nature portfolio

Peer Review File

The DoGA Consortium expression atlas of promoters and genes in 100 canine tissues

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

Hörtenhuber at al., report the development of the Dog Genome Annotation (DoGA) canine tissue biobank (132 tissue each sourced from 35 samples). A large subset of biobank, 596 samples representing 100 tissues, were subsequently processed via a STST2-seq pipeline to produce a repository of XX.

The authors attempt to demonstrate the utility of the repository by intersecting their data with OMIA (Online Mendelian Inheritance in Animals) and the association results from two recent publications aimed at resolving the contribution of breed membership to phenotype (i.e., Dutrow et al., DOI:10.1016/j.cell.2022.11.003 and Morrill et al., DOI:10.1126/science.abk0639). However, the results of these intersections are not put into context for the reader, rather this result is left for the reader to download and use for themselves. What was the point of these experiments? What does the author want the reader to know? The original authors of each paper and database have already characterised the location of their reported variants, so what is new?

While this single-cell transcriptional atlas will be of great use to the community, key methods and details are currently missing from the manuscript, making it not possible to assess its current quality or utility. Additional major comments are listed below.

General
• It is difficult to distinguish between the samples taken for the biobank and those subsequently used in the STRT experiment. A supplementary table with metadata for this paper is needed. Readers should not have to sift through the DoGA databases for this information.

• As individuals develop from an embryonic to adult stage, transcript usage will change. The authors combined the 25 day old or 30 day old embryos. Was this appropriate? What are the gross developmental changes at these stages which will be reflected in promoter usage? Could the authors comment of the distribution of embryonic signals in the UMAP?

• The authors perform their analyses for both CanFam3.1 and UU_Cfam_GSD_1.0. A general summary of data differences obtained from using both references is warranted. What were the gains from using UU_Cfam_GSD_1.0, were there features that made CanFam3.1 more suitable for some analyses?

• In the presentation of tissue enrichment of promoters, could this not be visualised or written to a table so that readers could explore this data for themselves? How do the authors expect that these values will change with the addition of biological replicates, additional tissues, and for additional lifestages?

• Many terms are not qualified, e.g., “promoter resolution”, “novel” (in the context of what?), “robustly expressed”, organ versus organ system (where in the paper are these terms the same/ different?), “reliably expressed”

Missing a section describing limitations, e.g.,
• First exons and promotors in dogs are extremely GC rich, (See e.g., Axelsson et al., DOI:10.1101/gr.124123.111). Are there any concerns that PCR biases underrepresent GC-rich templates?
• No statistics were given for the technical replicates. How can the reader assess the effects of nonlinear PCR amplification, batch effects, or other technical artifact on the reproducibility of the data?
• Table 2 in the manuscript lists whole genome sequencing data, but the generation of this data or its subsequent use it never mentioned. Could this have been used to address potential amplification errors in STRT-seq or have been used to look for allele-specific expression?

FROM THE MANUSCRIPT SECTIONS
ABSTRACT.

• All released dog references have included functional annotation. While not “exhaustive”, this could not be classed as missing.

RESULTS
• Any comment on the number of reads (Gb or Tb of data) generated and what fraction could be mapped to each reference?
• Any comment of the number of mRNA molecules per well on average in each tissue type? This could be of interest to other researches interested in using the technology for their own experiments.
• Any comments on the sequencing result differences between embryonic and “adult” tissues? Did embryonic tissues require additional lanes of sequencing to obtain sufficient data?
• What effect did the STRT2 library prep have on expression normalisation and results?
• Were there any technical replicates. If yes, where are those results. If not, why were they not needed? How much variation is expected from single cell “true” biology vs technical artefacts?
• How were the results of the UMAP interpreted? What determined a cluster as being indicative of muscle, cardiac, or CNS? This is never described. What is each dot in this plot?
• Unsupervised clustering is not described in methods.
• Were there differences between sexes, ages, breeds etc. What was being pooled?
• What does “reliably expressed” mean? How was this qualified?
• FANTOM5 promoter annotations are not mentioned specifically in the methods, but are referenced in the results. How were the methods used here similar or different to FANTOM5?
• What determines a “robust” set of regions? What is a region?
• Table 1. What makes are promoter “robust” vs “comprehensive”.
• How is a promoter classed as “novel”
• How is the breakdown of “primary”, “alternative”, “novel” by tissue? Can we learn something about biology?
• Is there an expression ratio difference required to define primary versus alternative? If the expression levels were similar, could this not point to overlapping but separate gene transcripts rather than alternative promoters?
• What are the promoter count difference for the authors different classed for CanFam3.1 versus UU_Cfam_GSD_1.0? Only the latter is reported.
• What is “close to” in base pairs?
• “Enriched” measured how?
• The ARAC and chromatin data was not available for all tissues sampled in DoGA. How does Table 1 account for this? Are the percentages for when there was overlap? Or is it a fraction of the whole?
• The first sentence of the “Tissue enrichment of genes” section is circular. What are the authors trying to say?
• The human Protein Atlas is not mentioned in the methods. How int eh results are the approaches similar? What is different? Does “ubiquitous” here meaning all tissues? All adult tissues?
• What is the comparison that is being made to the human data? The count of promoter regions? The tissues? The genes? This section is hard to follow and needs to be restructured for clarity.
• What is meant by an annotated gene symbol for the comparison between human and dog? For which annotations?
• Supplementary figure 2 is the stations, it is not related to OMIA.
• In the discussion of the OMIA overlap, how important is the use of alternative transcripts from the same promoter for tissue usage? Does that impact the interpretation? Can other annotation help to resolve this issue?
• There are no table results or illustrations to help the reader interpret the results from the lineage and behaviour sections. What is new here? Could lineage and behaviour SNP overlap alternative promoters from the primary transcript? How do the authors interpret these results? Why did the authors conduct these intersections?
• How is the case-example for cardia tissue different to the human protein atlas section?

DISCUSSION
Extremely limited.

METHODS
Postmortem examination…

• A summary is needed before the authors explain the fine details of BioBank collection. A "simple" metadata table showing which samples (age, sex, breed/dog/wolf/embryo/cell line) where taken for which tissue type is needed. This could then be colour or highlighted to show which samples were taken forward for STRT2-Seq. Supplementary Table 1 is mentioned in the results, but it does not contain sufficient information.
• What is "routine tissue processing"? A routine in one location is not the same for another. If it is routine for this group, cite an appropriate reference.
• Same applies to "no histological changes". The authors mention some specific stains and blood measures, but it is not clear what they are examining in each case. A sheet stating the histopathology and/ or blood measures for each tissue would be appreciated.
• Standard terminology from ECVP, for which year? Reference?

RNA Isolation
• What modifications were necessary?
• What kinds of detailed information is available for each tissue?

STRT2 library prep
• How was the protocol modified? Authors should give at least a brief description of the method, e.g., input RNA, PCR cycling, sequencing technology, pooling strategy and quantification, Spike-In of what and how much? The reader should not have to read Ezer et al., (DOI: 10.1016/j.xpro.2021.100995) to get this information.

STRT2 data processing
• Which software versions are used? Utility is not consistent between versions.
• Provide the GenBank assembly accession for each reference.
• What is the QC process for "good" or "bad" reads?
• DNAse1 treatment is not perfect. Did the authors quantify background from gDNA for example? If not, why? If yes, how?
• How did the authors deal with reads that mapped to repeats? Are these masked? What could these reads represent? Are they of interest biologically?
• Did the authors make use of strand information in gene processing?

STRT clusters
• Which annotation for which reference is used? canFam3.1 contains many gaps, as noted by the authors, whereas UU_Cfam_GSD_1.0 is long-read assembly and so is more complete, especially at promoters and first exons. Each reference has an NCBI annotation, plus canine tissue specific short- or long-read annotations. Clearly, the gene annotation used will impact the result.

External validation
• What ChiP-seq data was accessed? The reader should not have to access Roller et al., to see for which chromatin marks (DOI: 10.1186/s13059-021-02260-y).

Promoter validation
• Was this section performed only in relations to CanFam3.1? If yes, why? Did the five genes also have first exons in UU_Cfam_GSD_1.0?
• What is "robustly expressed". Can the authors this place this in context of expression across tissues?

Expression atlas
Which version of OMIA was access when? OMIA is CanFam3.1 referenced, how were these positions converted to UU_Cfam_GSD_1.0 reference?

