% (142,033) as being either in transcribed regions or disrupting splicing.

3.3. Overlap With Existing Gene Annotations

There are two existing gene annotations for CanFam 3.1. The most widely used is an Ensembl annotation generated using a standard Ensembl mammalian genebuild pipeline. It incorporated RNA-seq data, provided by the Broad Institute, for one sample of each of ten tissues (blood, brain, heart, kidney, liver, lung, ovary, skeletal muscle, skin, and testis) [21]. RNA-seq libraries for this dataset were generated using two RNA selection techniques (poly-A and duplex-specific nuclease), adding sensitivity particularly for shorter noncoding transcripts. The Ensembl gene build contains 39,074 transcripts in 32,704 genes (60 Mb total), including 19,856 coding, 11,898 non-coding, and 950 pseudogenes. This same RNA-seq data was also used for the Hoeppner et al. improved canine genome annotation, published in 2014, which used a less conservative 194,671 transcripts in 22,172 coding genes, 7224 lincRNA candidates, and 5295 antisense transcripts, as well as 82,039 other transcripts (249 Mb total) [20].

Our new BarkBase RNA-seq data, generated from libraries made using poly-A selection only, contains 151,787 transcripts in 37,106 StringTie-assembled genes. On the single base level, there is a high degree of overlap (over 90%) with the Ensembl canine reference annotation. In addition, BarkBase contains 84 Mb of unique sequence, compared to 5 Mb in the Ensembl annotation. (Figure 3). BarkBase also has a large degree of overlap with the Hoeppner et al. data set (103 Mb), with the Hoeppner et al. data containing an additional 146 Mb of sequence.

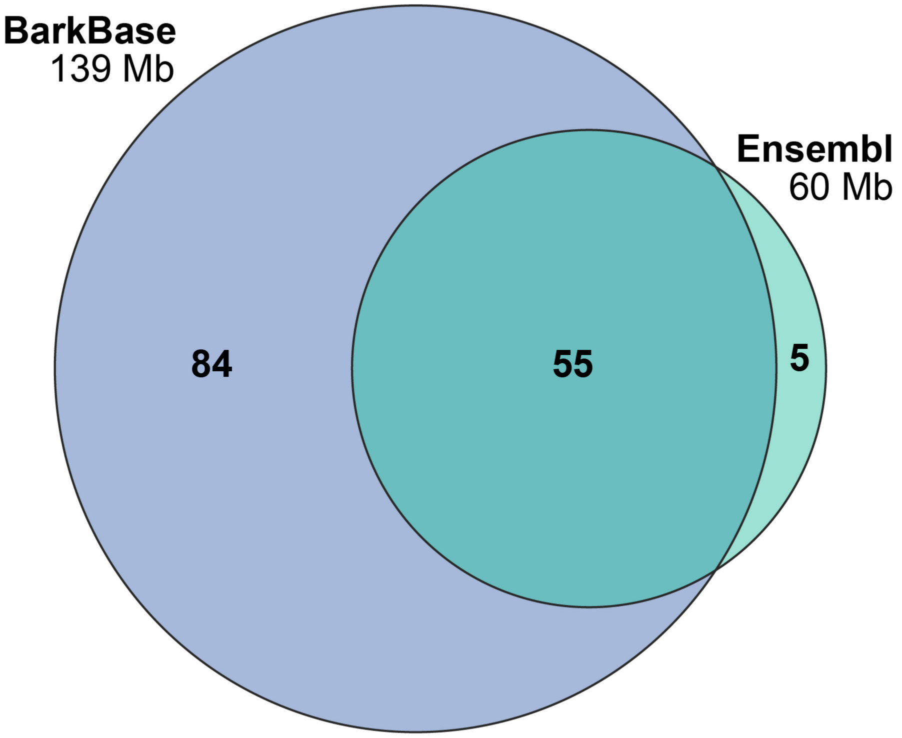


Figure 3. BarkBase captures novel transcripts. Overlapping the transcriptome from BarkBase and Ensembl shows most bases are captured in both datasets. BarkBase contains 84 Mb of transcribed sequence not included in the existing annotation, highlighting its utility to improve the annotation of the canine genome.

We examined the sensitivity of BarkBase, the Ensembl annotation, and the Hoeppner et al. data to identify exons present in the other data sets (Table S3). Sensitivity was calculated as the total number of exons in the reference annotation identified by the query dataset, divided by the total number of exons in the reference dataset [68]. BarkBase captures most of the Ensembl exons (87.6%) and about half (56.6%) of the exons in the Hoeppner et al. annotation. Reciprocally, both the Ensembl and Hoeppner et al. annotations captured about half of the exons in BarkBase (47.1% and 55.2%, respectively). This suggests that BarkBase includes exons missed in previous annotations, and will expand the catalog of annotated dog exons, transcripts, and genes.

Of the 32,704 Ensembl dog genes (19,856 coding, 11,898 non-coding, and 950 pseudogenes), 69% overlapped at least one transcript in BarkBase (Table S4). We capture the large majority of coding genes annotated in Ensembl (90%), but only 34% of the non-coding genes. The genes missing from BarkBase are largely non-coding genes without human orthologs. Of the 10,118 Ensembl genes that are not in BarkBase, 77% are noncoding, compared to only 18% of those that are in BarkBase. The genes missing from BarkBase are also less likely to have a 1:1 human ortholog (13% vs 67%).

There are 1314 Ensembl dog genes (44% coding) that have human orthologs but are missing from BarkBase (Table S5). We performed enrichment analysis of this gene set using GOseq and found 52 significantly enriched Gene Ontology (GO) terms. The top scoring term was "olfactory receptor activity" ($p = 1.8 \times 10^{-104}$, Table S6), followed by other terms relating to olfaction and sensory processes, including "detection of chemical stimulus involved in sensory perception of smell" ($p = 1.8 \times 10^{-104}$) and "G-protein coupled receptor activity" ($p = 1.4 \times 10^{-63}$). Less enriched terms include "mRNA binding involved in posttranscriptional gene silencing" ($p = 2.4 \times 10^{-12}$), "sexual reproduction" ($p = 5.0 \times 10^{-3}$) and "fertilization" ($p = 6.5 \times 10^{-3}$), and "male gamete generation" ($p = 4.6 \times 10^{-2}$). This suggests that many of the Ensembl genes not seen in BarkBase may be missed due to the lack of testis tissue in BarkBase, as testis is known to express olfactory [83] as well as taste receptors [84].

In contrast, the Ensembl genes found in BarkBase are highly enriched for GO terms related to broadly important cellular processes, including "catalytic activity", "protein binding", "intracellular", "cell", "nucleus", and "cytoplasm." All six terms had a reported p-value of 0 (Table S7). This enrichment is likely due to the gene set being tested contains 69% of all genes in the Ensembl reference, particularly coding genes. Many gene sets in the GO terms have all or nearly all members represented in this set, leading to highly significant enrichment of larger gene sets.

BarkBase captures 14,518 transcripts in 10,691 genes that are missing from the Ensembl canine annotation. Of these, 8051 are multi-exon transcripts. Approximately 80% of these were captured in Hoeppner et al., but we found 1520 novel transcripts. Of these, most (1090) are also missing from another recently published catalog of canine long non-coding RNAs [25]. Using BLAST, we tested all our genes against known protein-coding sequences, finding 769 potential novel lncRNAs (GTF file available at https://data.broadinstitute.org/barkbase/).

Of the transcripts that did not overlap the Ensembl or Hoeppner annotations, 15 transcripts overlapped five known human protein coding genes: The salivary protein encoding gene *STATH*, the eosinophil expressed pathogen response gene *EPX*, homeobox transcription factor *HOXD13*, transcriptional repressor *FEZF1*, and the calcium signaling gene *AHNAK2*. Of these human genes, four are not annotated in the canine genome, while *AHNAK2* does not map uniquely, perhaps indicating a set of paralogous sequences in the dog. *AHNAK2* is annotated at two locations, and our data adds two 5′ exons to the current gene annotation (Table S8).

3.4. Variability Among Adult Dogs and Tissues

BarkBase includes data for up to five dogs per tissue, offering opportunity to assess how patterns of expression vary between individuals and between healthy tissues. Previous datasets, were either disease-focused (primarily cancer-focused) or included just one individual per tissue [20,32,85–88]. With our current data set, we have sufficient power to detect over 80% of genes with at least a two-fold change in expression between tissues, but are still underpowered to comprehensively detect differentially expressed genes [89]. We can also assess a correlation of gene expression patterns between tissues on a whole transcriptome level.

Overall, we found that in adult dogs, gene expression is tightly matched to tissue type. We first looked at the fraction of the transcriptome contributed by the most highly expressed genes, an approach used to assess similarity in human tissue samples [90]. We saw clear similarity among dogs and diversity between tissues. In all of the tissue types, the transcriptome is dominated by a fairly small number of genes, with the top 1000 expressed genes comprising at least 25% of transcripts in all but one of the tissues types (Figure 4A). In pancreas, the pattern is even more extreme, with 75% of transcripts coming from just 40 genes, broadly consistent with findings from the human pancreas [91]. In the embryonic tissues, the pattern of strong similarity within tissues persisted, even though the samples were from different embryonic timepoints (Figure 4B). This may reflect the fairly narrow window of gestational ages sampled, or that normal embryonic development involves changes affecting just a small subset of transcribed genes. The embryonic tissues were also notably similar to their adult counterparts, suggesting the similarity among embryonic timepoints reflects their having acquired transcriptomic features of the relevant adult tissue type (Figure 4C). A single embryonic head sample is an outlier, with the top 1000 genes contributing a lower proportion of expressed transcripts, a pattern that could result if a less homogenous tissue sample were obtained during microdissection.

We next examined how similar the gene expression profiles were between samples. We found that, in both adult dogs and embryos, samples from a given tissue cluster across individuals, and are distinct from samples of other tissue types (Figure 5A). Samples from similar tissue types (for example, various samples from brain or heart) also tended to cluster closer together. This is consistent with the expectation that specific tissues have distinct transcriptomes that support their specific physiological roles and that are broadly consistent across individuals [92].

Of our five adult dogs, two had strikingly similar phenotypes. Both were purebred Malinois, of similar ages (three and four years old), and were euthanized because of behavioral problems. Genetic analysis suggests relatedness equivalent to third degree relatives (12% of genome identical by descent).The other three adult dogs were over the age of ten. One dog (Adult 1) weighed just 4 kg, ~10 fold smaller than the other four adults.

We saw no indications that the overall pattern of gene expression in tissues correlated with breed, age, or size. The two Malinois did not cluster together when the data from 23 tissue types was concatenated, and the small dog was not an outlier (Figure 5B). When tissues were examined separately, the tree topology was highly variable. The two Malinois did not appear to cluster together more frequently than other pairs of dogs, nor was the small dog an outlier (Figure 6). This is not surprising, as we are comparing gene expression across all genes, and studies in humans show both that expression levels are strongly heritable for only a small subset of genes [93], and that subtle changes in expression, or changes involving rarer cell types, may be masked in tissue-level data [94].

