IMPROVING CANINE GENOME ANNOTATION BY IDENTIFICATION OF OVER-MERGED GENES AND RETROGENES IN CURRENT DATABASES

by

YECHENG HUANG

(Under the Direction of Shaying Zhao)

ABSTRACT

The dog represents an important model animal in biomedical, evolutionary and behavioral research. The current canine gene annotation has been greatly improved but still significantly lag behind in accuracy compared to those of the human and the mouse. To address this, we identified several significant issues in the current canine gene annotations. These include gene over-merging, gene name mislabeling, retrogene misclassification, coding region misidentification, and UTR fragmentation. We systematically addressed these issues by a novel pipeline that also utilizes our own data.

INDEX WORDS: CANINE, GENOME ASSEMBLY, GENOME ANNOTATION, OVER-MERGING, RETROGENE

IMPROVING CANINE GENOME ANNOTATION BY IDENTIFICATION OF OVER-MERGED GENES AND RETROGENES IN CURRENT DATABASES

by

YECHENG HUANG

B.S., University of Nanjing, CHINA, 1998

M.S., University of Iowa, 2000

M.C.S., University of Iowa, 2001

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2016

IMPROVING CANINE GENOME ANNOTATION BY IDENTIFICATION OF OVER-MERGED GENES AND RETROGENES IN CURRENT DATABASES

by

YECHENG HUANG

Major Professor: Shaying Zhao

Committee: Jonathan Arnold

Liming Cai

Liang Liu

Electronic Version Approved:

Suzanne Barbour

Dean of the Graduate School

The University of Georgia

December 2016

DEDICATION

To my loving family, Kui Wang, Timothy Huang and Daniel Huang, who made this possible.

ACKNOWLEDGMENTS

I am grateful to my committee members Shaying Zhao, Jonathan Arnold, Liming Cai, Liang Liu; Shaying lab members; team members at GACRC/EITS for their support and guidance.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 INTRODUCTION

The dog serves as an important model organism in diverse fields including biomedical research, evolution, development, and behavior [1-5]. For example, spontaneous cancers in pet dogs represent one of the best cancer models [3, 4, 6-11]. As companion animals, dogs share the same environment as humans and are exposed to many of the same carcinogens. Indeed, environmental toxins, advancing age and obesity are also risk factors for canine cancer [6]. Therefore, dogs develop many same or similar diseases as human. Many cancers are naturally occurring and heterogeneous, capturing the essence of human cancer, which genetically modified or xenograft rodent models cannot replicate. Dogs also better resemble humans in biology, e.g., similar telomere and telomerase activities [12] and frequent spontaneous epithelial cancers [6], unlike mice [13]. Numerous anatomic and clinical similarities are noted for the same types/subtypes of cancer between the two species, and similar treatment schemes are used [4, 7, 8]. Furthermore, the large population of pet dogs (~70 million estimated in the US) provides abundant resource facilitating basic and clinical research. Because of these advantages, we have successfully developed a novel dog-human comparison strategy to address a central aim of cancer research – cancer driver-passenger discrimination [9, 10, 14-17].

The sequencing of the genome of man's best friend in 2005[18] has reinforced the position of the dog as an important animal model to study human physiology and disease. However, to fully utilize the great potential of the dog, an accurate version of canine gene annotation is essential. This however remains yet to be achieved [19, 20], as illustrated

by protein-coding genes as follows. Like other mammalian genomes, the canine genome is large, about 2.7Gb [18], and complex, with >50% of the genome made of transposable elements (>50%) and other repetitive sequences [18]. Only 1-2% encodes protein-coding genes, approximately 20,000 in total and scattering in the genome with the gene density varying greatly. The genes contain nine exons per gene on average. Exons, ranged from a few to several thousand base pairs, are separated by introns of 10bp to 800kb in size. Clearly, to precisely locate each exon for each of ~20,000 genes in the dog genome is a daunting task.