Reviewer #2 (Remarks to the Author):

The authors define a need for a promoter and gene expression map in dogs. The Dog Gene Atlas (DoGA) serves as a good resource for researchers in the fields of comparative genomics, veterinary medicine, and molecular biology.In this project, the authors have established a tissue biobank consisting of samples from 35 animals, including dogs, embryos, and wolves. The project has identified 100,000 promoters in 100 tissues, annotated promoter regions, and validated them using epigenomic data. The authors also highlight tissue enrichment of genes and promoters, their conservation between dogs and humans, and their possible applications for studying Mendelian diseases in animals, lineage- and behavior-specific SNPs, and cardiac tissue enrichment and conservation.
The manuscript lacks enough description and interpretation of their various data, also needs more figures to fully present their data. The authors should also provide biological meanings of their findings, including strength and weakness of the paper.
1. The DoGA STRT2-seq database (35 animals and 100 tissues) represents an impressive amount of data to profile tissue-specific gene expression and promoter maps. However, there are previous studies that have tried to establish a canine epigenomic database, such as publications by Barkbase (PMID: 31181663) and EpiC Dog (https://doi.org/10.1101/2022.07.22.501075). In the Barkbase project, bulk RNA-seq and ATAC-seq (on several tissues) data were produced on 27 tissues from more than four breeds. EpiC Dog made an epigenome map by performing bulk RNA-seq, five types of histone modification ChIP-seq, and MBD-seq on 11 tissues of beagle. Therefore, it is necessary for this manuscript to cite and mention previous efforts and to explain to the readers what are the differences and strengths of DoGA's project compared to previous projects.

2. There are more than five 5’-end RNA-sequencing methods. (https://www.nature.com/articles/s41592-018-0014-2/). Representatively, there are two projects (ENCODE (Encyclopedia of DNA Elements) project and Fantom (Functional Annotation of the Mammalian Genome) project) to create a comprehensive atlas of gene expression in both humans and mice using 5’-end RNA-sequencing technology, and both adopted the CAGE method. In contrast, both STRT2-seq and CAGE-seq methods are adopted in DoGA. The authors should mention in this manuscript what strategy was used to select both methods that are used to confirm similar results, and what parts were identified differentially using STRT2-seq.

3. The authors have created an unprecedented resource in the field of canine research through rigorous efforts, and through the establishment of a data coordination center and genome browser track hubs, it is effectively proposed as a resource for the scientific community and has an “open science” potential value to the whole community. However, it seems necessary to present more organized results about data quality evaluation and biological meaning.

4. It is necessary to evaluate the quality of STRT2-seq data of DoGA at each stage of pre-processing and processing and then visualize them so that data users can refer to it. And the authors should show the data quality level of DoGA in comparison with the standards generally accepted (e.g. ENCODE or FANTOM project).
5. It is recommended to refer to the following two papers for quality check and visualization.
1) Functional annotation of human long noncoding RNAs via molecular phenotyping. (PMID: 32718982) and 2) A promoter level mammalian expression atlas. (PMID: 24670764)

5. As mentioned in this publication (PMID: 29867192; Comprehensive comparative analysis of 5’-end RNA-sequencing methods), although bulk RNA-seq is a powerful approach for quantifying gene expression, discovering previously unknown transcripts, and determining splice isoforms, it is often difficult to reliably identify more than one TSS per gene. Therefore, it is believed that dog genome annotation can be improved with a previously unexpected resolution by using DoGA's STRT2-seq and CAGE datasets.
6. According to Figure 3 of BarkBase's publication (PMID: 31181663), there are significant gaps in the existing genome annotation.
7. As the authors noted in Table 1 in this manuscript, many novel promoters have been identified. However, this manuscript should build an improved genome annotation that can be used by various canine researchers who conduct follow-up studies by integrating analysis with various previously published bulk RNA-seq rather than simply reporting the promoter location.

8. The authors utilized an approach used in the Human Protein Atlas (HPA) to detect ubiquitous and tissue-enriched expression of the promoters and genes. Through this analysis, genes can be divided into a total of six categories (Tissue-Enriched, Group-Enriched, Tissue-Enhanced, Expressed-in-all, Mixed, and Not-Expressed). The factors that can change the results in this analysis are the minimum expression cutoff, the fold-change cutoff for expression comparison, and the number of tissues for "Group Enriched" classification. In the case of HPA (Uhlén et al. 2015), the minimum expression criterion is 1 (TPM or FPKM), the fold-change criterion is five-fold, and the number of tissues to distinguish group-enriched is 2-7 tissues. The authors should present an accurate criterion by presenting only the fold change criterion (3-fold higher) in the method.

8. It is recommended to visualize the results for the six categories and to present the differences from existing reports by conducting functional analysis on ubiquitous and tissue-enriched for reliability of the results.

9. In Figure 3A, results are not shown.

10. The authors presented several usage examples for the utilization of DoGA resources and some visual data. In addition, it was built so that detailed information can be checked in the "DoGA Expression Atlas". In these examples, the authors overlapped the promoters with lineage- and behavior-specific SNPs. Though overlapping statistics and visual figures are presented in the manuscript or "DoGA Expression Atlas", but it’s difficult to understand.
11. The authors should present the enrichment analysis of canine GWAS summary statistics related with lineage, morphological and behavioral traits for tissue-specific active promoters as a demonstration of resource utility for canine studies similar to the examples below.
- Figure 9, Integrative analysis of 111 reference human epigenomes. (PMID: 25693563)
- Figure S31, Ancestry-inclusive dog genomics challenges popular breed stereotypes (PMID: 35482869)
- Figure 3, An atlas of dynamic chromatin landscapes in mouse fetal development. (PMID: 32728240)

11. It is expected that DoGA's resources will be used for overall research such as dog genomics, transcriptomics, and epigenomics. Many of the web pages cannot be accessed yet, but it is expected that scientists will be able to better utilize DoGA's data by optimizing the access and speed of the mentioned web pages (for example, DoGA Expression Atlas; https://expression-atlas.doggenomeannotation.org/dogaatlas/).

Promoter Manuscript – Rebuttal letter

Reviewer #1 (Remarks to the Author):

Hörtenhuber at al., report the development of the Dog Genome Annotation (DoGA) canine tissue biobank (132 tissue each sourced from 35 samples). A large subset of biobank, 596 samples representing 100 tissues, were subsequently processed via a STST2-seq pipeline to produce a repository of XX.

The authors attempt to demonstrate the utility of the repository by intersecting their data with OMIA (Online Mendelian Inheritance in Animals) and the association results from two recent publications aimed at resolving the contribution of breed membership to phenotype (i.e., Dutrow et al., DOI:10.1016/j.cell.2022.11.003 and Morrill et al., DOI:10.1126/science.abk0639). However, the results of these intersections are not put into context for the reader, rather this result is left for the reader to download and use for themselves. What was the point of these experiments? What does the author want the reader to know? The original authors of each paper and database have already characterised the location of their reported variants, so what is new?

Response: We thank the reviewer for the comments. We have now substantially revised the Results and Discussion parts to better communicate and interpret the use case examples.

While this single-cell transcriptional atlas will be of great use to the community, key methods and details are currently missing from the manuscript, making it not possible to assess its current quality or utility. Additional major comments are listed below.

Response: We have now revised the Methods and added many details to better evaluate the quality of the study.

General

• It is difficult to distinguish between the samples taken for the biobank and those subsequently used in the STRT experiment. A supplementary table with metadata for this paper is needed. Readers should not have to sift through the DoGA databases for this information.

Response: We replaced the Supplementary Table 1 with an extensive project metadata file from DCC, describing the samples in detail, including sequencing quality control.

• As individuals develop from an embryonic to adult stage, transcript usage will change. The authors combined the 25 day old or 30 day old embryos. Was this appropriate?

Response: We did not combine the embryos but agree with the reviewer that it was not clearly expressed in the previous version. We have now performed additional analyses and

provide additional insights to developmental stage-specific promoter usage in embryos with new figures and revised texts.

What are the gross developmental changes at these stages which will be reflected in promoter usage?

Response: There are substantial significant changes and we have included descriptions and new analyses in the Results and Discussion with references.

Could the authors comment of the distribution of embryonic signals in the UMAP?

Response: They cluster very well with each other and distinct from the adult samples, but near the majority of the tissues as expected (see Figure 3a).

• The authors perform their analyses for both CanFam3.1 and UU_Cfam_GSD_1.0. A general summary of data differences obtained from using both references is warranted. What were the gains from using UU_Cfam_GSD_1.0, were there features that made CanFam3.1 more suitable for some analyses?

Response: We have now performed the analyses in both CanFma3 and Canfam4 and present the comparison data in Table 1. CanFam4 became available during the later stages of this study, and we had already performed promoter validation experiments in CanFam3. We believe it is beneficial to provide annotation in both references, since the community still uses also the older CanFam3 version.

• In the presentation of tissue enrichment of promoters, could this not be visualised or written to a table so that readers could explore this data for themselves?