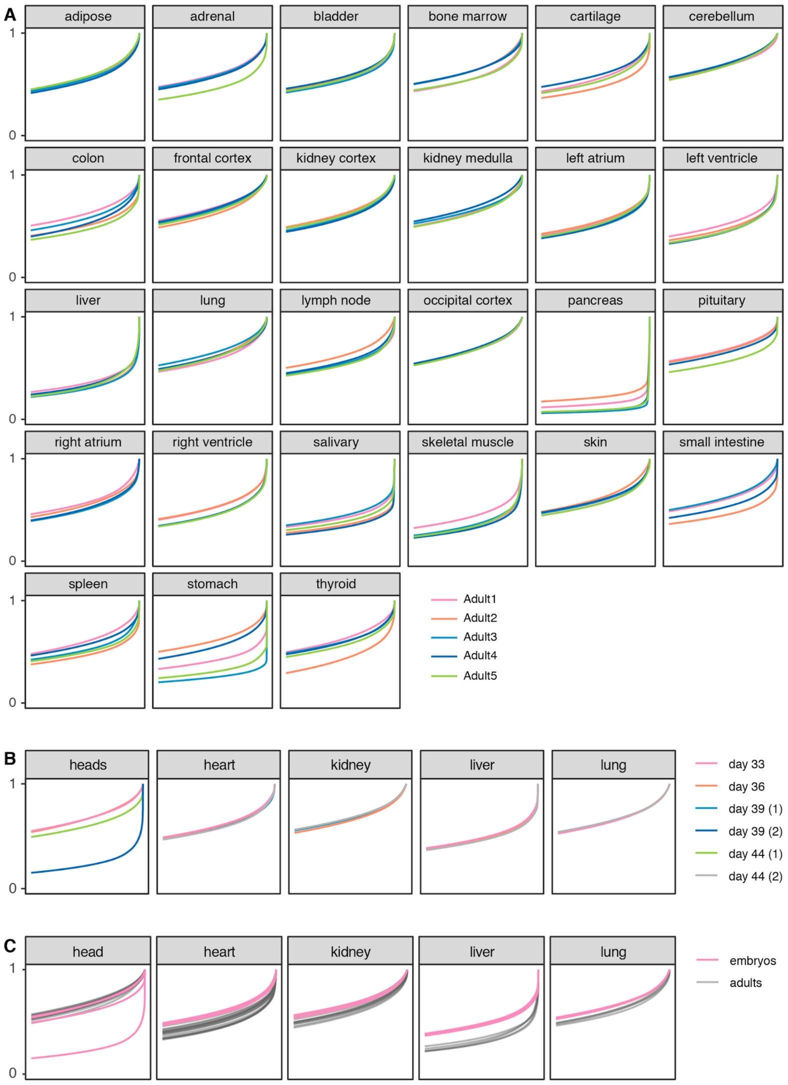


Figure 4. Cumulative transcriptome expression is matched to tissue type. Cumulative sum of fraction of tissue-specific transcriptomes represented by individual genes in (**A**) canine embryos at four gestational time points; and (**B**) up to five individual adult dogs. Single-gene counts per million (CPM) values were divided by sample-sum CPM, sorted in increasing order, and the cumulative sum calculated. Cumulative values are shown for the 1000 top-expressed genes in each sample. Data sampled from a given embryonic tissue at different gestational time points are very similar, perhaps reflecting the fairly narrow time window of sampling. Combining data from adult and embryonic samples (**C**) reveals strong similarity of data from given tissue types across individuals and developmental stages.

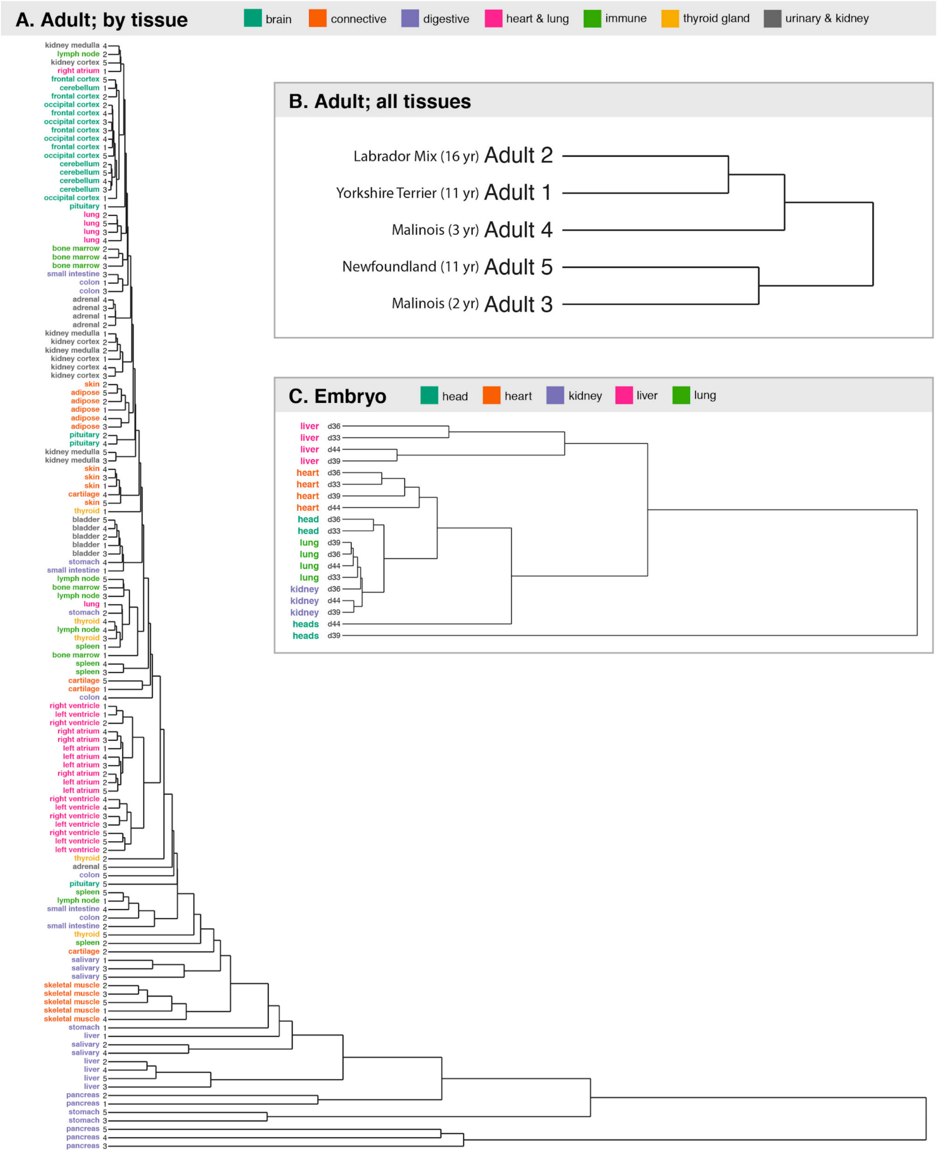


Figure 5. Transcriptome data from five individuals clusters primarily by tissue type. Hierarchical clustering of RNA-seq data from (**A**) single tissues of five adult dogs; (**B**) five adult dogs, based on data concatenated across 21 tissues; and (**C**) embryonic tissues sampled at four gestational time points. Clustering is based on Euclidean distances among samples. Overall, in data from both adults and embryos, samples of a given tissue cluster across individuals. As observed in cumulative analysis, embryonic samples of a given tissue type cluster despite variation in gestational time points, perhaps reflecting the fairly narrow time window of sampling.

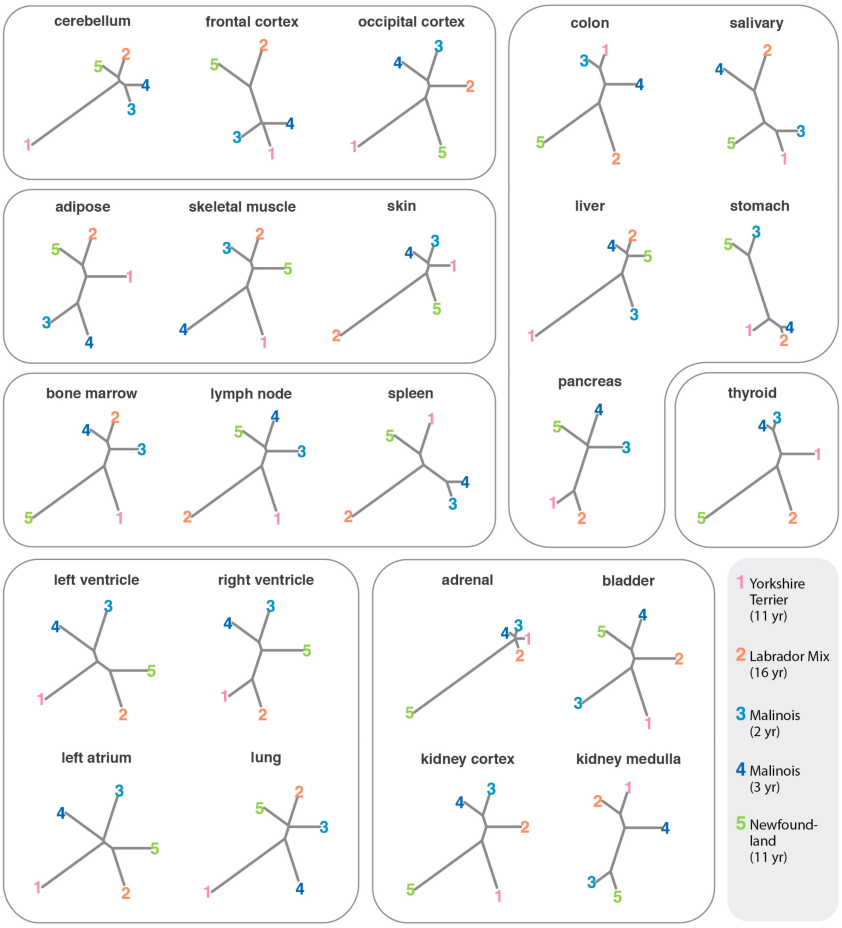
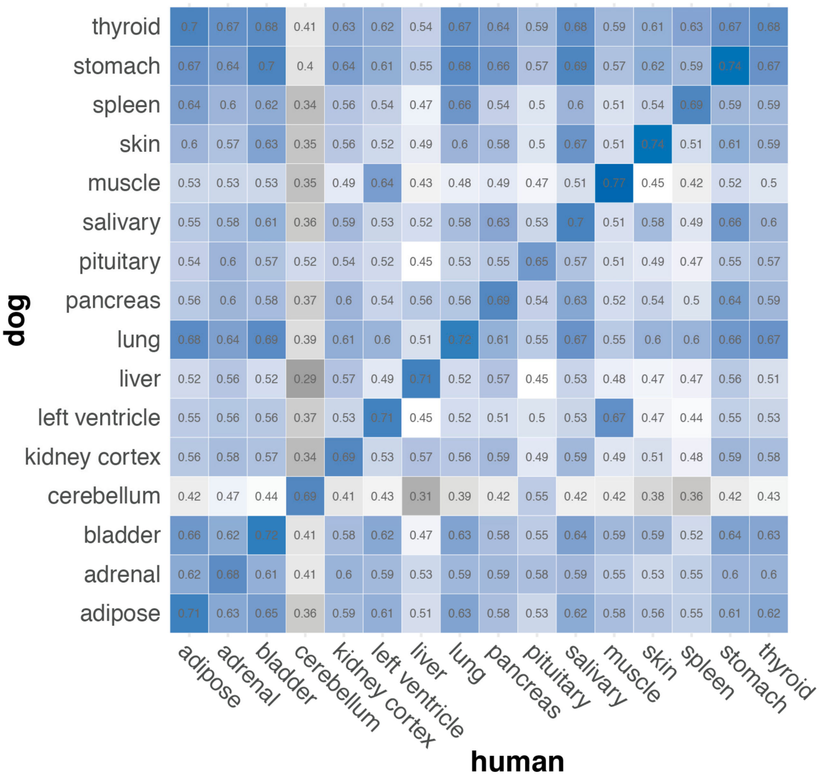


Figure 6. The relationship between samples within a single tissue type is highly variable. Clustering is based on Euclidean distances among samples, with no consistent clustering by age or breed observed. Outlines group tissues of a given class.