Gene identification in a sequenced mammalian genome is traditionally achieved by evidenced-based approaches. This is usually done by mapping already-existing gene-related sequences (e.g., expressed sequence tags or ESTs, mRNA or cDNA sequences, protein sequences) to the genome, in combination with *ab initio* gene predication[3-5] by software programs such as Gene Scan and others. Human and mouse represents the best-annotated mammalian genomes. This is made possible by several large human or mouse EST or cDNA sequencing projects, including the mammalian gene collection (mgc.nci.nih.gov/), and the data generated by numerous scientists worldwide devoting to specific genes in past decades. In fact, even before the publication of the human and mouse genomes in 2001[21] and 2002, databases have already been established to assemble and annotate human genes, including the gene indices, Unigene, FANTOM db (fantom.gsc.riken.jp). These resources greatly facilitates the annotation of the human[21] and mouse genomes.

For the dog, this unfortunately is not the case, with only 382,638 ESTs and <2,500 curated Refseq transcripts. Hence, the initial canine gene annotation was primarily achieved by mapping RefSeq transcripts, EST, mRNA, and protein sequences from the human and other species to the dog genome[18, 19]. This is how the canine XenoRefGene database is built at the University of Santa Cruz (UCSC) genome site. While providing an unprecedentedly large number of canine genes, this annotation has a number of significant issues.

Recently emerged next-generation sequencing (NGS) technologies have revolutionized genome sequencing and gene annotation. RNA-seq[22], whole transcriptome shotgun sequencing, is especially valuable, as it can efficiently identify alternative splicing, discover new genes and missing/incomplete exons. Various groups, including the Broad Institute [20] and us[15, 16], have performed RNA-seq on various canine tissues and cells. As a result, two new canine gene annotations, entirely built based on canine RNA-seq data, have been released by the Ensembl genome site and the Broad Institute[20]. We refer these two annotations as Ensembl and Broad hereafter. However, as software tools analyzing RNA-seq are imperfect and still needs improving, significant issues still exist with these annotations.

To improve the canine gene annotation, we set out to compare the current three canine gene annotation databases, XenoRefGene, Ensembl and Broad. We identified and effectively addressed several significant issues in these databases, as described by the study below.

CHAPTER 2 METHOD OF ASSEMBLY AND ANNOTATION

Public canine annotation databases downloaded.

The three public canine gene annotation databases used in our analysis are all based on the canFam3 genome assembly. These include Canis_familiearis.CanFam3.1.81.gtf, downloaded from the Ensembl genome sites (http://useast.ensembl.org/info/data/ftp/index.html), Broad improved V1 (Broad V1)[20], and UCSC XenoRefGene, all downloaded from the UCSC genome site (http://www.genome.ucsc.edu/). The XenoRefGene data were further processed to remove all nonhuman transcripts.

Canine transcript assembly building

Our own RNA-seq data are from 9 normal canine skin and mammary tissues samples. As previously described [15, 16], each sample was sequenced with the Illumina technology, yielding 54 to 64 million paired-end sequence reads of 50 bp.

RNA-seq data quality was checked by FastQC version 0.11.2[23]. Then, RNA sequences were assembled with reference-dependent and independent approaches, as outlined in Figure S1. For reference-dependent approach, we mapped RNA-seq reads to the reference genome CanFam 3.1.81 by Bowtie version 1.1.1[24], TopHat version 2.0.13[25]. Lastly, transcripts were assembled with Cufflinks version 2.2.1[22] [26] with the Ensembl annotation database (3.1.81). The reference-independent assembly was achieved by using Trinity version 2.0.6[27] to construct transcripts de novo for each

sample. Then, each trinity transcript was aligned to the canine reference genome with Blat version 36[28].

To build the UGA raw assembly, Trinity transcripts and Cufflinks transcripts of each of the 9 samples were merged via Cuffcmp from Cuffinks package version 2.2.1 [**Figure 1**]. Then, UGA raw assembly was merged with the Ensembl annotation Canis_familiearis.CanFam3.1.81.gtf and the BroadV1 annotation via Cuffcmp. Finally, UGA transcripts that share identical intron regions [**Figure 2**] with either Ensembl or Broad transcripts were identified, classified as transcripts with higher confidence level, and subject to further annotation analyses. [**Figure 1**].