Response: We added a Figure 3b and 3c to visualize the number of tissue-enriched promoters. Additionally, we added the full table of tissue-enriched genes and promoters from the expression atlas to the Supplementary Materials (Supp. Table 4).

How do the authors expect that these values will change with the addition of biological replicates, additional tissues, and for additional lifestages?

Response: We expect more promoters, although not significantly, to be found with more tissues. We included this in the Discussion now.

• Many terms are not qualified, e.g., “promoter resolution”, “novel” (in the context of what?), “robustly expressed”, organ versus organ system (where in the paper are these terms the same/ different?), “reliably expressed”

Response: We have clarified these terms in the manuscript.

Missing a section describing limitations, e.g.,

Response: We have added a comprehensive section in the end of the Discussion.

• First exons and promotors in dogs are extremely GC rich, (See e.g., Axelsson et al., DOI:10.1101/gr.124123.111). Are there any concerns that PCR biases underrepresent GC-rich templates? No statistics were given for the technical replicates. How can the reader assess the effects of nonlinear PCR amplification, batch effects, or other technical artifact on the reproducibility of the data?

Response: The STRT method was described in its original form 2011 as a single-cell RNA sequencing method (https://pubmed.ncbi.nlm.nih.gov/21543516/). It has been adopted later also as small-input 5' targeted RNAseq method that has been applied in several studies including both cultured cells and tissue samples, both from human, mouse and dog (e.g., https://pubmed.ncbi.nlm.nih.gov/27212086/, https://pubmed.ncbi.nlm.nih.gov/27355630/, https://pubmed.ncbi.nlm.nih.gov/30032427/, https://pubmed.ncbi.nlm.nih.gov/30974166/, https://pubmed.ncbi.nlm.nih.gov/31551465/, https://pubmed.ncbi.nlm.nih.gov/31619476/, https://pubmed.ncbi.nlm.nih.gov/32150541/, https://pubmed.ncbi.nlm.nih.gov/36339249/), modified to accommodate globin-rich samples, including validation of non-biased recovery of non-globin RNAs (https://pubmed.ncbi.nlm.nih.gov/27515369/), refined for library bias correction (https://pubmed.ncbi.nlm.nih.gov/31409293/), and described in detail and validated for dog RNA sequencing (https://pubmed.ncbi.nlm.nih.gov/34950881/).

• Table 2 in the manuscript lists whole genome sequencing data, but the generation of this data or its subsequent use it never mentioned.
Could this have been used to address potential amplification errors in STRT-seq or have been used to look for allele-specific expression?

Response: We have added text concerning the WGS data to the Methods, Results and Discussion. We agree that it can be used for allele-specific expression analyses in the future (Table 2).

FROM THE MANUSCRIPT SECTIONS

ABSTRACT.
• All released dog references have included functional annotation. While not "exhaustive", this could not be classed as missing.

Response: We have rephrased this statement.

RESULTS
• Any comment on the number of reads (Gb or Tb of data) generated and what fraction could be mapped to each reference?

Response: We have now prepared a QC report (Supplementary Table 2).

• Any comment of the number of mRNA molecules per well on average in each tissue type? This could be of interest to other researches interested in using the technology for their own experiments.

Response: We don't know the exact number of mRNA molecules per tissue type, however our approach using spike-in molecules enables this analysis for other researchers.

• Any comments on the sequencing result differences between embryonic and "adult" tissues? Did embryonic tissues require additional lanes of sequencing to obtain sufficient data?
Response: No difference in handling embryos and adults sequencing libraries was required.

• What effect did the STRT2 library prep have on expression normalisation and results?
Response: We found a high correlation between samples, so no different normalization was needed.

• Were there any technical replicates. If yes, where are those results. If not, why were they not needed? How much variation is expected from single cell "true" biology vs technical artefacts?
Response: Yes, we have technical replicates, see the metadata for details (Supplementary Table 1) and Suppl. Figure 3 for a correlation matrix between technical replicates, biological replicates and unrelated samples.

• How were the results of the UMAP interpreted? What determined a cluster as being indicative of muscle, cardiac, or CNS? This is never described. What is each dot in this plot?
Response: We clarified the UMAP plot, indicating that no clustering was performed, only coloring based on the metadata.

• Unsupervised clustering is not described in methods.
Response: We didn't use any clustering, the color coding in the UMAP plot is based on metadata related to the organ system. We have rephrased this in the results and figure caption.

• Were there differences between sexes, ages, breeds etc. What was being pooled?

Response: We did not pool any samples for this study. We added suppl. Figure 4 visualizing the effects of sex, age, breed in our samples.

• What does "reliably expressed" mean? How was this qualified?
Response: We added a definition in the text. It means highly expressed in three biological replicates.

• FANTOM5 promoter annotations are not mentioned specifically in the methods, but are referenced in the results. How were the methods used here similar or different to FANTOM5?
Response: We modeled our analysis with the same concepts and used the same approach for clustering promotor regions and define promoter activity.

• What determines a "robust" set of regions? What is a region?
Response: The robust set is a subset expressed across three replicates. A region means here a span of the genome. We clarified this in the manuscript.

• Table 1. What makes are promoter “robust” vs “comprehensive”.
Response: Promoters in the comprehensive set should have, similarly to the robust promoters, a high expression, but were not replicated across three replicates. We clarified it in the manuscript.

• How is a promoter classed as “novel”
Response: We clarified in the manuscript. Novel promoters are promoter regions with a distance higher than 500 bp from the nearest gene model.

• How is the breakdown of “primary”, “alternative”, “novel” by tissue? Can we learn something about biology?

Response:
We included this information in the table off tissue enriched promoters (supp. Table 4). It indicates alternative promoter usage on a tissue level. The novel promoters indicate a lack of gene annotations. Deeper understanding of the biological significance and role of the alternative and novel promoters in each tissue is beyond the scope of this manuscript.

• Is there an expression ratio difference required to define primary versus alternative? If the expression levels were similar, could this not point to overlapping but separate gene transcripts rather than alternative promoters?
Response: No expression ratio is required, it is just a ranking based on expression, where the promoter with the highest expression is assigned as primary (similar to FANTOM5). Our promoter clustering (paraclu) is based on a certain distance between their TSSs, which minimizes the risk of over-clustering potentially overlapping promoters. The clustering approach is applied on the expression of all samples summed up; therefore, we don’t distinguish between highly correlated promoters. Nevertheless, two identified clusters indicate two distinct transcription starting site, which indicates a distinct biological function. The possibility of false-positives cannot be completely excluded.

• What are the promoter count difference for the authors different classed for CanFam3.1 versus UU_Cfam_GSD_1.0? Only the latter is reported.
Response: We added this information to Table 1.

• What is “close to” in base pairs?
Response: We have clarified this in the manuscript.

• “Enriched” measured how?
Response: We have rephrased the text.

• The ATAC and chromatin data was not available for all tissues sampled in DoGA. How does Table 1 account for this? Are the percentages for when there was overlap? Or is it a fraction of the whole?
Response: We added a clarification in the text.

• The first sentence of the “Tissue enrichment of genes” section is circular. What are the authors trying to say?
Response: We rephrased this in the manuscript.

• The human Protein Atlas is not mentioned in the methods. How int eh results are the approaches similar? What is different? Does “ubiquitous” here meaning all tissues? All adult tissues?
Response: We added a reference to the HPA in the methods and clarified the similarities and differences. Compared to HPA we don’t identify “group-enriched” promoters but only “expressed-in-all”, meaning in all tissues (including embryos) and “tissue-enriched”.

• What is the comparison that is being made to the human data? The count of promoter regions? The tissues? The genes? This section is hard to follow and needs to be restructured for clarity.
Response: We have rewritten the paragraph for clarification.

• What is meant by an annotated gene symbol for the comparison between human and dog? For which annotations?
Response: We revised the text.

• Supplementary figure 2 is the stations, it is not related to OMIA.
Response: We corrected the reference to Figure 3.

• In the discussion of the OMIA overlap, how important is the use of alternative transcripts from the same promoter for tissue usage? Does that impact the interpretation? Can other annotation help to resolve this issue?
Response: Our data and approach does not allow to distinct alternative transcripts and their tissue usage, therefore we don’t know if it has any impact. Combining our data with data sets, which allow identification of alternative transcripts usage on a tissue level, for example barkbase.

• There are no table results or illustrations to help the reader interpret the results from the lineage and behaviour sections. What is new here? Could lineage and behaviour SNP overlap alternative promoters from the primary transcript? How do the authors interpret these results? Why did the authors conduct these intersections?
Response: We have revised the Results and Discussion to better discuss the results.