3.5. Comparison of Expression Profiles in Canine and Human Tissues

Substantial congruity between human and canine tissue-specific transcriptomes is evident from a comparison of our new BarkBase RNA-seq data to human RNA-seq data publicly available through the GTEx Portal [95]. We assessed this simply by measuring the rank-order correlation of the single-gene expression values for genes expressed in dog and human tissues. While this correlation metric varies only moderately across tissues, the strongest correlation was almost always between the same tissue type in dogs and human (100% of human tissues and 94% of dog tissues; Figure 7). Human and dog skeletal muscle had the highest correlation ($r^2 = 0.77$), and pituitary gland the lowest ($r^2 = 0.65$). The only tissue not to match the same tissue type in the other species was dog thyroid tissue, which matched human adipose ($r^2 = 0.70$) slightly better than human thyroid ($r^2 = 0.68$). Comparing cerebellum with any tissue type other than cerebellum ($r^2 = 0.69$) yielded some of the lowest correlations, with all except pituitary ($r^2 = 0.52$–0.55) between $r^2 = 0.29$ and $r^2 = 0.47$. This is consistent with the tight clustering of cerebellum samples as a distinct clade on our tissue dendrogram (Figure 5A) and with previous reports that cerebellum is a complex tissue composed of a large number of cell types with transcriptomic profiles distinct from those in other organs [96].



	adipose	adrenal	bladder	cerebellum	kidney cortex	left ventricle	liver	lung	pancreas	pituitary	salivary	muscle	skin	spleen	stomach	thyroid
thyroid	0.7	0.67	0.68	0.41	0.63	0.62	0.54	0.67	0.64	0.59	0.68	0.59	0.61	0.63	0.67	0.68
stomach	0.67	0.64	0.7	0.4	0.64	0.61	0.55	0.68	0.66	0.57	0.69	0.57	0.62	0.59	0.74	0.67
spleen	0.64	0.6	0.62	0.34	0.56	0.54	0.47	0.66	0.54	0.5	0.6	0.51	0.54	0.69	0.59	0.59
skin	0.6	0.57	0.63	0.35	0.56	0.52	0.49	0.6	0.58	0.5	0.67	0.51	0.74	0.51	0.61	0.59
muscle	0.53	0.53	0.53	0.35	0.49	0.64	0.43	0.48	0.49	0.47	0.51	0.77	0.45	0.42	0.52	0.5
salivary	0.55	0.58	0.61	0.36	0.59	0.53	0.52	0.58	0.63	0.53	0.7	0.51	0.58	0.49	0.66	0.6
pituitary	0.54	0.6	0.57	0.52	0.54	0.52	0.45	0.53	0.55	0.65	0.57	0.51	0.49	0.47	0.55	0.57
pancreas	0.56	0.6	0.58	0.37	0.6	0.54	0.56	0.56	0.69	0.54	0.63	0.52	0.54	0.5	0.64	0.59
lung	0.68	0.64	0.69	0.39	0.61	0.6	0.51	0.72	0.61	0.55	0.67	0.55	0.6	0.6	0.66	0.67
liver	0.52	0.56	0.52	0.29	0.57	0.49	0.71	0.52	0.57	0.45	0.53	0.48	0.47	0.47	0.56	0.51
left ventricle	0.55	0.56	0.56	0.37	0.53	0.71	0.45	0.52	0.51	0.5	0.53	0.67	0.47	0.44	0.55	0.53
kidney cortex	0.56	0.58	0.57	0.34	0.69	0.53	0.57	0.56	0.59	0.49	0.59	0.49	0.51	0.48	0.59	0.58
cerebellum	0.42	0.47	0.44	0.69	0.41	0.43	0.31	0.39	0.42	0.55	0.42	0.42	0.38	0.36	0.42	0.43
bladder	0.66	0.62	0.72	0.41	0.58	0.62	0.47	0.63	0.58	0.55	0.64	0.59	0.59	0.52	0.64	0.63
adrenal	0.62	0.68	0.61	0.41	0.6	0.59	0.53	0.59	0.59	0.58	0.59	0.55	0.53	0.55	0.6	0.6
adipose	0.71	0.63	0.65	0.36	0.59	0.61	0.51	0.63	0.58	0.53	0.62	0.58	0.56	0.55	0.61	0.62

Figure 7. Gene expression levels correlate between dog and human tissues. Heatmap showing Spearman correlation between the genes expressed in canine and human tissues, after filtering for minimum expression (median CPM > 1) and unique orthology mapping between species. In all cases except one (dog thyroid), comparison of dog tissue to the corresponding human tissue had the highest Spearman coefficient, suggesting broad conservation of the transcriptome in these tissues across species.

Same tissue comparisons between dog and human yielded higher correlations (mean ^{2} = 0.71 ± 0.03) than comparing the different tissues between dog and human (mean = 0.55 ± 0.08) and than comparing different tissues within dog or human (0.62 ± 0.11). Together, these results suggest that tissue-specific dog transcriptomes have much in common with human transcriptomes of the corresponding tissues. Precisely why our data indicated a difference between the same tissue in dog and human will require further work, including determining whether the source of the deviation is primarily artifactual, arising from differences in sampling, data generation and analysis, or whether it reflects true biological differences.

3.6. RNA-Sequencing of Embryonic Tissues at Multiple Time Points

RNA-seq data from embryos revealed strong overlap of data for individual tissues assayed at various gestational stages as assessed by both cumulative transcription across the 1000 top-expressed genes (Figure 4), and by tissue-level clustering (Figure 5). To identify the top differentially expressed functional categories in each tissue, we first mapped all genes with FDR < 0.1 to their known human ortholog. In total, 158 genes with human orthologs were differentially expressed between embryonic data 36 and day 44 across head, heart, kidney, and liver, with the number varying by tissue type: 71 genes for head, 41 genes for heart, two genes for kidney, and 44 genes for liver. In the lung, 8365 genes were differentially expressed. These two embryonic time points were both different individuals and different sexes, contributing to differences in expression.

We analyzed these sets with an ingenuity pathway analysis and identified the top "Diseases and Bio Functions" (Table 1). The most significantly enriched set was "epithelial neoplasm" in the lung ($p = 2 \times 10^{-323}$), consistent with ongoing differentiation of lung epithelial cells across mammalian gestation [97]. In the head, both top sets were related to organ enlargement: hypertrophy ($p = 9.8 \times 10^{-10}$) and visceromegaly ($p = 1.6 \times 10^{-9}$). In the heart, the top set, "Morphogenesis of embryonic skeleton" ($p = 4.4 \times 10^{-12}$), included seven *HOXA* genes and one *HOXB* gene. *HOX* genes encode proteins whose spatial and temporal regulation guides embryonic development, and *HOXA* gene mutations are implicated in congenital heart defects [98,99]. In the kidneys, there was only one gene underlying the functional enrichment, *EMX2*, a gene involved in both urogenital and neurological development [100]. In liver, the top function was a liver-related disease (Hepatitis B virus-related hepatocellular carcinoma, $p = 9.2 \times 10^{-5}$). Other top sets included hyperphenylalaninemia ($p = 2.3 \times 10^{-4}$), a disease resulting from a lack of a liver-specific enzyme, and metabolism of acylglycerol ($p = 3.2 \times 10^{-4}$), a pathway involving the liver.

Table 1. Functional enrichment among genes differentially expressed in embryonic tissues at d36 and d44 reflects organ-specific roles. Genes differentially expressed in each of the five individual tissues sampled at embryonic d36 as compared d44 (FDR < 0.1) were analyzed for functional enrichment using IPA.

	Category	Diseases or Functions	*p*	No. of Genes	Genes
head	Organismal Injury and Abnormalities	Hypertrophy	9.8×10^{-10}	15	*TRIM55, CXCL12, TNNI3K, TNNT2, ADGRG1, GATA6, JPH2, INHBA, RPS6KA2, NR3C1, SLC25A4, CAST, RRAD, TRIM63, IL33*
		Visceromegaly	1.6×10^{-9}	18	*TRIM55, CXCL12, TNNI3K, TNNT2, NEXN, ADGRG1, GATA6, JPH2, INHBA, TBX20, NR3C1, SLC25A4, CAST, RRAD, SSTR2, BIK, TRIM63, IL33*
	Cardiovascular Disease, Cardiovascular System Development and Function, Organ Morphology, Organismal Development, Organismal Injury and Abnormalities	Enlargement of heart	3.7×10^{-9}	15	*TRIM55, CXCL12, TNNI3K, TNNT2, NEXN, ADGRG1, GATA6, JPH2, INHBA, TBX20, SLC25A4, CAST, RRAD, TRIM63, IL33*
		Abnormal morphology of heart	5.0×10^{-9}	17	*TRIM55, CXCL12, TNNI3K, TNNT2, NEXN, ADGRG1, GATA6, JPH2, DHRS3, INHBA, TBX20, RPS6KA2, SLC25A4, CAST, RRAD, TRIM63, IL33*
		Muscular hypertrophy	5.7×10^{-9}	10	*INHBA, TRIM55, RPS6KA2, CAST, RRAD, ADGRG1, GATA6, JPH2, TRIM63, IL33*
		Hypertrophy of heart	1.6×10^{-7}	11	*CXCL12, INHBA, TRIM55, TNNI3K, SLC25A4, TNNT2, RRAD, ADGRG1, GATA6, TRIM63, IL33*
	Cardiovascular System Development and Function	Morphology of cardiovascular system	6.4×10^{-9}	19	*CXCL12, TRIM55, TNNI3K, TNNT2, PLA2G7, NEXN, ADGRG1, GATA6, JPH2, DHRS3, INHBA, TBX20, RPS6KA2, SLC25A4, RRAD, CAST, SSTR2, TRIM63, IL33*
	Cardiovascular Disease, Cardiovascular System Development and Function	Abnormal morphology of cardiovascular system	8.2×10^{-9}	18	*TRIM55, CXCL12, TNNI3K, TNNT2, NEXN, ADGRG1, GATA6, JPH2, DHRS3, INHBA, TBX20, RPS6KA2, SLC25A4, CAST, RRAD, SSTR2, TRIM63, IL33*
	Organismal Development, Organismal Injury and Abnormalities	Abnormal morphology of thoracic cavity	2.9×10^{-8}	18	*TRIM55, CXCL12, TNNI3K, TNNT2, NEXN, ADGRG1, GATA6, JPH2, DHRS3, INHBA, TBX20, RPS6KA2, NR3C1, SLC25A4, CAST, RRAD, TRIM63, IL33*