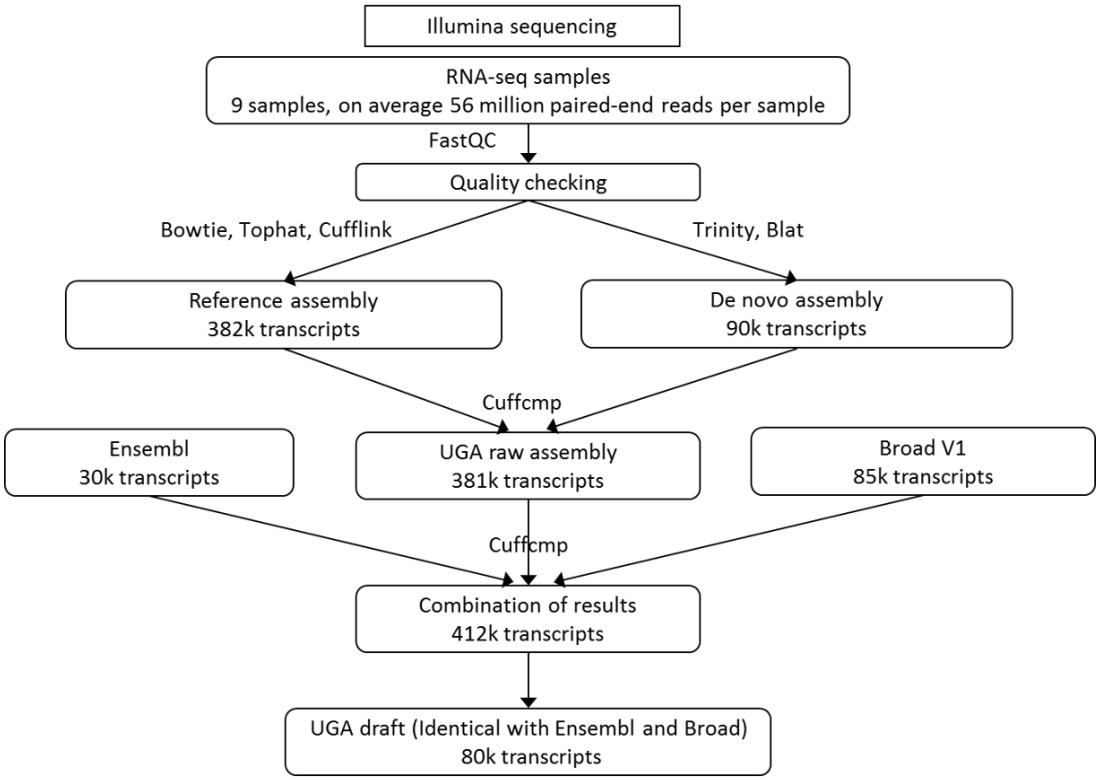

Figure 1: We identified about 80,000 canine transcripts of a high confident level via novel assembly pipeline and UGA data