• How is the case-example for cardia tissue different to the human protein atlas section?
Response: The cardiac tissue case-example shows the conservation of alternative promoter usage, whereas the case-example with the HPA show conservation of tissue-enrichment on a gene level.

DISCUSSION
Extremely limited.
Response: We have now completely revised the discussion, including a section on limitations and more in depth comparison with previous studies.

METHODS
Postmortem examination…
A summary is needed before the authors explain the fine details of BioBank collection.
Response: We added a biobank section at the beginning of methods and revised the text.

A “simple” metadata table showing which samples (age, sex, breed/dog/wolf/embryo/cell line) where taken for which tissue type is needed. This could then be colour or highlighted to show which samples were taken forward for STRT2-Seq. Supplementary Table 1 is mentioned in the results, but it does not contain sufficient information.

Response: Original Suppl Table 1 is now replaced with an extensive metadata table and indicators of the samples used for STRT2-seq.

• What is “routine tissue processing”? A routine in one location is not the same for another. If it is routine for this group, cite an appropriate reference.
Response: We have revised the text.

• Same applies to “no histological changes”. The authors mention some specific stains and blood measures, but it is not clear what they are examining in each case. A sheet stating the histopathology and/ or blood measures for each tissue would be appreciated.
Response: We have clarified this in the Methods. Full reports of each animal are available in the DCC.

• Standard terminology from ECVP, for which year? Reference?
Response: We have clarified this.

RNA Isolation
• What modifications were necessary?
Response: We have clarified this.

• What kinds of detailed information is available for each tissue?
Response: This can be found in a new Suppl Table 1 we added.

STRT2 library prep
• How was the protocol modified? Authors should give at least a brief description of the method, e.g., input RNA, PCR cycling, sequencing technology, pooling strategy and quantification, Spike-In of what and how much? The reader should not have to read Ezer et al., (DOI: 10.1016/j.xpro.2021.100995) to get this information.
Response: We added a summary of the method to the manuscript.

STRT2 data processing
• Which software versions are used? Utility is not consistent between versions.
Response: We expanded the STRT2 data processing section including software versions.

• Provide the GenBank assembly accession for each reference.
Response: We added the accession numbers for the reference genomes in the methods section.

• What is the QC process for “good” or “bad” reads?

Response: The reads qualified by Illumina's base-caller were used as many as possible, however lower-quality reads among PCR duplicates suggested by the Unique Molecular Identifier were excluded by Picard's MarkDuplicates.

• DNAse1 treatment is not perfect. Did the authors quantify background from gDNA for example? If not, why? If yes, how?
Response: Quality of template RNA was analyzed with Tape Station and/or Bioanalyzer and no significant amount of gDNA was detected.

• How did the authors deal with reads that mapped to repeats? Are these masked? What could these reads represent? Are they of interest biologically?
Response:
Multimapped reads were removed during the data processing using Picard's MarkDuplicates. The re-analysis would be interesting when the dog pan-genomes (using the long-read DNA sequencer) are published.

• Did the authors make use of strand information in gene processing?
Response: Yes, our analysis is strand specific.

STRT clusters
• Which annotation for which reference is used? canFam3.1 contains many gaps, as noted by the authors, whereas UU_Cfam_GSD_1.0 is long-read assembly and so is more complete, especially at promoters and first exons. Each reference has an NCBI annotation, plus canine tissue specific short- or long-read annotations. Clearly, the gene annotation used will impact the result.
Response: We clarified the gene annotation used in the methods.

External validation
• What ChiP-seq data was accessed? The reader should not have to access Roller et al., to see for which chromatin marks (DOI: 10.1186/s13059-021-02260-y).
Response: We clarified which chromatin marks where targeted by the ChIP-seq data.

Promoter validation
• Was this section performed only in relations to CanFam3.1? If yes, why? Did the five genes also have first exons in UU_Cfam_GSD_1.0?
Response: All expect one was found in CanFam4. We have clarified this in the Supplementary material.

• What is "robustly expressed". Can the authors this place this in context of expression across tissues?
Response: It means highly expressed across three biological replicates (independent of their tissue type). We clarified this in the text.

Expression atlas
Which version of OMIA was access when? OMIA is CanFam3.1 referenced, how were these positions converted to UU_Cfam_GSD_1.0 reference?
Response: We clarified this in the Methods section.

Reviewer #2 (Remarks to the Author):

The authors define a need for a promoter and gene expression map in dogs. The Dog Gene Atlas (DoGA) serves as a good resource for researchers in the fields of comparative genomics, veterinary medicine, and molecular biology.In this project, the authors have established a tissue biobank consisting of samples from 35 animals, including dogs, embryos, and wolves. The project has identified 100,000 promoters in 100 tissues, annotated promoter regions, and validated them using epigenomic data. The authors also highlight tissue enrichment of genes and promoters, their conservation between dogs and humans, and their possible applications for studying Mendelian diseases in animals, lineage- and behavior-specific SNPs, and cardiac tissue enrichment and conservation.
The manuscript lacks enough description and interpretation of their various data, also needs more figures to fully present their data. The authors should also provide biological meanings of their findings, including strength and weakness of the paper.

Response: We thank the reviewer for encouraging comments. We have revised the manuscript with new analyses, results, methodological details, interpretation, and more comprehensive discussion, including a limitations section. The reviewers' comments have been very helpful, and we believe the manuscript is now much improved.

1. The DoGA STRT2-seq database (35 animals and 100 tissues) represents an impressive amount of data to profile tissue-specific gene expression and promoter maps. However, there are previous studies that have tried to establish a canine epigenomic database, such as publications by Barkbase (PMID: 31181663) and EpiC Dog (https://doi.org/10.1101/2022.07.22.501075). In the Barkbase project, bulk RNA-seq and ATAC-seq (on several tissues) data were produced on 27 tissues from more than four breeds. EpiC Dog made an epigenome map by performing bulk RNA-seq, five types of histone modification ChIP-seq, and MBD-seq on 11 tissues of beagle. Therefore, it is necessary for this manuscript to cite and mention previous efforts and to explain to the readers what are the differences and strengths of DoGA's project compared to previous projects.

Response: This is a fair comment. We have added a comprehensive discussion with proper references comparing the different data sets and the significance of our findings as a complementary data set.

2. There are more than five 5′-end RNA-sequencing methods. (https://www.nature.com/articles/s41592-018-0014-2/). Representatively, there are two projects (ENCODE (Encyclopedia of DNA Elements) project and Fantom (Functional Annotation of the Mammalian Genome) project) to create a comprehensive atlas of gene

expression in both humans and mice using 5′-end RNA-sequencing technology, and both adopted the CAGE method. In contrast, both STRT2-seq and CAGE-seq methods are adopted in DoGA. The authors should mention in this manuscript what strategy was used to select both methods that are used to confirm similar results, and what parts were identified differentially using STRT2-seq.

Response: We motivated our choice of STRT2-seq in the manuscript with greater detail, emphasizing the low-input character of our libraries.

3. The authors have created an unprecedented resource in the field of canine research through rigorous efforts, and through the establishment of a data coordination center and genome browser track hubs, it is effectively proposed as a resource for the scientific community and has an “open science” potential value to the whole community. However, it seems necessary to present more organized results about data quality evaluation and biological meaning.
Response: We have added QC reports (see Suppl. Data). We have improved, revised and added additional use case-examples with new results and discussion to expand the biological findings.

4. It is necessary to evaluate the quality of STRT2-seq data of DoGA at each stage of pre-processing and processing and then visualize them so that data users can refer to it.
And the authors should show the data quality level of DoGA in comparison with the standards generally accepted (e.g. ENCODE or FANTOM project). It is recommended to refer to the following two papers for quality check and visualization.
1) Functional annotation of human long noncoding RNAs via molecular phenotyping. (PMID: 32718982) and 2) A promoter level mammalian expression atlas. (PMID: 24670764)

Response: We performed additional evaluation similar to the ones presented in the mentioned articles supporting the quality of our data (Supplementary Figure 3)

5. As mentioned in this publication (PMID: 29867192; Comprehensive comparative analysis of 5′-end RNA-sequencing methods), although bulk RNA-seq is a powerful approach for quantifying gene expression, discovering previously unknown transcripts, and determining splice isoforms, it is often difficult to reliably identify more than one TSS per gene. Therefore, it is believed that dog genome annotation can be improved with a previously unexpected resolution by using DoGA's STRT2-seq and CAGE datasets.
Response: We agree and added this point in the Discussion.

6. According to Figure 3 of BarkBase's publication (PMID: 31181663), there are significant gaps in the existing genome annotation.
Response: We agree, and we added this point to the Introduction and Discussion.