Table 1. *Cont.*

	Category	Diseases or Functions	p	No. of Genes	Genes
heart	Organismal Development	Abnormal morphology of body cavity	8.0×10^{-8}	22	*TRIM55, RBMS1, CXCL12, TNNI3K, TNNT2, MAPK8IP2, NEXN, ADGRG1, GATA6, JPH2, DHRS3, INHBA, TBX20, RPS6KA2, NR3C1, SLC25A4, RRAD, CAST, BIK, SSTR2, TRIM63, IL33*
	Skeletal and Muscular System Development and Function	Morphogenesis of embryonic skeleton	4.4×10^{-12}	7	*HOXB8, HOXA6, HOXA3, HOXA7, HOXA4, HOXA2, HOXA5*
		Morphology of axial skeleton	1.1×10^{-8}	8	*HSD11B2, HOXB8, HOXA3, HOXA6, HOXB9, HOXA4, HOXA5, mir-196*
		Fusion of bone	1.6×10^{-8}	6	*HOXA6, HOXA3, HOXB9, HOXA7, HOXA4, HOXA5*
		Morphology of skeleton	1.8×10^{-8}	9	*HSD11B2, HOXB8, HOXA6, HOXA3, HOXB9, HOXA4, HOXA2, mir-196, HOXA5*
	Embryonic Development, Organismal Development	Patterning of rostrocaudal axis	1.2×10^{-11}	8	*HOXB8, HOXA6, HOXA3, HOXB9, HOXA7, HOXA4, HOXA2, HOXA5*
	Organismal Development	Abnormal morphology of body cavity	9.2×10^{-9}	17	*TRIM55, HSD11B2, MYH7, SMYD1, TNNC1, TNNI3K, HOXA3, HOXB9, ATP2A2, HOXA5, PDZK1, TBX20, SLC25A4, HOXA7, HOXA2, TRIM63, SGPP2*
	Cardiovascular System Development and Function, Organ Development, Organ Morphology, Skeletal and Muscular System Development and Function	Contraction of cardiac muscle	3.9×10^{-8}	6	*MYH7, TNNC1, TNNI3K, ATP2A2, TRIM63, SRL*
	Organ Morphology, Skeletal and Muscular System Development and Function	Quantity of rib	6.0×10^{-8}	5	*HOXA6, HOXB9, HOXA4, HOXA5, mir-196*
	Cancer, Skeletal and Muscular Disorders, Tissue Morphology	Transformation of vertebrae	7.1×10^{-8}	5	*HOXA6, HOXB9, HOXA4, HOXA5, mir-196*
	Organismal Development, Organismal Injury and Abnormalities	Abnormal morphology of thoracic cavity	8.8×10^{-8}	13	*TRIM55, MYH7, SMYD1, TNNC1, TNNI3K, HOXA3, HOXB9, ATP2A2, HOXA5, TBX20, SLC25A4, HOXA7, TRIM63*
kidney	Cell Cycle	Cell division of neural stem cells	9.1×10^{-5}	1	*EMX2*
	Embryonic Development, Nervous System Development and Function, Organ Development, Organismal Development, Tissue Development	Development of hippocampal fissure	9.1×10^{-5}	1	*EMX2*
	Nervous System Development and Function, Organ Morphology, Organismal Development	Size of primary visual cortex	9.1×10^{-5}	1	*EMX2*
	Nervous System Development and Function, Neurological Disease, Organ Morphology, Organismal Development, Organismal Injury and Abnormalities	Abnormal morphology of medial ganglionic eminences	1.8×10^{-4}	1	*EMX2*
	Developmental Disorder, Embryonic Development, Tissue Morphology	Degeneration of Wolffian duct	1.8×10^{-4}	1	*EMX2*

Table 1. *Cont.*

	Category	Diseases or Functions	*p*	No. of Genes	Genes
liver	Cancer, Gastrointestinal Disease, Hepatic System Disease, Organismal Injury and Abnormalities	Hepatitis B virus-related hepatocellular carcinoma	9.2×10^{-5}	3	*TF, ALDOB, RGN*
	Cell-To-Cell Signaling and Interaction, Renal and Urological System Development and Function	Activation of kidney cells	9.8×10^{-5}	2	*TF, MST1*
	Organismal Injury and Abnormalities	Organ Degeneration	9.8×10^{-5}	8	*EFEMP1, GSTZ1, TF, GRID2, RP2, RGN, ZNF408, mir-22*
	Developmental Disorder, Hereditary Disorder, Metabolic Disease, Organismal Injury and Abnormalities	Hyperphenylalaninemia	2.3×10^{-4}	2	*GCH1, DNAJC12*
	Lipid Metabolism, Small Molecule Biochemistry	Metabolism of acylglycerol	3.2×10^{-4}	4	*ACSL5, SLC22A4, RGN, F2*
lung	Cancer,Organismal Injury and Abnormalities	Epithelial neoplasm	0.0	6961	*many*
		Non-hematological solid tumor	0.0	7039	*many*
		Nonhematologic malignant neoplasm	0.0	7021	*many*
		Carcinoma	0.0	6949	*many*
		Tumorigenesis of tissue	0.0	6969	*many*

3.7. ATAC-seq

ATAC-seq (assay for transposase-accessible chromatin using sequencing) assesses chromatin accessibility genome-wide [101,102]. Pairing ATAC-seq data with RNA-seq data places the transcriptome in the context of chromatin states genomewide. This enables identification of *cis*-regulatory elements, such as enhancers, promoters, and insulators, which are not detectable using RNA-seq alone, as well as validation of novel transcription start sites relevant to a specific set of RNAs [101,103].

BarkBase includes a pilot set of ATAC-seq data for 15 tissues sampled from five individual adult dogs (Figure 1). Of 46 ATAC-seq libraries created, 37 yielded sufficiently high quality data for inclusion. Each of these 37 samples has between 4600 and 136,000 ATAC-seq peaks (median 50,000). Known transcription start sites (TSS) are highly enriched for ATAC-seq peaks. A close look at two tissue types for which we have data from multiple individuals (pancreas and salivary gland) shows the pattern of enrichment is similar across individuals, but varies between tissue types (Figure 8A,B). Moreover, most of the peaks are either intronic or intergenic, illustrating the capacity of ATAC-seq to annotate potentially important genomic loci not found using RNA-seq alone. This pattern persists across all tissues tested, with 20–40% of peaks falling in distal intergenic regions and strong enrichment for transcription start sites (Figure 8C,D).

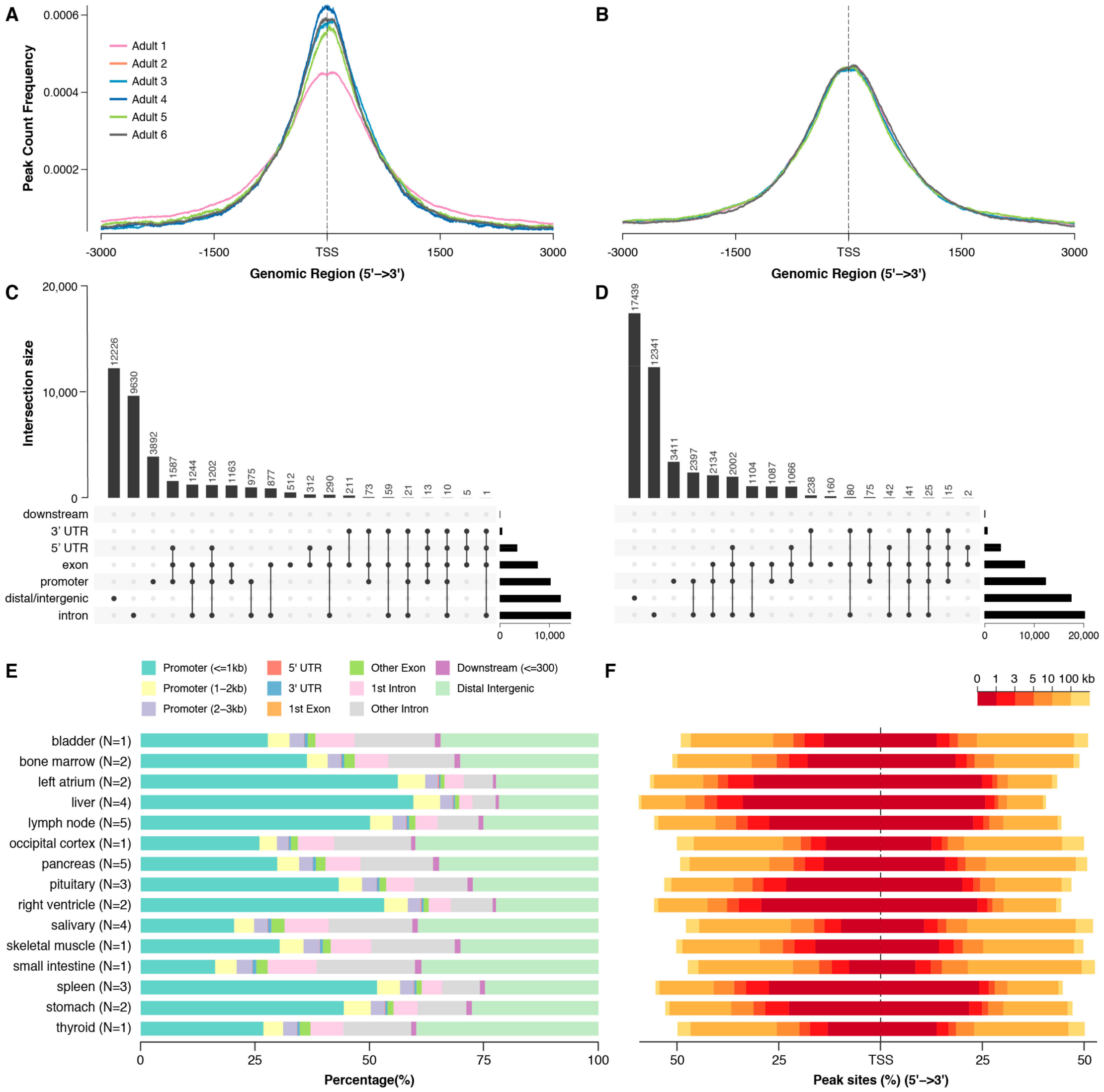


Figure 8. ATAC-seq maps transposase-accessible chromatin in canine tissues. Analysis of the two tissue types with ATAC-seq data for five individuals, pancreas (**A**) and salivary gland (**B**), reveals strong enrichment of peaks around known transcription start sites. This enrichment is consistent across individuals. Annotating the ATAC-seq peaks with ChIPseeker, using the Ensembl annotation of dog, shows, as expected, an overlap with known promoters in both (**C**) the pancreas ($n \cong 10{,}000$) and (**D**) salivary gland ($n \cong 12{,}000$), but there are more peaks in distal/intergenic regions, potentially marking novel promoters or distal regulatory elements. (**E**) Across all tissues, ATAC-seq peaks are most likely to be in annotated promoters, but a large proportion are far from genes. (**F**) In all tissues, the enrichment for ATAC-seq peaks falls off rapidly with increasing distance from a TSS.