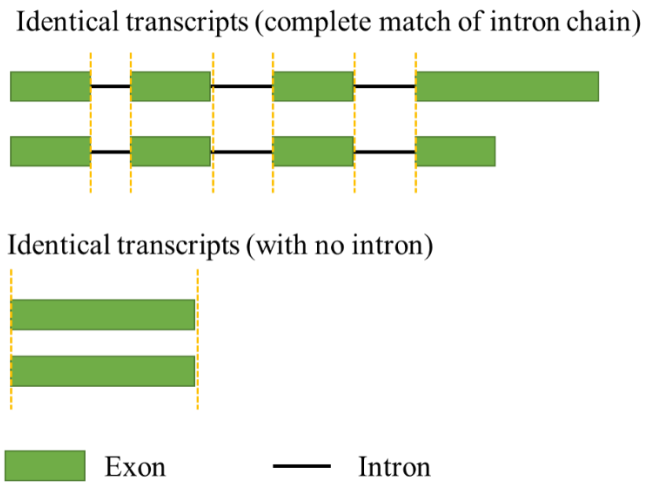

Figure 2: Identical transcripts with complete match of intron region chain

Annotation of canine transcripts

Our annotation pipeline is outlined in [**Figure 3**]. First, we identified the Open Reading Frame (ORF) of each transcript via TransDecoder version 2.0.1[29]. Second, we searched the ORF sequences against the manually curated protein database UniProtKB/Swiss-Prot Release 2015_08 [30] using NCBI Blast+ version 2.2.9[31]. To annotate each coding transcript, we selected the top hit among those with the blast expectation value (E) of < 1e-5. Then, we linked the UniProtKB/Swiss-Prot protein ID assigned by the TrasnDecoder.Predict function to the official gene symbol by querying the UniProt database. For those without official gene symbols found, the first gene name aliases or the UniProtKB/Swiss-Prot IDs were adapted. Finally, for transcripts with no blast hits from UniProtKB/Swiss-Prot, we continued to use their XenoRefGene, Ensembl, or Broad names and IDs. With results from Blast and Hmmer, the ORFs are aligned to genome data by utility cdna_alignment_orf_to_genome_orf in TransDecoder to generate the genome with the identified coding region. The genome data with coding region was further annotated by TrasnDecoder.Predict to generate raw genome annotation [**Figure 3**].