7. As the authors noted in Table 1 in this manuscript, many novel promoters have been identified. However, this manuscript should build an improved genome annotation that can be used by various canine researchers who conduct follow-up studies by integrating analysis with various previously published bulk RNA-seq rather than simply reporting the promoter location.

Response: We agree. We now provide UCSC trackhub tracks for canFam3.1 and canFam4 indicating the promoter types (https://genome-euro.ucsc.edu/s/DoGa/canFam4).

8. The authors utilized an approach used in the Human Protein Atlas (HPA) to detect ubiquitous and tissue-enriched expression of the promoters and genes. Through this analysis, genes can be divided into a total of six categories (Tissue-Enriched, Group-Enriched, Tissue-Enhanced, Expressed-in-all, Mixed, and Not-Expressed). The factors that can change the results in this analysis are the minimum expression cutoff, the fold-change cutoff for expression comparison, and the number of tissues for "Group Enriched" classification. In the case of HPA (Uhlén et al. 2015), the minimum expression criterion is 1 (TPM or FPKM), the fold-change criterion is five-fold, and the number of tissues to distinguish group-enriched is 2-7 tissues. The authors should present an accurate criterion by presenting only the fold change criterion (3-fold higher) in the method.
Response: We clarified our deviations from the approach in the Human Protein Atlas in the Methods.

8. It is recommended to visualize the results for the six categories and to present the differences from existing reports by conducting functional analysis on ubiquitous and tissue-enriched for reliability of the results.
Response: We added a visualization of the number of enriched promoters and genes in the different tissues (Figure 3b,c).

9. In Figure 3A, results are not shown.
Response: We clarified the insights from Figure 3a.

10. The authors presented several usage examples for the utilization of DoGA resources and some visual data. In addition, it was built so that detailed information can be checked in the "DoGA Expression Atlas". In these examples, the authors overlapped the promoters with lineage- and behavior-specific SNPs. Though overlapping statistics and visual figures are presented in the manuscript or "DoGA Expression Atlas", but it's difficult to understand.
Response: We clarified our approach to the SNP overlap and expanded our description of the results.

11. The authors should present the enrichment analysis of canine GWAS summary statistics related with lineage, morphological and behavioral traits for tissue-specific active promoters as a demonstration of resource utility for canine studies similar to the examples below.
- Figure 9, Integrative analysis of 111 reference human epigenomes. (PMID: 25693563)
- Figure S31, Ancestry-inclusive dog genomics challenges popular breed stereotypes (PMID: 35482869)
- Figure 3, An atlas of dynamic chromatin landscapes in mouse fetal development. (PMID: 32728240)
Response: Our data set is too large to show a similar visualization with all SNPs, but we added a figure showing similar results for a subset of the data (Figure 4c).

11. It is expected that DoGA's resources will be used for overall research such as dog genomics, transcriptomics, and epigenomics. Many of the web pages cannot be accessed yet, but it is expected that scientists will be able to better utilize DoGA's data by optimizing

the access and speed of the mentioned web pages (for example, DoGA Expression Atlas; https://expression-atlas.doggenomeannotation.org/dogaatlas/).

Response: We opened up all websites and tried to optimize the speed.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

This is a greatly improved version of the DoGA expression atlas manuscript. The authors have address many of the queries raised by both reviewers, but in doing so, some main text and supplementary material has become confused, and references lost.

The data contained within DoGA will be a great help to the canid community when available, as the use cases exemplify. Some polishing of the current manuscript is still warranted. Specific comments below.

Annotate the columns of Supplementary Table 1.
For example, is age in months, days? "steril_castr" is annotated with 1, 0, what do these mean?
No units are referenced for the stock concentrations. What are these values?

Page 9, paragraph 4. "In addition, the biobank contained cultured fibroblasts from 11 dogs and 2 Finnish wolves obtained by primary culture of skin explants and RNA samples of 12 whole embryos." Are these from the sample described above, or are these additional dogs?

Page 10, section 2.
Are all the STRT2 library preparation steps as described in Ezer et al, or are there additional changes from that published method outlined here?
"(Islam et al., 2012, 2014)" do not appear in the references.
This section also contains many short paragraphs.

Page 11, section 2.
"used by FANTOM5 [ref]." Where is the reference?
"NCBI Canis lupus familiaris Annotation Release 106" does not include CanFam3.1. Was annotation only performed here for UU_Cfam_GSD_1.0/CanFam4?

In the methods, "Expression Atlas" what is "robust" and what is "comprehensive"?

In the methods, "Whole genome sequencing", are the 10 canids from the tissue sampled dogs, or are these other samples? Could be good to link to "Availability of data and scripts here, or perhaps have this statement at the start of the methods to avoid confusion.

Page 3, section "Identification of 150,000 promoters in 100 tissues".
What is the difference between the robust and comprehensive set other than the numbers? Is it the three biological replicates? This could be rephrased.

In table 1, are novel promoters never primary or alternative? Are these a separate class in the "All". Surely novel promoters would be in the alternate or primary classes? Wouldn't it be interesting to know what fraction of primary or alternative promoters were novel?

The numbers in this paragraph don't match those in table 1. What is the difference?

In Figure 3a, there are 4 major clusters, multiple minor, but 12 organ systems. Does figure 3a reflect this, or are there other patterns at work?

How did the authors perform this, but eye, statistically? - "We studied the impacts of biological factors, such as sex, age, breed and did not observe any discernible effects (suppl. Figure 4)." UMAP reduces dimensionality, are these factors tied to higher order PCs?

What is meant by "We experimentally validated five promoter candidates using RT-PCR in relevant tissues including alternative promoter sites (Supplementary Data 2)." Did the results meet you expectation? How did these 5 candidates relate to the chromatin regions? Why does the reader care about these 5 candidates? Did they have different levels of support? Could the results not simply be included in this section and not Supplementary Data 2?

“In comparison with previous findings in humans24, we found 8,561 genes and 7,473 promoters enriched in one or more tissues, with testis having the most enriched promoter regions (2,653) and genes (5,340) (Figure 3b).” How is this compared to humans? What was the human result?

OMIA refers to the inheritance of the disorder, Mendelian phenotype, not the alleles, monogenic. There are polygenic/ complex traits in OMIA.

For the OMIA comparisons, would the variants implicated in disease likely impact transcript usage as indicated by primary and alternate promoters? For example, does RBP4 use the same promoters in retina, common bile duct and liver?

In “Example use-case: Conservation of tissue-enriched gene expression between dogs and humans”, for the tissue-enriched genes are these 1:1 orthologues or 1:many? It will affect the interpretation.

In “Example use-case: Overlap of promoters with lineage- and behaviour-specific SNPs”, how big does an overlap need to be given the border resolution issue of defining a promoter space?

In the Discussion. EpiC Dog (doi.org/.1126/sciadv.ade3399, epigenomics) and BarkBase (doi.org/10.3390/genes10060433, RNA Seq plus ATAC Seq) have different goals to DoGA and use different technologies. It is true that the number of tissues are less in these studies, but the range of experiment types are greater. The authors should qualify their discussion points.

Minor
Page 9, paragraph 1. “Tissue pieces or appr. A 1 x 1 …” What is the A?
Page 10, section 2. Grammar. Suggest, Previous studies have shown the unbiased performance of STRT2 in GC rich regions 40–46.
Page 11, section 3. Check et al.,
Be consistent. CanFam3.1 and canFam4 or UU_Cfam_GSD_1.0 or UU_Cfam_GSD_1.0/CanFam4. There is a mix of terms referring to these builds.
Many short paragraphs in the methods sections. Check formatting and grammar.
Be consistent. Sometimes full gene names are italicised, other times not. Choose one format.
Gene abbreviations should always be italicised. Currently they are not. Please check and correct.
In the methods, what is the date of access for OMIA?
In the methods, what are the references for the dog linage and dog behaviour studies?
Page 3, “Supplementary Table 2” is a list of primers not the list of samples or replicates.
Check use of decimal point or commas when defining numbers in the thousands. See paragraph 4 on page 3 as an example, i.e., 18.367.
The order of the supplementary figures does not match the text, e.g., suppl figure 4 is mentioned before suppl figure 3. Suppl figure 3 does not appear in the text at all.
“Supplementary Data 2” comes before Supplementary Data 1 or Supp. Data 1 in the main text. There is no mention of Supp Data 1 in the main text.
BarkBase and EpiC Dog not cited correctly in the discussion.
BarkBase not Barkbase

Reviewer #2 (Remarks to the Author):

The authors have responded to some extent to the reviewer's comments and made improvements to the manuscript. However, when compared to the extensive dataset generated from 100 tissues, the results still lack in-depth analysis. This deficiency pertains not only to the biological significance but also to the need for a more in-depth analysis of the STRT2-seq data itself and subsequent discussions following this analysis.