For those tissues for which we have multiple samples, we also find that large numbers of peaks are shared among tissues of the same type, as well as a significant number of peaks that are shared across all members of the same tissue, but are present in no other tissue. For example, we find 35,000 peaks shared across the five pancreas samples; of these, 8000 are unique to the pancreas. This combination suggests both that the peaks we see are reflective of real-world differences, and that we have captured regulatory elements specific to tissue type.

3.8. Novel Genes and ATAC

BarkBase includes 84 Mb of transcribed sequence that is not found in the existing Ensembl canine reference annotation. This includes 769 novel, multi-exon genes. We assessed the potential for

ATAC-seq to discern which genes are most likely to be real, focusing on the pancreas, the tissue for which we had the most data (Figure 1). After excluding genes with a mean expression under 0.16 CPM in all samples, we found 44 novel genes expressed in the pancreas in at least two dogs, 58 novel genes expressed in the salivary gland in at least two dogs, and 491 novel genes not expressed in the pancreas in any dogs. We observed that novel genes expressed in the pancreas tended to be closer in proximity to a pancreas ATAC-seq peak than novel genes expressed in a different tissue (salivary gland), or novel genes expressed in any tissue but pancreas (Figure 9).

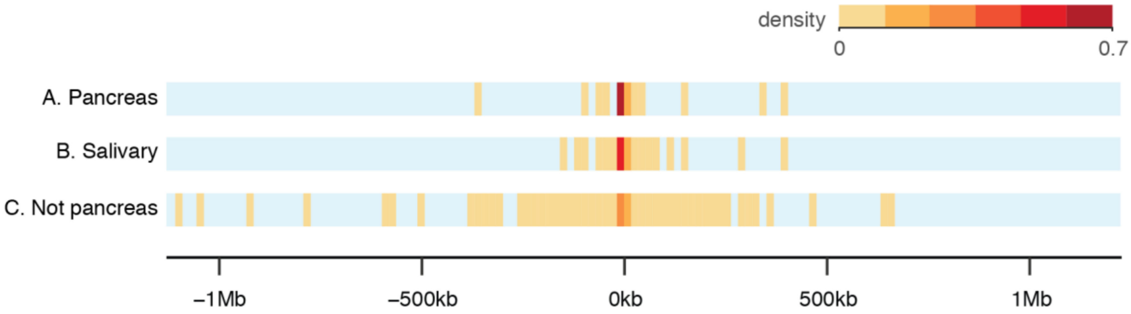


Figure 9. Integrating ATAC-seq with RNA-seq data can help validate novel genes. (**A**) Of the 44 novel genes expressed in the pancreas, most are less than 25 kb from a pancreas ATAC-seq peak. For those closest to ATAC-seq peaks, integrating RNA-seq and ATAC-seq provides additional evidence that they are real genes. (**B**) 58 novel genes expressed in the salivary gland (including 15 also expressed in pancreas) do not cluster as closely to pancreas ATAC-seq peaks, suggesting tissue specificity. (**C**) 491 novel genes not expressed in the pancreas are much more dispersed relative to the ATAC-seq peaks in the pancreas.

4. Discussion

BarkBase contains the largest and most comprehensive set of canine functional genomic data produced to date. Overall, data quality is high. Comparing either RNA-seq or ATAC-seq data from different individuals for a particular tissue type shows the high degree of congruence expected from previous work in humans [95,104]. RNA-seq data sets for up to 27 tissues from each of the five dogs are available for immediate download at BarkBase.org; corresponding ATAC-seq data sets discussed here are available as well. Libraries have been constructed for most of the remaining samples, and data will be posted on BarkBase as soon as sequencing is complete.

BarkBase improves substantially on earlier annotations of the dog genome. By analyzing data directly from dog tissues, rather than making inferences from human and lifting them over to the dog genome, our method improves sensitivity by enabling detection of genes expressed in dogs but not humans. With RNA-seq data from 150 samples from diverse tissues, including embryonic tissue, we detect novel genes not identified in earlier RNA-seq informed annotations. These newly identified genes include five orthologous to protein-coding genes known from human studies to function in health-relevant processes including tooth development (*STATH*) and immune function (*AHNAK2*, *EPX*). In addition, these five genes have been identified in various roles in cancer studies [105–108], highlighting the utility of BarkBase as a tool for advancing dog as a model for human medicine.

BarkBase is complementary to existing annotations, which include genes and transcripts that we miss. Integrating these "lost" transcripts should further improve the annotation. The genes missing from BarkBase are largely non-coding genes without human orthologs, including multi-exon lncRNAs. lncRNAs tend to have more tissue-specific expression patterns [109], including in dogs [25], and we may have missed them because of the difference in tissue types represented between our dataset and others. However, non-coding genes are also challenging to annotate. Using RNA-seq data alone, as was done for both Ensembl and BarkBase, is only the first step [110]. ATAC-seq data from BarkBase adds another level of information, guiding predictions of non-coding genes physically near the genomic elements that regulate them (Figure 9). We will integrate the ATAC-seq data into our gene predictions once the complete set of ATAC-seq data is available. Additional information will be required for the prediction of genes regulated across larger expanses of sequence [111].

Using the RNA-seq data from multiple individuals, we can compare patterns of gene expression across individuals at single-tissue resolution. We observe that samples tend to cluster by tissue type, not by individual (Figure 5A). Comparing the relationship between individuals for each tissue type, we see no evidence that particular individuals cluster together more or less often (Figure 6). This suggests that, at least within our small dataset, the effects of size, age, and breed do not substantially alter the overall transcriptional landscape. Additional data from much larger sets of individuals, with additional biological replicates and single-cell transcriptomes, may reveal subtle patterns and gene-specific differences not accessible from our present data set or analytic approach. Nonetheless, data already available through BarkBase indicate that tissue identity, not characteristics of the individual, is the strongest predictor of fundamental transcriptome features.

We note that any effort to discern breed effects on gene expression would require a much larger dataset controlled for environmental factors, such as age and lifestyle, or phenotypes, such as size (dwarfism and gigantism), that might also be expected to affect gene expression and can spuriously correlate with breed [112]. For example, in our data we observed that the shortest distance between any two organ-specific samples is between thyroid samples from Adult 3 and Adult 4, the two younger Malinois dogs. It is tempting to speculate that this is associated with the Malinois' reported higher risk of thyroid disease, or with age-related changes in thyroid function, but we have no ability to distinguish age effects from breed effects [113,114]. Work investigating how germline variants affect gene expression in humans suggests sample sizes in the hundreds may be required [115].

We also saw very little difference in overall gene expression patterns between embryonic time points, with samples clustering almost perfectly by tissue type (Figure 5C). The two exceptions were the two head samples that had pooled multiple individuals, likely reflecting the increased variability in gene expression in these samples. This was, at first, unexpected, as the transcriptome of individual embryonic tissues is known to shift substantially across embryonic development [116]. However, we also observed that the cumulative distribution of transcripts closely matched the adult tissues (Figure 4C). One possible explanation is that our sampling of embryos more than halfway through gestation, coupled with their spread across a fairly narrow, nine-day time window, does not capture the dramatic epigenomic and transcriptomic shifts that characterize preimplantation and early developmental shifts in, for example, mouse [116,117] and human [118]. Instead, we confirm that tissue-specific embryonic transcriptomes are broadly stable over a ~10d window midway through gestation and embryonic transcriptomes, and are broadly similar to transcriptomes of matched adult tissues.

Higher resolution comparisons through differential gene expression (DGE) revealed developmental changes not evident from overall gene expression patterns. For example, IPA analysis of genes differing in expression between heart tissue sampled at these two gestational time points revealed enrichment for genes associated with abnormalities of heart development. DGE in liver revealed enrichment of genes associated with metabolism, and in lung revealed genes associated with epithelial neoplasm and cancers. These findings indicate that BarkBase will be a powerful tool for identifying specific genes and classes of genes whose activity shifts during gestation, and suggest future efforts should focus on expanding on this work to capture the full arc of transcriptomic changes across canine development in utero.

5. Conclusions

BarkBase dramatically expands genomic resources for dogs, improving the annotation of the canine genome and revealing close similarities between dog transcriptomes and tissue-matched data from humans. BarkBase samples already span the typical canine lifespan, including development in utero, and includes individuals from several different breeds, but a more comprehensive data set will be required for inference of potential associations of transcriptomic features to dog age, breed, and/or environmental factors. We therefore offer BarkBase both as a powerful resource that is available to all researchers for immediate use, and as a paradigm for ongoing collection of data to further enhance the value of dog as a powerful natural model for human disease.

Supplementary Materials: The following are available online at http://www.mdpi.com/2073-4425/10/6/433/s1: Table S1. All samples in BarkBase. Table S2. Variant Calling Background Information. Table S3. Sensitivity to detect annotated exons. Table S4. Ensembl genes detected in BarkBase. Table S5. Ensembl genes not seen in BarkBase. Table S6. GO enrichment of Ensembl genes not seen in BarkBase. Table S7. GO enrichment of Ensembl genes detected in BarkBase. Table S8. Human genes overlapped by novel BarkBase transcripts.

Author Contributions: Conceptualization, C.S., L.F., K.L.-T., V.N.M.-W., A.H. and E.K.K.; Data curation, K.J.M., D.P.G., J.H., R.S., J.T.-M., J.J., X.L. and K.M.M.; Formal analysis, K.J.M., D.P.G., J.H., R.S., J.T.-M., X.L., K.M.M. and E.K.K.; Funding acquisition, C.S., L.F., K.L.-T., V.N.M.-W., A.H. and E.K.K.; Investigation, K.J.M., D.P.G., J.H., R.S., J.T.-M., J.A., X.L., K.M.M. and E.K.K.; Methodology, K.J.M., D.P.G., V.N.M.-W., A.H. and E.K.K.; Project administration, J.J. and E.K.K.; Resources, K.J.M., D.P.G., J.H., R.S., J.T.-M., J.A., L.J.A., M.K., B.L., V.N.M.-W. and A.H.; Software, K.J.M., D.P.G., J.H., R.S. and J.T.-M.; Supervision, V.N.M.-W., A.H. and E.K.K.; Validation, K.J.M., D.P.G., J.H., R.S. and J.T.-M.; Visualization, K.J.M. and E.K.K.; Writing—original draft, K.J.M., D.P.G., J.H. and E.K.K.; Writing—review & editing, K.J.M., D.P.G., J.H., R.S., J.T.-M., J.J., J.A., X.L., K.M.M., K.L.-T., V.N.M.-W., A.H. and E.K.K.