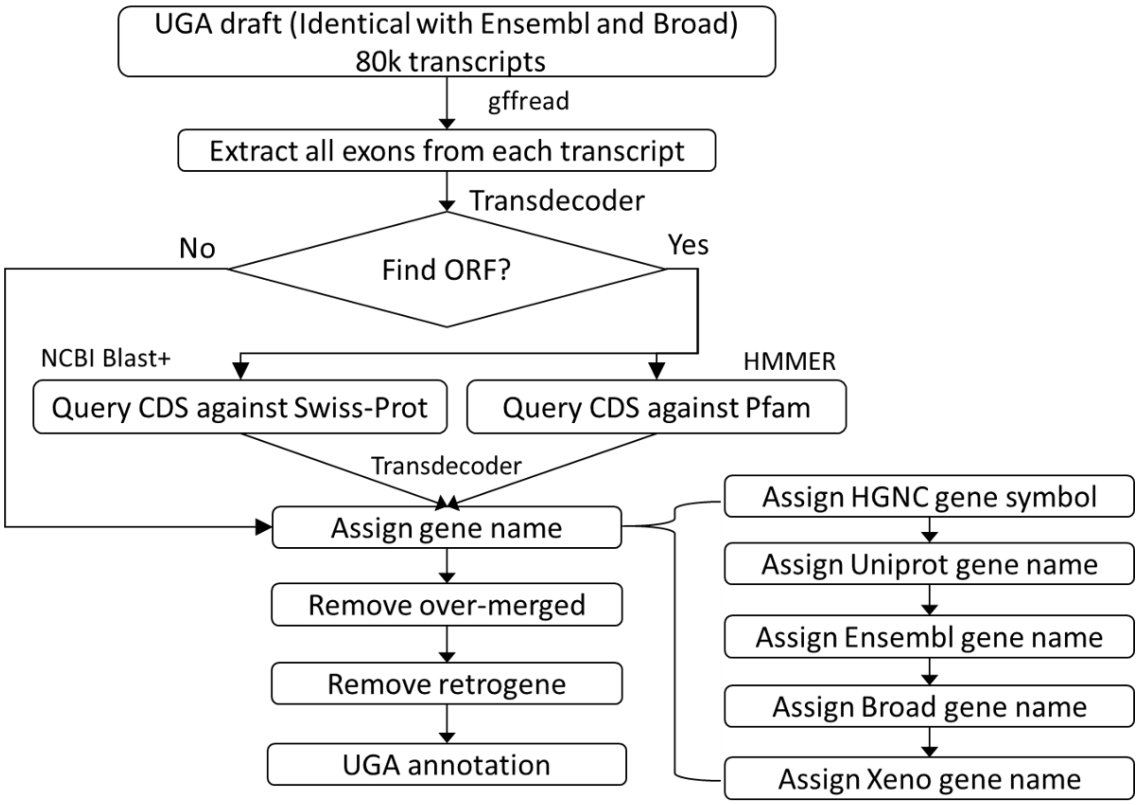

Figure 3: We built pipeline to maximally annotate each high confident transcript with official gene symbol

Over-merging discovery

Our preliminary analysis indicate that over-merging **Figure 4** is less prominent in the Ensembl and XenoRefGene databases. We hence first identified over-merged Ensembl or XenRefGene transcripts by comparing transcripts within each database and between the two databases following steps outlined in **Figure 5**. Then, we removed the over-merged transcripts from Ensembl or XenoRefGene database. For the Broad V1 and the UGA annotations, we compared transcripts within each database and also to transcripts of the Ensemsbl and XenoRefGene database after over-merging correction.

Our strategy for over-merging discovery is outlined in **Figure 5** and is briefly described as follows. First, a transcript, called "the parent transcript", was compared to each child transcript", defined as those transcripts that are on same chromosome, in the same strand and with at least one bp overlap with the parent transcript. If a parent transcript is found to span two or more child transcripts with two distinct gene symbols/names, then the parent transcript will be classified as an over-merged candidate. Then, a set of filter described below were applied to reduce false positives.

The first filter is unannotated transcripts. Specifically, if a child transcript gene name starts with "ENSCAF" (Ensembl IDs) or "CFRNASEQ_PROT" (Broad IDs) and without any official/common gene symbol/names assigned, these transcripts were removed from the over-merging candidates. The 2nd filter is known read-through genes. If the gene name of the parent transcript is composed of those of the first and second child transcripts linked by a "- ", e.g., SYS1-DBNDD2, it was then classified as read-through transcript. The 3rd filter is transcripts from complex gene families, including ZNF or OR genes. The 4th filter is delta, the distance between the CDS end of the head or first child and the CDS beginning of tail or 2nd child transcript [**Figure 6**]. If delta is negative, parent transcript could be a read-through transcript and will be removed from the over-merging transcript candidates.

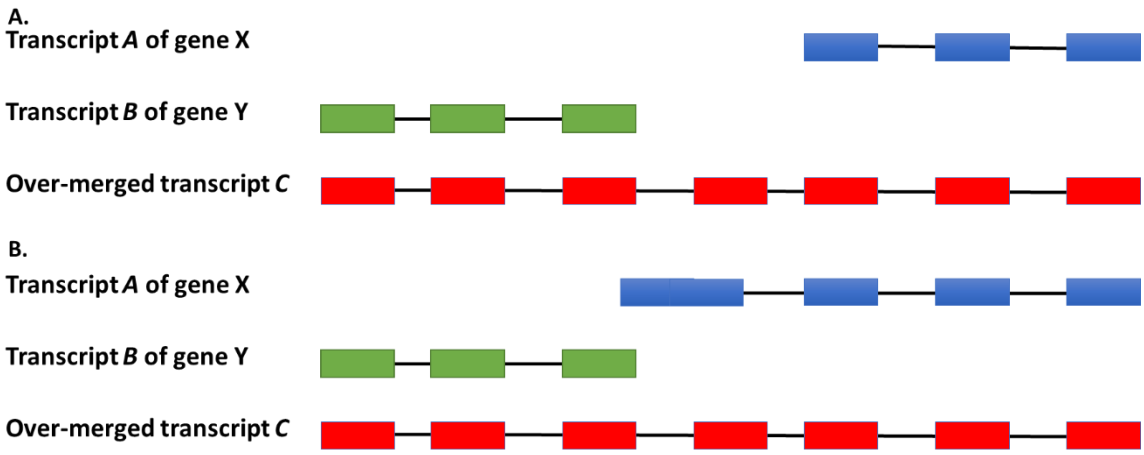

Figure 4: Over-merging definition. In case A, delta is negative, where in case B, delta is positive. Both cases are over-merging. (refer to Figure 6 for delta definition)

.