For both the known and newly identified promoters, it would be valuable to demonstrate how well they correlate with histone marks assessed by EpicDog and to what extent differences are

observed. The manuscript should highlight their general features and provide specific examples within a given tissue context.

Main Figures 3 and 4 are too basic and not suitable for the format of Nature Communications. The authors should aim to provide a more profound analysis and enhance the presentation. Additionally, Figure 3a remains unclear, as nothing is visible in the UMAP plot.

On page 3, line 7 from the bottom, '18.367' should be corrected to '18,367.'

When citing other papers, there is no need to devalue the work of others by solely comparing the quantity of tissue types. Different studies have distinct focuses and use varying types of data. In the discussion, the authors should rephrase this aspect.

The Supplementary Figures should be properly formatted and integrated into the paper. They currently appear too preliminary. Since the numbers are deleted, and we can not locate the Supplementary Figures.

The authors should also clearly define their contributions, including those of newly added authors."

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

This is a greatly improved version of the DoGA expression atlas manuscript. The authors have address many of the queries raised by both reviewers, but in doing so, some main text and supplementary material has become confused, and references lost.

Response: Thank you for the feedback. The additional suggestions were greatly appreciated and further improved the manuscript.

The data contained within DoGA will be a great help to the canid community when available, as the use cases exemplify. Some polishing of the current manuscript is still warranted. Specific comments below.

Annotate the columns of Supplementary Table 1.

For example, is age in months, days? “steril_castr” is annotated with 1, 0, what do these mean? No units are referenced for the stock concentrations. What are these values?

Response: We added the missing units to the table.

Page 9, paragraph 4. “In addition, the biobank contained cultured fibroblasts from 11 dogs and 2 Finnish wolves obtained by primary culture of skin explants and RNA samples of 12 whole embryos.” Are these from the sample described above, or are these additional dogs?

Response: We clarified this in the methods section.

Page 10, section 2.

Are all the STRT2 library preparation steps as described in Ezer et al, or are there additional changes from that published method outlined here?

Response: The publication is based on this study, so they are the same steps.

“(Islam et al., 2012, 2014)” do not appear in the references.

This section also contains many short paragraphs.

Response: We merged several paragraphs.

Page 11, section 2.

“used by FANTOM5 [ref].” Where is the reference?

“NCBI Canis lupus familiaris Annotation Release 106” does not include CanFam3.1. Was annotation only performed here for UU_Cfam_GSD_1.0/CanFam4?

Response: We clarified the used annotation in the methods section

In the methods, “Expression Atlas” what is “robust” and what is “comprehensive”?

Response: We clarified the terms in the section.

In the methods, “Whole genome sequencing”, are the 10 canids from the tissue sampled dogs, or are these other samples? Could be good to link to “Availability of data and scripts here, or perhaps have this statement at the start of the methods to avoid confusion.

Response: Yes, these are the same. We added a mention of the availability in this section.

Page 3, section “Identification of 150,000 promoters in 100 tissues”.

What is the difference between the robust and comprehensive set other than the numbers? Is it the three biological replicates? This could be rephrased.

Response: We rephrased the statement to make it clearer.

In table 1, are novel promoters never primary or alternative? Are these a separate class in the “All”. Surely novel promoters would be in the alternate or primary classes? Wouldn’t it be interesting to know what fraction of primary or alternative promoters were novel?

Response: We added Figure 3a to clarify our promoter types. The identification of alternative promoters inside the class of novel promoters is non-trivial and needs some further computational analysis, because one needs to find a good clustering of related novel promoters.

The numbers in this paragraph don’t match those in table 1. What is the difference?

Response: We corrected the numbers.

In Figure 3a, there are 4 major clusters, multiple minor, but 12 organ systems. Does figure 3a reflect this, or are there other patterns at work?

Response: We clarified Figure 3a in both the Results and the Discussion sections.

How did the authors perform this, but eye, statistically? - “We studied the impacts of biological factors, such as sex, age, breed and did not observe any discernible effects (suppl. Figure 4).” UMAP reduces dimensionality, are these factors tied to higher order PCs?

Response: The reviewer is correct, that this is only based on the visual inspection of the UMAP plots. We clarified this in the text.

What is meant by “We experimentally validated five promoter candidates using RT-PCR in relevant tissues including alternative promoter sites (Supplementary Data 2).” Did the results meet you expectation? How did these 5 candidates relate to the chromatin regions? Why does the reader care about these 5 candidates? Did they have different levels of support? Could the results not simply be included in this section and not Supplementary Data 2?

Response: We transferred the relevant parts from the supplementary data into the result section. We randomly selected 5 candidates for validation of our STRT2-based promoter candidates.

“In comparison with previous findings in humans24, we found 8,561 genes and 7,473 promoters enriched in one or more tissues, with testis having the most enriched promoter regions (2,653) and genes (5,340) (Figure 3b).” How is this compared to humans? What was the human result?

Response: The Human Protein Atlas showed a similar high enrichment in testis samples.

OMIA refers to the inheritance of the disorder, Mendelian phenotype, not the alleles, monogenic. There are polygenic/ complex traits in OMIA.

Response: It’s true that there are also more complex traits in OMIA, we only looked at monogenic traits and clarified this in the text.

For the OMIA comparisons, would the variants implicated in disease likely impact transcript usage as indicated by primary and alternate promoters? For example, does RBP4 use the same promoters in retina, common bile duct and liver?

Response: Thanks for this excellent suggestion. We expanded on it in the results and discussion sections and added a new supplementary figure for clarification.

In “Example use-case: Conservation of tissue-enriched gene expression between dogs and humans”, for the tissue-enriched genes are these 1:1 orthologues or 1:many? It will affect the interpretation.

Response: We used 1:1 orthologues and clarified this in the text.

In “Example use-case: Overlap of promoters with lineage- and behaviour-specific SNPs”, how big does an overlap need to be given the border resolution issue of defining a promoter space?

Response: We used a 1bp overlap.

In the Discussion. EpiC Dog (doi.org/.1126/sciadv.ade3399, epigenomics) and BarkBase (doi.org/10.3390/genes10060433, RNA Seq plus ATAC Seq) have different goals to DoGA and use different technologies. It is true that the number of tissues are less in these studies, but the range of experiment types are greater. The authors should qualify their discussion points.

Response: The reviewer is correct and we modified our discussion to highlight the complementary nature of the different studies. All three studies are fantastic additions to the functional annotation of the dog genome.

Minor

Page 9, paragraph 1. “Tissue pieces or appr. A 1 x 1 …” What is the A?

Page 10, section 2. Grammar. Suggest, Previous studies have shown the unbiased performance of STRT2 in GC rich regions 40–46.

Page 11, section 3. Check et al.,

Be consistent. CanFam3.1 and canFam4 or UU_Cfam_GSD_1.0 or UU_Cfam_GSD_1.0/CanFam4. There is a mix of terms referring to these builds.

Many short paragraphs in the methods sections. Check formatting and grammar.

Be consistent. Sometimes full gene names are italicised, other times not. Choose one format. Gene abbreviations should always be italicised. Currently they are not. Please check and correct.

In the methods, what is the date of access for OMIA?

In the methods, what are the references for the dog linage and dog behaviour studies?

Page 3, “Supplementary Table 2” is a list of primers not the list of samples or replicates.

Check use of decimal point or commas when defining numbers in the thousands. See paragraph 4 on page 3 as an example, i.e., 18.367.

The order of the supplementary figures does not match the text, e.g., suppl figure 4 is mentioned before suppl figure 3. Suppl figure 3 does not appear in the text at all.

“Supplementary Data 2” comes before Supplementary Data 1 or Supp. Data 1 in the main text. There is no mention of Supp Data 1 in the main text.

BarkBase and EpiC Dog not cited correctly in the discussion.

BarkBase not Barkbase

Response: We corrected these errors.

Reviewer #2 (Remarks to the Author):

The authors have responded to some extent to the reviewer's comments and made improvements to the manuscript. However, when compared to the extensive dataset generated from 100 tissues, the results still lack in-depth analysis. This deficiency pertains not only to the biological significance but also to the need for a more in-depth analysis of the STRT2-seq data itself and subsequent discussions following this analysis.

Response: Thank you for the feedback and additional useful suggestion. We now have added new analyses of the data, figures, clarifications, and about 500 new wolf samples in the biobank part.

For both the known and newly identified promoters, it would be valuable to demonstrate how well they correlate with histone marks assessed by EpicDog and to what extent differences are observed. The manuscript should highlight their general features and provide specific examples within a given tissue context.