Funding: This research was funded in part by the NIH under OD grant 5R24OD018250, NCI grant 5R37CA218570; NIMH grant R21MH109938 and NHGRI grant 1R01HG008742.

Acknowledgments: The authors would like to thank all the dog owners and dogs, without whom this work would not have been possible. In particular, we thank the dog owners who chose to participate in the Deceased Dog Donation program at Cummings School of Veterinary Medicine after losing their beloved pet. At the Cummings School of Veterinary Medicine, we also thank Asma Chaudri, who assisted with sample collection and the Deceased Dog Donation program staff. We thank SciStories LLC., for the canine anatomy illustrations and the BarkBase website; E Ostrander, C Drögemüller, J Modiano, JR Mickleson, K Minor and B. Klein, who shared dog whole genome sequence data to support variant calling; the Broad Institute Genomics Platform. The Genotype-Tissue Expression (GTEx) Project, from which we downloaded human RNA-seq data on 04/28/19 for comparison with our data from dog, was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Schiffman, J.D.; Breen, M. Comparative oncology: What dogs and other species can teach us about humans with cancer. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **2015**, *370*, 20140231. [CrossRef] [PubMed]
2. Noh, H.J.; Tang, R.; Flannick, J.; O'Dushlaine, C.; Swofford, R.; Howrigan, D.; Genereux, D.P.; Johnson, J.; van Grootheest, G.; Grünblatt, E.; et al. Integrating evolutionary and regulatory information with a multispecies approach implicates genes and pathways in obsessive-compulsive disorder. *Nat. Commun.* **2017**, *8*, 774. [CrossRef] [PubMed]
3. Wilbe, M.; Jokinen, P.; Truvé, K.; Seppala, E.H.; Karlsson, E.K.; Biagi, T.; Hughes, A.; Bannasch, D.; Andersson, G.; Hansson-Hamlin, H.; et al. Genome-wide association mapping identifies multiple loci for a canine SLE-related disease complex. *Nat. Genet.* **2010**, *42*, 250–254. [CrossRef] [PubMed]
4. Karlsson, E.K.; Sigurdsson, S.; Ivansson, E.; Thomas, R.; Elvers, I.; Wright, J.; Howald, C.; Tonomura, N.; Perloski, M.; Swofford, R.; et al. Genome-wide analyses implicate 33 loci in heritable dog osteosarcoma, including regulatory variants near CDKN2A/B. *Genome Biol.* **2013**, *14*, R132. [CrossRef] [PubMed]
5. Schoenebeck, J.J.; Ostrander, E.A. Insights into morphology and disease from the dog genome project. *Annu. Rev. Cell Dev. Biol.* **2014**, *30*, 535–560. [CrossRef]
6. Kaeberlein, M.; Creevy, K.E.; Promislow, D.E.L. The dog aging project: Translational geroscience in companion animals. *Mamm. Genome* **2016**, *27*, 279–288. [CrossRef]
7. Karlsson, E.K.; Lindblad-Toh, K. Leader of the pack: Gene mapping in dogs and other model organisms. *Nat. Rev. Genet.* **2008**, *9*, 713–725. [CrossRef]

8. Lindblad-Toh, K.; Wade, C.M.; Mikkelsen, T.S.; Karlsson, E.K.; Jaffe, D.B.; Kamal, M.; Clamp, M.; Chang, J.L.; Kulbokas, E.J., 3rd; Zody, M.C.; et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **2005**, *438*, 803–819. [CrossRef]
9. Parker, H.G.; Kim, L.V.; Sutter, N.B.; Carlson, S.; Lorentzen, T.D.; Malek, T.B.; Johnson, G.S.; DeFrance, H.B.; Ostrander, E.A.; Kruglyak, L. Genetic structure of the purebred domestic dog. *Science* **2004**, *304*, 1160–1164. [CrossRef]
10. Karlsson, E.K.; Baranowska, I.; Wade, C.M.; Salmon Hillbertz, N.H.C.; Zody, M.C.; Anderson, N.; Biagi, T.M.; Patterson, N.; Pielberg, G.R.; Kulbokas, E.J., 3rd; et al. Efficient mapping of mendelian traits in dogs through genome-wide association. *Nat. Genet.* **2007**, *39*, 1321–1328. [CrossRef]
11. Ostrander, E.A.; Wang, G.-D.; Larson, G.; vonHoldt, B.M.; Davis, B.W.; Jagannathan, V.; Hitte, C.; Wayne, R.K.; Zhang, Y.-P. Dog10K: An international sequencing effort to advance studies of canine domestication, phenotypes, and health. *Natl. Sci. Rev.* **2019**. [CrossRef]
12. Plassais, J.; Kim, J.; Davis, B.W.; Karyadi, D.M.; Hogan, A.N.; Harris, A.C.; Decker, B.; Parker, H.G.; Ostrander, E.A. Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. *Nat. Commun.* **2019**, *10*, 1489. [CrossRef] [PubMed]
13. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **2012**, *489*, 57–74. [CrossRef]
14. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **2013**, *45*, 580–585. [CrossRef] [PubMed]
15. Roadmap Epigenomics, Consortium; Kundaje, A.; Meuleman, W.; Ernst, J.; Bilenky, M.; Yen, A.; Heravi-Moussavi, A.; Kheradpour, P.; Zhang, Z.; Wang, J.; et al. Integrative analysis of 111 reference human epigenomes. *Nature* **2015**, *518*, 317–330.
16. 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* **2012**, *337*, 1190–1195. [CrossRef]
17. Hoffman, M.M.; Ernst, J.; Wilder, S.P.; Kundaje, A.; Harris, R.S.; Libbrecht, M.; Giardine, B.; Ellenbogen, P.M.; Bilmes, J.A.; Birney, E.; et al. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.* **2013**, *41*, 827–841. [CrossRef]
18. Ernst, J.; Kheradpour, P.; Mikkelsen, T.S.; Shoresh, N.; Ward, L.D.; Epstein, C.B.; Zhang, X.; Wang, L.; Issner, R.; Coyne, M.; et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **2011**, *473*, 43–49. [CrossRef]
19. Gjoneska, E.; Pfenning, A.R.; Mathys, H.; Quon, G.; Kundaje, A.; Tsai, L.-H.; Kellis, M. Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature* **2015**, *518*, 365–369. [CrossRef]
20. Hoeppner, M.P.; Lundquist, A.; Pirun, M.; Meadows, J.R.S.; Zamani, N.; Johnson, J.; Sundström, G.; Cook, A.; FitzGerald, M.G.; Swofford, R.; et al. An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts. *PLoS ONE* **2014**, *9*, e91172. [CrossRef]
21. EMBL-EBI Ensembl gene annotation project (e!68): Canis lupus familiaris. Available online: https://useast.ensembl.org/info/genome/genebuild/2012_07_dog_genebuild.pdf (accessed on 1 May 2019).
22. Cunningham, F.; Achuthan, P.; Akanni, W.; Allen, J.; Amode, M.R.; Armean, I.M.; Bennett, R.; Bhai, J.; Billis, K.; Boddu, S.; et al. Ensembl 2019. *Nucleic Acids Res.* **2019**, *47*, D745–D751. [CrossRef] [PubMed]
23. Raney, B.J.; Dreszer, T.R.; Barber, G.P.; Clawson, H.; Fujita, P.A.; Wang, T.; Nguyen, N.; Paten, B.; Zweig, A.S.; Karolchik, D.; et al. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics* **2014**, *30*, 1003–1005. [CrossRef] [PubMed]
24. Kent, W.J. BLAT—The BLAST-like alignment tool. *Genome Res.* **2002**. [CrossRef] [PubMed]
25. Le Béguec, C.; Wucher, V.; Lagoutte, L.; Cadieu, E.; Botherel, N.; Hédan, B.; De Brito, C.; Guillory, A.-S.; André, C.; Derrien, T.; et al. Characterisation and functional predictions of canine long non-coding RNAs. *Sci. Rep.* **2018**, *8*, 13444. [CrossRef]
26. Wucher, V.; Legeai, F.; Hédan, B.; Rizk, G.; Lagoutte, L.; Leeb, T.; Jagannathan, V.; Cadieu, E.; David, A.; Lohi, H.; et al. FEELnc: A tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Res.* **2017**, *45*, e57. [CrossRef] [PubMed]

27. Haeussler, M.; Zweig, A.S.; Tyner, C.; Speir, M.L.; Rosenbloom, K.R.; Raney, B.J.; Lee, C.M.; Lee, B.T.; Hinrichs, A.S.; Gonzalez, J.N.; et al. The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res.* **2019**, *47*, D853–D858. [CrossRef]
28. Lindblad-Toh, K.; Garber, M.; Zuk, O.; Lin, M.F.; Parker, B.J.; Washietl, S.; Kheradpour, P.; Ernst, J.; Jordan, G.; Mauceli, E.; et al. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **2011**, *478*, 476–482. [CrossRef]
29. Rands, C.M.; Meader, S.; Ponting, C.P.; Lunter, G. 8.2% of the Human genome is constrained: Variation in rates of turnover across functional element classes in the human lineage. *PLoS Genet.* **2014**, *10*, e1004525. [CrossRef]
30. Villar, D.; Berthelot, C.; Aldridge, S.; Rayner, T.F.; Lukk, M.; Pignatelli, M.; Park, T.J.; Deaville, R.; Erichsen, J.T.; Jasinska, A.J.; et al. Enhancer evolution across 20 mammalian species. *Cell* **2015**, *160*, 554–566. [CrossRef]
31. Megquier, K.; Turner-Maier, J.; Swofford, R.; Kim, J.-H.; Sarver, A.L.; Wang, C.; Sakthikumar, S.; Johnson, J.; Koltookian, M.; Lewellen, M.; et al. Genomic analysis reveals shared genes and pathways in human and canine angiosarcoma. *BioRxiv* **2019**, 570879. [CrossRef]
32. Tonomura, N.; Elvers, I.; Thomas, R.; Megquier, K.; Turner-Maier, J.; Howald, C.; Sarver, A.L.; Swofford, R.; Frantz, A.M.; Ito, D.; et al. Genome-wide association study identifies shared risk loci common to two malignancies in golden retrievers. *PLoS Genet.* **2015**, *11*, e1004922. [CrossRef] [PubMed]
33. Hendricks, W.P.D.; Zismann, V.; Sivaprakasam, K.; Legendre, C.; Poorman, K.; Tembe, W.; Perdigones, N.; Kiefer, J.; Liang, W.; DeLuca, V.; et al. Somatic inactivating PTPRJ mutations and dysregulated pathways identified in canine malignant melanoma by integrated comparative genomic analysis. *PLoS Genet.* **2018**, *14*, e1007589. [CrossRef] [PubMed]
34. Ostrander, E.A.; Dreger, D.L.; Evans, J.M. Canine Cancer Genomics: Lessons for Canine and Human Health. *Annu. Rev. Anim. Biosci.* **2019**, *7*, 449–472. [CrossRef] [PubMed]
35. Elvers, I.; Turner-Maier, J.; Swofford, R.; Koltookian, M.; Johnson, J.; Stewart, C.; Zhang, C.-Z.; Schumacher, S.E.; Beroukhim, R.; Rosenberg, M.; et al. Exome sequencing of lymphomas from three dog breeds reveals somatic mutation patterns reflecting genetic background. *Genome Res.* **2015**, *25*, 1634–1645. [CrossRef] [PubMed]
36. Sakthikumar, S.; Elvers, I.; Kim, J.; Arendt, M.L.; Thomas, R.; Turner-Maier, J.; Swofford, R.; Johnson, J.; Schumacher, S.E.; Alföldi, J.; et al. SETD2 Is Recurrently Mutated in Whole-Exome Sequenced Canine Osteosarcoma. *Cancer Res.* **2018**, *78*, 3421–3431. [CrossRef] [PubMed]
37. Shearin, A.L.; Hedan, B.; Cadieu, E.; Erich, S.A.; Schmidt, E.V.; Faden, D.L.; Cullen, J.; Abadie, J.; Kwon, E.M.; Gröne, A.; et al. The MTAP-CDKN2A locus confers susceptibility to a naturally occurring canine cancer. *Cancer Epidemiol. Biomarkers Prev.* **2012**, *21*, 1019–1027. [CrossRef] [PubMed]
38. Tang, R.; Noh, H.J.; Wang, D.; Sigurdsson, S.; Swofford, R.; Perloski, M.; Duxbury, M.; Patterson, E.E.; Albright, J.; Castelhano, M.; et al. Candidate genes and functional noncoding variants identified in a canine model of obsessive-compulsive disorder. *Genome Biol.* **2014**, *15*, R25. [CrossRef]
39. Sarviaho, R.; Hakosalo, O.; Tiira, K.; Sulkama, S.; Salmela, E.; Hytönen, M.K.; Sillanpää, M.J.; Lohi, H. Two novel genomic regions associated with fearfulness in dogs overlap human neuropsychiatric loci. *Transl. Psychiatry* **2019**, *9*, 18. [CrossRef]
40. Tengvall, K.; Kierczak, M.; Bergvall, K.; Olsson, M.; Frankowiack, M.; Farias, F.H.G.; Pielberg, G.; Carlborg, Ö.; Leeb, T.; Andersson, G.; et al. Genome-wide analysis in German shepherd dogs reveals association of a locus on CFA 27 with atopic dermatitis. *PLoS Genet.* **2013**, *9*, e1003475. [CrossRef]
41. Vieira, N.M.; Elvers, I.; Alexander, M.S.; Moreira, Y.B.; Eran, A.; Gomes, J.P.; Marshall, J.L.; Karlsson, E.K.; Verjovski-Almeida, S.; Lindblad-Toh, K.; et al. Jagged 1 Rescues the Duchenne Muscular Dystrophy Phenotype. *Cell* **2015**, *163*, 1204–1213. [CrossRef]
42. Hayward, J.J.; Castelhano, M.G.; Oliveira, K.C.; Corey, E.; Balkman, C.; Baxter, T.L.; Casal, M.L.; Center, S.A.; Fang, M.; Garrison, S.J.; et al. Complex disease and phenotype mapping in the domestic dog. *Nat. Commun.* **2016**, *7*, 10460. [CrossRef] [PubMed]
43. Raffan, E.; Dennis, R.J.; O'Donovan, C.J.; Becker, J.M.; Scott, R.A.; Smith, S.P.; Withers, D.J.; Wood, C.J.; Conci, E.; Clements, D.N.; et al. A Deletion in the Canine POMC Gene Is Associated with Weight and Appetite in Obesity-Prone Labrador Retriever Dogs. *Cell Metab.* **2016**, *23*, 893–900. [CrossRef]
44. Becker, D.; Minor, K.M.; Letko, A.; Ekenstedt, K.J.; Jagannathan, V.; Leeb, T.; Shelton, G.D.; Mickelson, J.R.; Drögemüller, C. A GJA9 frameshift variant is associated with polyneuropathy in Leonberger dogs. *BMC Genomics* **2017**, *18*, 662. [CrossRef] [PubMed]

45. Lin, L.; Faraco, J.; Li, R.; Kadotani, H.; Rogers, W.; Lin, X.; Qiu, X.; de Jong, P.J.; Nishino, S.; Mignot, E. The sleep disorder canine narcolepsy is caused by a mutation in the hypocretin (orexin) receptor 2 gene. *Cell* **1999**, *98*, 365–376. [CrossRef]
46. Acland, G.M.; Aguirre, G.D.; Ray, J.; Zhang, Q.; Aleman, T.S.; Cideciyan, A.V.; Pearce-Kelling, S.E.; Anand, V.; Zeng, Y.; Maguire, A.M.; et al. Gene therapy restores vision in a canine model of childhood blindness. *Nat. Genet.* **2001**, *28*, 92–95. [CrossRef] [PubMed]
47. Meyers-Wallen, V.N.; Boyko, A.R.; Danko, C.G.; Grenier, J.K.; Mezey, J.G.; Hayward, J.J.; Shannon, L.M.; Gao, C.; Shafquat, A.; Rice, E.J.; et al. XX Disorder of Sex Development is associated with an insertion on chromosome 9 and downregulation of RSPO1 in dogs (Canis lupus familiaris). *PLoS ONE* **2017**, *12*, e0186331. [CrossRef]
48. Meyers-Wallen, V. Canine Embryonic Atlas at Cornell University. Available online: https://www.vet.cornell.edu/canine-atlas (accessed on 1 May 2019).
49. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [CrossRef]
50. DePristo, M.A.; Banks, E.; Poplin, R.; Garimella, K.V.; Maguire, J.R.; Hartl, C.; Philippakis, A.A.; del Angel, G.; Rivas, M.A.; Hanna, M.; et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **2011**, *43*, 491–498. [CrossRef]
51. McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **2010**, *20*, 1297–1303. [CrossRef]
52. Purcell, S.; Chang, C. PLINK2 (v1.90b6.9). Available online: www.cog-genomics.org/plink/2.0/ (accessed on 1 May 2019).
53. Chang, C.C.; Chow, C.C.; Tellier, L.C.; Vattikuti, S.; Purcell, S.M.; Lee, J.J. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* **2015**, *4*, 7. [CrossRef]
54. Cingolani, P.; Platts, A.; Wang, L.L.; Coon, M.; Nguyen, T.; Wang, L.; Land, S.J.; Lu, X.; Ruden, D.M. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **2012**, *6*, 80–92. [CrossRef]
55. ENCODE Consortium Current ENCODE Experiment Guidelines. Available online: https://www.encodeproject.org/about/experiment-guidelines/ (accessed on 28 May 2019).
56. Pertea, M.; Kim, D.; Pertea, G.M.; Leek, J.T.; Salzberg, S.L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **2016**, *11*, 1650–1667. [CrossRef]
57. Kim, D.; Langmead, B.; Salzberg, S.L. HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* **2015**, *12*, 357–360. [CrossRef] [PubMed]
58. Pertea, M.; Pertea, G.M.; Antonescu, C.M.; Chang, T.-C.; Mendell, J.T.; Salzberg, S.L. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **2015**, *33*, 290–295. [CrossRef] [PubMed]
59. Robinson, M.D.; Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **2010**, *11*, R25. [CrossRef]
60. Robinson, M.D.; McCarthy, D.J.; Smyth, G.K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **2010**, *26*, 139–140. [CrossRef] [PubMed]
61. McCarthy, D.J.; Chen, Y.; Smyth, G.K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **2012**, *40*, 4288–4297. [CrossRef]
62. Law, C.W.; Alhamdoosh, M.; Su, S.; Smyth, G.K.; Ritchie, M.E. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. *F1000Res.* **2016**, *5*. [CrossRef]
63. R: The R Project for Statistical Computing. Available online: https://www.R-project.org/ (accessed on 28 May 2019).
64. Eisenberg, E.; Levanon, E.Y. Human housekeeping genes, revisited. *Trends Genet.* **2013**, *29*, 569–574. [CrossRef] [PubMed]
65. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15*, 550. [CrossRef]
66. Kinsella, R.J.; Kähäri, A.; Haider, S.; Zamora, J.; Proctor, G.; Spudich, G.; Almeida-King, J.; Staines, D.; Derwent, P.; Kerhornou, A.; et al. Ensembl BioMarts: A hub for data retrieval across taxonomic space. *Database* **2011**, *2011*, bar030. [CrossRef]
67. Jones, E.; Oliphant, T.; Peterson, P. SciPy: Open Source Scientific Tools for Python. Available online: http://www.scipy.org/ (accessed on 1 May 2019).