Response: We added a comparison to the manuscript as well as to the expression atlas.

Main Figures 3 and 4 are too basic and not suitable for the format of Nature Communications. The authors should aim to provide a more profound analysis and enhance the presentation. Additionally, Figure 3a remains unclear, as nothing is visible in the UMAP plot.

Response: We have revised Figures 3 and 4 to make them more readable.

On page 3, line 7 from the bottom, '18.367' should be corrected to '18,367.'

When citing other papers, there is no need to devalue the work of others by solely comparing the quantity of tissue types. Different studies have distinct focuses and use varying types of data. In the discussion, the authors should rephrase this aspect.

Response: We didn't want to devalue other studies, just highlight the differences as requested by the reviewers. We clarified the complementary nature and value of the different studies in the discussion section.

The Supplementary Figures should be properly formatted and integrated into the paper. They currently appear too preliminary. Since the numbers are deleted, and we can not locate the Supplementary Figures.

Response: We updated the supplementary figures to integrate them better into the paper.

The authors should also clearly define their contributions, including those of newly added authors."

Response: We added this section to the manuscript

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The DoGA project contains a wealth of information that will be of use to not only the canine community, but to cross species genetics.

This round of the DoGA manuscripts includes some comments between the co-authors and so it may not be the completed version. some elements to be addressed are below. Particularly, Supplementary Files 1 and 2 were found, but not the described Sup Figures including UMAPs. Did this not upload properly or is there another issue at play.

Page 3, Results - A large canine tissue biobank
What is DoGA DCC? Data co-ordination centre? A web link would be helpful.
What are Sup Figs 1,2? "Supplementary_Data_1_and_2_1710500996_1" only contains images of PCRs. Is this mislabeled?

Page 3/4, Results – Identification of 100,000 promoters in 100 tissues
What was the annotation files used for genes? Could be good to add a note here so that the reader doesn't have flip to the methods (there it is included – NCBI v106).
Are there 4 or 5 major groups in the plot? What does four/five mean and why then circle six groups in the plot?
Where is Sup Figs 3? Is this an additional UMAP? "Supplementary_Data_1_and_2_1710500996_1" only contains images of PCRs.

Page 4, Results – Validation of identified promoters
TNNI3 – missing italics

Page 4, Results – Tissue enrichment or genes and their promoters
"In comparison with previous findings in humans…", What is the comparison? The number of genes, the number in teste?
Given the use of CanFam3.1 and CanFam4 throughout the text (including in the previous paragraph), it would be good to clarify that this section is for CanFam4.
For the tissue enriched genes and promoters, is this based on a canonical transcript, or is it merged data?
EpicDog was produced on CanFam3.1. Is the use of EpicDog here lifted to CanFam4, remapped, or something else?

Page 5, Results – Use-case OMIA
What do you want the reader to take from the RBP4 example? A narrowing of the transcript usage for retina, or that the transcript is used across tissues? This use-case seems to lack the question you wanted the user to ask and so how the data would be used. A small refinement of text could be warranted.

Page 5, Results – Use-case Conservation
Are these 1:1 orthologues? If yes, determined how? If not, state this so that the reader can better use the data.

Page 5, Results – Use-case SNPs
How does the co-localisation of SNPs and promoters take you past what was known before?

Check use, CanFam 3 or CanFam3.1?

Page 12, tissue enrichment.
A 1:1 orthologues for gene, let alone promoter is challenging. However, a table for human and dog 1:1 orthologues genes, and gene losses is available as part of TOGA. Perhaps the authors could check this list and update their text? TOGA: Bogdan M. Kirilenko et al. ,Integrating gene annotation with orthology inference at scale. Science380,e abn3107 (2023). DOI:10.1126/science.abn3107

Page 14. Who is Dc in the contributions? Also, some listed authors do not have contributions, e.g., Jeffrey Schoenebeck

Page 15. Table 1. “The highest-expressed promoter region of a gene was classified as primary; additionally, if present, it was categorized as an alternative promoter”. Are there some words missing? Is the primary also an alternate?

Annotations in Supplementary Table 1 are still missing. Without this information the metadata is not complete. For example, what is “steril_castr” 0, 1, 2? Is age in months? The stick concentration does not have units.

Reviewer #2 (Remarks to the Author):

I have had the opportunity to review your manuscript submitted to Nature Communications, which integrates a vast amount of data across dogs, wolves, and humans to provide insightful information on promoters and associated genes relevant to canine research. This study presents a significant contribution to the field, given its unprecedented scale of data collection from a wide variety of organs and tissues for the study of promoters and expression in dogs. It is evident that compiling and interpreting this data required considerable effort and dedication. Despite the commendable work, I believe certain areas could be improved to enhance the manuscript's completeness and impact. Below are my comments and suggestions for your consideration:

Figures 1 and 2 Utilization: Currently, Figures 1 and 2 occupy a substantial amount of space with information primarily about the samples used. I suggest condensing this information into a single figure to streamline the presentation and make efficient use of figure space.

Methodology on Promoter Identification: The manuscript lacks a detailed methodological description of the promoter identification process. It is crucial to provide a comprehensive explanation of the criteria and methodologies employed for promoter identification to ensure clarity and reproducibility of the findings.

Enhancement of "Example Use-Case" Sections: The "Example use-case" sections spanning pages 5-6 effectively illustrate the research's significance and applications. I recommend dedicating a main figure to each example use-case to visually represent the findings and their implications more clearly. For instance, if specific phenotypic SNPs in the promoter region are linked to "Sorry when wrong" behavior, illustrations could include epigenetic peaks in the IGV browser for related genes, associations with certain genes, or gene ontology of significant genes.

Additional Information on Newly Discovered Promoter Candidates: In the Results section, line 116 mentions the discovery of 14,394 new promoter candidates. Further elaboration on these findings is needed, including the criteria for their identification, how these regions were previously understood, reasons for their prior anonymity, their epigenetic features, and their tissue specificity.

Your study provides valuable insights into canine genetics and has the potential to significantly advance our understanding of gene expression and regulation in dogs and related species. The suggestions mentioned above are intended to further refine and highlight the importance of your work. I look forward to seeing the enhanced version of your manuscript and believe that with these adjustments, your research will make an even more substantial contribution to the field.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The DoGA project contains a wealth of information that will be of use to not only the canine community, but to cross species genetics.

This round of the DoGA manuscripts includes some comments between the co-authors and so it may not be the completed version. some elements to be addressed are below. Particularly, Supplementary Files 1 and 2 were found, but not the described Sup Figures including UMAPs. Did this not upload properly or is there another issue at play.

Page 3, Results - A large canine tissue biobank
What is DoGA DCC? Data co-ordination centre? A web link would be helpful.
Response: We removed the first mention of the DCC and added a link to the resource section.

What are Sup Figs 1,2? “Supplementary_Data_1_and_2_1710500996_1” only contains images of PCRs. Is this mislabeled?
Response: Sorry for the confusion. We clarified the files list and provide four different supplementary files:

- “Supplementary Figures 1-5.docx”. We added a new figure in this revision and the file is now renamed as “Supplementary Figures 1-6.docx”.
- “Supplementary_Tables 1-4.xslx”.
- “Supplementary Data 1. QC report for STRT2 sequencing with MultiQC report.html”,
- “Supplementary Data 2.docx” with PCR validation results.

All supplementary figures, tables and data are referenced in the text when appropriate.

Page 3/4, Results – Identification of 100,000 promoters in 100 tissues
What was the annotation files used for genes? Could be good to add a note here so that the reader doesn’t have flip to the methods (there it is included – NCBI v106).
Response: We added the missing reference.

Are there 4 or 5 major groups in the plot? What does four/five mean and why then circle six groups in the plot?
Response: We improved the annotation and description of this figure by including all samples. There are, in total, six groups, as now clarified in the text.

Where is Sup Figs 3? Is this an additional UMAP?
“Supplementary_Data_1_and_2_1710500996_1” only contains images of PCRs.
Response: Sorry for the confusion. The Sup Figs 3 is included in the “Supplementary Figures 1-5.docx”, which is renamed “Supplementary Figures 1-6.docx” during this revision.
Supplementary Data 1 a separate large html file about STRT2 qc results.
Supplementary Data 2 contains PCR validations results.

Page 4, Results – Validation of identified promoters
TNNI3 – missing italics
Response: We corrected this mistake.

Page 4, Results – Tissue enrichment or genes and their promoters
“In comparison with previous findings in humans…”, What is the comparison? The number of genes, the number in teste?