68. GffCompare: Program for Processing GTF/GFF Files. Available online: https://ccb.jhu.edu/software/stringtie/gffcompare.shtml (accessed on 1 May 2019).
69. Kent, W.J.; Sugnet, C.W.; Furey, T.S.; Roskin, K.M.; Pringle, T.H.; Zahler, A.M.; Haussler, D. The human genome browser at UCSC. *Genome Res.* **2002**, *12*, 996–1006. [CrossRef]
70. Pruitt, K.D.; Brown, G.R.; Hiatt, S.M.; Thibaud-Nissen, F.; Astashyn, A.; Ermolaeva, O.; Farrell, C.M.; Hart, J.; Landrum, M.J.; McGarvey, K.M.; et al. RefSeq: An update on mammalian reference sequences. *Nucleic Acids Res.* **2014**, *42*, D756–D763. [CrossRef]
71. Pruitt, K.D. NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **2004**, *33*, D501–D504. [CrossRef]
72. Young, M.D.; Wakefield, M.J.; Smyth, G.K.; Oshlack, A. Gene ontology analysis for RNA-seq: Accounting for selection bias. *Genome Biol.* **2010**, *11*, R14. [CrossRef]
73. Huber, W.; Carey, V.J.; Gentleman, R.; Anders, S.; Carlson, M.; Carvalho, B.S.; Bravo, H.C.; Davis, S.; Gatto, L.; Girke, T.; et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **2015**, *12*, 115–121. [CrossRef]
74. Quinlan, A.R.; Hall, I.M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **2010**, *26*, 841–842. [CrossRef]
75. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: architecture and applications. *BMC Bioinf.* **2009**, *10*, 421. [CrossRef]
76. O'Leary, N.A.; Wright, M.W.; Brister, J.R.; Ciufo, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D.; et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **2016**, *44*, d733–d745. [CrossRef]
77. Gaspar, J.M. ATAC-seq Guidelines. Available online: https://informatics.fas.harvard.edu/atac-seq-guidelines-old-version.html (accessed on 1 May 2019).
78. ataqv: A toolkit for QC and visualization of ATAC-seq results. Available online: https://github.com/ParkerLab/ataqv (accessed on 1 May 2019).
79. Yu, G.; Wang, L.-G.; He, Q.-Y. ChIPseeker: An R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **2015**, *31*, 2382–2383. [CrossRef]
80. Aken, B.L.; Ayling, S.; Barrell, D.; Clarke, L.; Curwen, V.; Fairley, S.; Fernandez Banet, J.; Billis, K.; García Girón, C.; Hourlier, T.; et al. The Ensembl gene annotation system. *Database* **2016**. [CrossRef]
81. Roadmap Epigenomics Project Visual Browser. Available online: http://www.roadmapepigenomics.org/data/visualbrowser/adult (accessed on 1 May 2019).
82. Leinonen, R.; Sugawara, H.; Shumway, M. International Nucleotide Sequence Database Collaboration The sequence read archive. *Nucleic Acids Res.* **2011**, *39*, D19–D21. [CrossRef]
83. Kang, N.; Koo, J. Olfactory receptors in non-chemosensory tissues. *BMB Rep.* **2012**, *45*, 612–622. [CrossRef]
84. Li, F. Taste perception: From the tongue to the testis. *Mol. Hum. Reprod.* **2013**, *19*, 349–360. [CrossRef]
85. Mooney, M.; Bond, J.; Monks, N.; Eugster, E.; Cherba, D.; Berlinski, P.; Kamerling, S.; Marotti, K.; Simpson, H.; Rusk, T.; et al. Comparative RNA-Seq and microarray analysis of gene expression changes in B-cell lymphomas of Canis familiaris. *PLoS ONE* **2013**, *8*, e61088. [CrossRef]
86. Maeda, S.; Tomiyasu, H.; Tsuboi, M.; Inoue, A.; Ishihara, G.; Uchikai, T.; Chambers, J.K.; Uchida, K.; Yonezawa, T.; Matsuki, N. Comprehensive gene expression analysis of canine invasive urothelial bladder carcinoma by RNA-Seq. *BMC Cancer* **2018**, *18*, 472. [CrossRef]
87. Gorden, B.H.; Kim, J.-H.; Sarver, A.L.; Frantz, A.M.; Breen, M.; Lindblad-Toh, K.; O'Brien, T.D.; Sharkey, L.C.; Modiano, J.F.; Dickerson, E.B. Identification of three molecular and functional subtypes in canine hemangiosarcoma through gene expression profiling and progenitor cell characterization. *Am. J. Pathol.* **2014**, *184*, 985–995. [CrossRef]
88. Scott, M.C.; Temiz, N.A.; Sarver, A.E.; LaRue, R.S.; Rathe, S.K.; Varshney, J.; Wolf, N.K.; Moriarity, B.S.; O'Brien, T.D.; Spector, L.G.; et al. Comparative Transcriptome Analysis Quantifies Immune Cell Transcript Levels, Metastatic Progression, and Survival in Osteosarcoma. *Cancer Res.* **2018**, *78*, 326–337. [CrossRef]
89. Schurch, N.J.; Schofield, P.; Gierliński, M.; Cole, C.; Sherstnev, A.; Singh, V.; Wrobel, N.; Gharbi, K.; Simpson, G.G.; Owen-Hughes, T.; et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA* **2016**, *22*, 839–851. [CrossRef]
90. Ramsköld, D.; Wang, E.T.; Burge, C.B.; Sandberg, R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.* **2009**, *5*, e1000598. [CrossRef]

91. Danielsson, A.; Pontén, F.; Fagerberg, L.; Hallström, B.M.; Schwenk, J.M.; Uhlén, M.; Korsgren, O.; Lindskog, C. The Human Pancreas Proteome Defined by Transcriptomics and Antibody-Based Profiling. *PLoS ONE* **2014**, *9*, e115421. [CrossRef]
92. Melé, M.; Ferreira, P.G.; Reverter, F.; DeLuca, D.S.; Monlong, J.; Sammeth, M.; Young, T.R.; Goldmann, J.M.; Pervouchine, D.D.; Sullivan, T.J.; et al. Human genomics. The human transcriptome across tissues and individuals. *Science* **2015**, *348*, 660–665. [CrossRef]
93. Wright, F.A.; Sullivan, P.F.; Brooks, A.I.; Zou, F.; Sun, W.; Xia, K.; Madar, V.; Jansen, R.; Chung, W.; Zhou, Y.-H.; et al. Heritability and genomics of gene expression in peripheral blood. *Nat. Genet.* **2014**, *46*, 430–437. [CrossRef]
94. Zeisel, A.; Muñoz-Manchado, A.B.; Codeluppi, S.; Lönnerberg, P.; La Manno, G.; Juréus, A.; Marques, S.; Munguba, H.; He, L.; Betsholtz, C.; et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **2015**, *347*, 1138–1142. [CrossRef]
95. GTEx Consortium Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **2015**, *348*, 648–660. [CrossRef]
96. McKenzie, A.T.; Wang, M.; Hauberg, M.E.; Fullard, J.F.; Kozlenkov, A.; Keenan, A.; Hurd, Y.L.; Dracheva, S.; Casaccia, P.; Roussos, P.; et al. Brain Cell Type Specific Gene Expression and Co-expression Network Architectures. *Sci. Rep.* **2018**, *8*, 8868. [CrossRef]
97. Rackley, C.R.; Stripp, B.R. Building and maintaining the epithelium of the lung. *J. Clin. Investig.* **2012**, *122*, 2724–2730. [CrossRef]
98. Mallo, M. Reassessing the Role of Hox Genes during Vertebrate Development and Evolution. *Trends Genet.* **2018**, *34*, 209–217. [CrossRef]
99. Di-Poï, N.; Koch, U.; Radtke, F.; Duboule, D. Additive and global functions of HoxA cluster genes in mesoderm derivatives. *Dev. Biol.* **2010**, *341*, 488–498. [CrossRef]
100. Stelzer, G.; Rosen, N.; Plaschkes, I.; Zimmerman, S.; Twik, M.; Fishilevich, S.; Stein, T.I.; Nudel, R.; Lieder, I.; Mazor, Y.; et al. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr. Protoc. Bioinf.* **2016**, *54*, 1.30.1–1.30.33.
101. Buenrostro, J.D.; Wu, B.; Chang, H.Y.; Greenleaf, W.J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr. Protoc. Mol. Biol.* **2015**, *109*, 21–29. [PubMed]
102. Buenrostro, J.D.; Giresi, P.G.; Zaba, L.C.; Chang, H.Y.; Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **2013**, *10*, 1213–1218. [CrossRef] [PubMed]
103. Ackermann, A.M.; Wang, Z.; Schug, J.; Naji, A.; Kaestner, K.H. Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes. *Mol. Metab.* **2016**, *5*, 233–244. [CrossRef] [PubMed]
104. Wang, E.T.; Sandberg, R.; Luo, S.; Khrebtukova, I.; Zhang, L.; Mayr, C.; Kingsmore, S.F.; Schroth, G.P.; Burge, C.B. Alternative isoform regulation in human tissue transcriptomes. *Nature* **2008**, *456*, 470–476. [CrossRef] [PubMed]
105. Cantile, M.; Franco, R.; Tschan, A.; Baumhoer, D.; Zlobec, I.; Schiavo, G.; Forte, I.; Bihl, M.; Liguori, G.; Botti, G.; et al. HOX D13 expression across 79 tumor tissue types. *Int. J. Cancer* **2009**, *125*, 1532–1541. [CrossRef]
106. Song, I.-S.; Oh, N.S.; Kim, H.T.; Ha, G.-H.; Jeong, S.-Y.; Kim, J.-M.; Kim, D.-I.; Yoo, H.-S.; Kim, C.-H.; et al. Human ZNF312b Promotes the Progression of Gastric Cancer by Transcriptional Activation of the K-ras Gene. *Cancer Res.* **2009**, *69*, 3131–3139. [CrossRef]
107. Inoue, K.; Tsubamoto, H.; Isono-Nakata, R.; Sakata, K.; Nakagomi, N. Itraconazole treatment of primary malignant melanoma of the vagina evaluated using positron emission tomography and tissue cDNA microarray: A case report. *BMC Cancer* **2018**, *18*. [CrossRef]
108. Wang, M.; Li, X.; Zhang, J.; Yang, Q.; Chen, W.; Jin, W.; Huang, Y.-R.; Yang, R.; Gao, W.-Q. AHNAK2 is a Novel Prognostic Marker and Oncogenic Protein for Clear Cell Renal Cell Carcinoma. *Theranostics* **2017**, *7*, 1100–1113. [CrossRef]
109. Derrien, T.; Johnson, R.; Bussotti, G.; Tanzer, A.; Djebali, S.; Tilgner, H.; Guernec, G.; Martin, D.; Merkel, A.; Knowles, D.G.; et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res.* **2012**, *22*, 1775–1789. [CrossRef]
110. Mattick, J.S.; Rinn, J.L. Discovery and annotation of long noncoding RNAs. *Nat. Struct. Mol. Biol.* **2015**, *22*, 5–7. [CrossRef]

111. Schoenfelder, S.; Fraser, P. Long-range enhancer–promoter contacts in gene expression control. *Nat. Rev. Genet.* **2019**, *1*. [CrossRef]
112. Viñuela, A.; Brown, A.A.; Buil, A.; Tsai, P.-C.; Davies, M.N.; Bell, J.T.; Dermitzakis, E.T.; Spector, T.D.; Small, K.S. Age-dependent changes in mean and variance of gene expression across tissues in a twin cohort. *Hum. Mol. Genet.* **2018**, *27*, 732–741. [CrossRef]
113. Chaker, L.; Cappola, A.R.; Mooijaart, S.P.; Peeters, R.P. Clinical aspects of thyroid function during ageing. *Lancet Diabetes Endocrinol.* **2018**, *6*, 733–742. [CrossRef]
114. Diagnostic Center for Population & Animal Health. *Canine Thyroid Registry*; Michigan State University: East Lansing, MI, USA, 2010.
115. Consortium, G. GTEx Consortium Genetic effects on gene expression across human tissues. *Nature* **2017**, *550*, 204–213. [CrossRef]
116. Werber, M.; Wittler, L.; Timmermann, B.; Grote, P.; Herrmann, B.G. The tissue-specific transcriptomic landscape of the mid-gestational mouse embryo. *Development* **2014**, *141*, 2325–2330. [CrossRef]
117. Park, S.-J.; Shirahige, K.; Ohsugi, M.; Nakai, K. DBTMEE: A database of transcriptome in mouse early embryos. *Nucleic Acids Res.* **2015**, *43*, D771–D776. [CrossRef]
118. Sahakyan, A.; Plath, K. Transcriptome Encyclopedia of Early Human Development. *Cell* **2016**, *165*, 777–779. [CrossRef]