Given the use of CanFam3.1 and CanFam4 throughout the text (including in the previous paragraph), it would be good to clarify that this section is for CanFam4.
Response: Thanks for the suggestions. We clarified this in the text, including the reference genome.

For the tissue enriched genes and promoters, is this based on a canonical transcript, or is it merged data?
Response: We don't understand this question. We used our robust sets of gene-level expressions and promoters to identify tissue-enriched elements.
EpicDog was produced on CanFam3.1. Is the use of EpicDog here lifted to CanFam4, remapped, or something else?
Response: We clarified our remapping to CanFam4 in the results section.

Page 5, Results – Use-case OMIA
What do you want the reader to take from the RBP4 example? A narrowing of the transcript usage for retina, or that the transcript is used across tissues? This use-case seems to lack the question you wanted the user to ask and so how the data would be used. A small refinement of text could be warranted.
Response: We added a sentence to clarify our point of the example.

Page 5, Results – Use-case Conservation
Are these 1:1 orthologues? If yes, determined how? If not, state this so that the reader can better use the data.
Response: Yes, these are 1:1 orthologues based on gene symbols.

Page 5, Results – Use-case SNPs
How does the co-localisation of SNPs and promoters take you past what was known before?
Response: Our data helps to prioritize SNPs based on promoters and their tissue expression patterns. For example, our data gives insights into expression patterns across different brain regions for promoters overlapping behavioral SNPs. We further clarified this in the discussion.

Check use, CanFam 3 or CanFam3.1?
Response: We changed the vocabulary in the whole text text to use the correct CanFam3.1.

Page 12, tissue enrichment.
A 1:1 orthologues for gene, let alone promoter is challenging. However, a table for human and dog 1:1 orthologues genes, and gene losses is available as part of TOGA. Perhaps the authors could check this list and update their text? TOGA: Bogdan M. Kirilenko et al. ,Integrating gene annotation with orthology inference at scale. Science380,e abn3107 (2023). DOI:10.1126/science.abn3107
Response: Thank you for the reference. We are aware of the difficulty concerning orthologue identification, but such in-depth analysis is out of the scope of this study. However, we now highlight this opportunity for future studies in the discussion.

Page 14. Who is Dc in the contributions? Also, some listed authors do not have contributions, e.g., Jeffrey Schoenebeck
Response: We updated the contributors list to include all authors.

Page 15. Table 1. "The highest-expressed promoter region of a gene was classified as primary; additionally, if present, it was categorized as an alternative promoter". Are there some words missing? Is the primary also an alternate?
Response: Thanks for pointing out this inconsistency. We clarified the sentence.

Annotations in Supplementary Table 1 are still missing. Without this information the

metadata is not complete. For example, what is “steril_castr” 0, 1, 2? Is age in months? The stick concentration does not have units.
Response: We updated the units and clarified the entries for “steril_castr”.

Reviewer #2 (Remarks to the Author):

I have had the opportunity to review your manuscript submitted to Nature Communications, which integrates a vast amount of data across dogs, wolves, and humans to provide insightful information on promoters and associated genes relevant to canine research. This study presents a significant contribution to the field, given its unprecedented scale of data collection from a wide variety of organs and tissues for the study of promoters and expression in dogs. It is evident that compiling and interpreting this data required considerable effort and dedication. Despite the commendable work, I believe certain areas could be improved to enhance the manuscript’s completeness and impact. Below are my comments and suggestions for your consideration:

Figures 1 and 2 Utilization: Currently, Figures 1 and 2 occupy a substantial amount of space with information primarily about the samples used. I suggest condensing this information into a single figure to streamline the presentation and make efficient use of figure space.
Response: We disagree. Figure 1 serves as a necessary overview of the DoGA project, including the DoGA biobank, which will be referenced in a series of upcoming papers utilizing the resources and manuscript, while figure 2 gives a summary of the inviduadal sampled tissues and tissue enrichment results from this study.

Methodology on Promoter Identification: The manuscript lacks a detailed methodological description of the promoter identification process. It is crucial to provide a comprehensive explanation of the criteria and methodologies employed for promoter identification to ensure clarity and reproducibility of the findings.
Response: We added a more detailed description of the promoter identification in the methods section.

Enhancement of "Example Use-Case" Sections: The "Example use-case" sections spanning pages 5-6 effectively illustrate the research’s significance and applications. I recommend dedicating a main figure to each example use-case to visually represent the findings and their implications more clearly. For instance, if specific phenotypic SNPs in the promoter region are linked to "Sorry when wrong" behavior, illustrations could include epigenetic peaks in the IGV browser for related genes, associations with certain genes, or gene ontology of significant genes.
Response: Thank you for the suggestions. We have revised Figure 4 to include visualization for every use case and expanded the section on behavioral SNPs.

Additional Information on Newly Discovered Promoter Candidates: In the Results section, line 116 mentions the discovery of 14,394 new promoter candidates. Further elaboration on these findings is needed, including the criteria for their identification, how these regions were previously understood, reasons for their prior anonymity, their epigenetic features, and their tissue specificity.
Response: We improved our description of the features of the novel promotor regions, including epigenetic marks (Table 1) and tissue specificity. Additionally, we discuss reasons why they have not been discovered until now.

Your study provides valuable insights into canine genetics and has the potential to significantly advance our understanding of gene expression and regulation in dogs and related species. The suggestions mentioned above are intended to further refine and highlight the importance of your work. I look forward to seeing the enhanced version of your

manuscript and believe that with these adjustments, your research will make an even more substantial contribution to the field.

Response: Thank you for your encouraging comments. We hope the revised version is ready for publication to give the community access to our resources.

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

I appreciate the efforts the author have made in this round of review, including the including of correlation plots. There remain just a very few elements to address.

Could the authors comment on why the numbers of wolves in the Biobank change between review versions?
Page 3, with the re-inclusion of a wolve, the Results section should now include 43 and not 42 animals.
Page 4 and Supp Figure 3. The second UMAP is for Animal IDs and not breed as per the caption. Should there be 9 breeds (as per Table S1), or 17 samples (dogs and embryos) as per the sampling. The UMAP has 11 elements in the legend.
Is the overlap for epigenetic marks in Table 1 only for robust promoters? Does this carry over to the primary, alternative and novel? If yes, clarify the table text.
Page 4, TNNI3 gene should be italicised
Page 4, canFam4 not CanFam4.1
Page 7, RBP4 gene names in italics.
There are issues with the author list and their contributions. Where is Işıl Takan and who are RFAA and ABK?
With the current labels, Supp. Fig 5 cannot be understood. The X and Y labels are unreadable. What do the stars mean in the Corr. plot?
Be consistent and use one. “Supplementary Figure”, “Suppl. Figure” or “Supplementary Fig.” in the text.
Supplementary Figures are out of order in the text, e.g., Supp. Fig 6 is mentioned before Supp. Fig 5.

Reviewer #2 (Remarks to the Author):

The authors responded well to my comments. I think the manuscript is now acceptable.

Response to the Reviewer

Could the authors comment on why the numbers of wolves in the Biobank change between review versions?

Response: The DoGA project continue and we keep collecting new animals and add them to the biobank, hence updating numbers. The sample resource is also highly valuable for the community.

Page 3, with the re-inclusion of a wolve, the Results section should now include 43 and not 42 animals.

Response: We updated the numbers in the text and Fig 1, since we have collected more wolves during the review process.

Page 4 and Supp Figure 3. The second UMAP is for Animal IDs and not breed as per the caption. Should there be 9 breeds (as per Table S1), or 17 samples (dogs and embryos) as per the sampling. The UMAP has 11 elements in the legend.

Response: We updated the Supplementary Figures to also include a plot showing the distribution of 9 breeds in this projection. The UMAP based on animal_id shows 11 elements, due to the collapse of the embryo samples into two groups based on age.

Is the overlap for epigenetic marks in Table 1 only for robust promoters? Does this carry over to the primary, alternative and novel? If yes, clarify the table text.

Response: The overlap was for robust promoter candidates. We clarified this in the figure caption**.**

Page 4, TNNI3 gene should be italicised
Page 4, canFam4 not CanFam4.1
Page 7, RBP4 gene names in italics.

Response: Fixed.

There are issues with the author list and their contributions. Where is Işıl Takan and who are RFAA and ABK?

Response: Fixed.

With the current labels, Supp. Fig 5 cannot be understood. The X and Y labels are unreadable. What do the stars mean in the Corr. plot?

Response: We updated the figure legend for clarification.

Be consistent and use one. “Supplementary Figure”, “Suppl. Figure” or “Supplementary Fig.” in the text.

Response: Fixed.

Supplementary Figures are out of order in the text, e.g., Supp. Fig 6 is mentioned before Supp. Fig 5.

Response: Fixed.