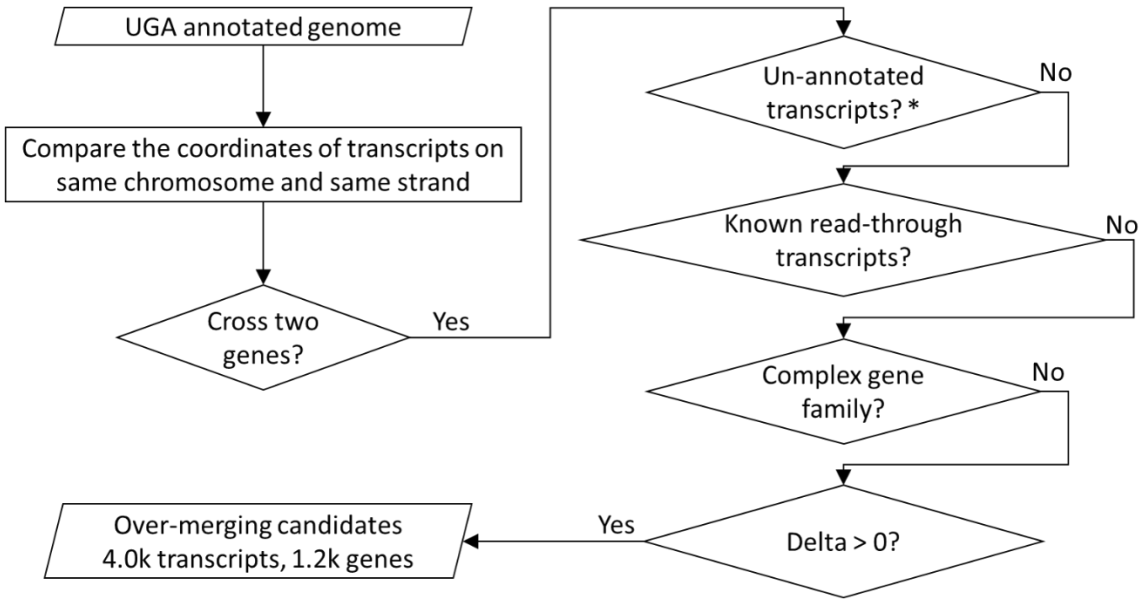

Figure 5: We built a pipeline for over-merged transcript identification

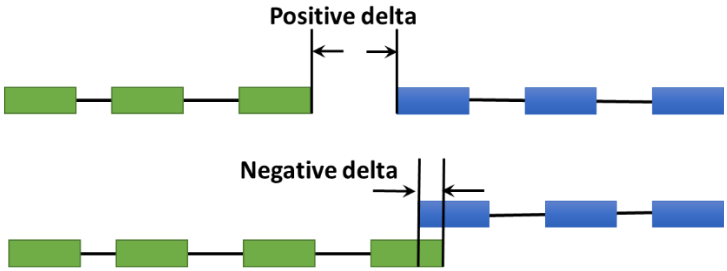

Figure 6: Definition of delta, the gap between transcripts on same chromosome and save strand

Retrogene identification

We followed the pipeline outlined in **Figure 8** for retrogene finding. A retrogene formed when as a processed (introless) transcript was reverse transcribed and reinserted back to the genome [**Figure 7**]. Briefly, we searched each single coding exon transcript against all multi-exon transcripts (after removing over-merged candidates) in each annotation database with NCBI blast using cutoffs of E ≤ 1e-5 (-max_target_seqs 1 -outfmt 6 -evalue 1e-5). If a multi-exon transcript match was identified as the top hit, we would examine their genomic coordinates. If no overlapping was identified, the single exon transcript would then be classified as retrogene.

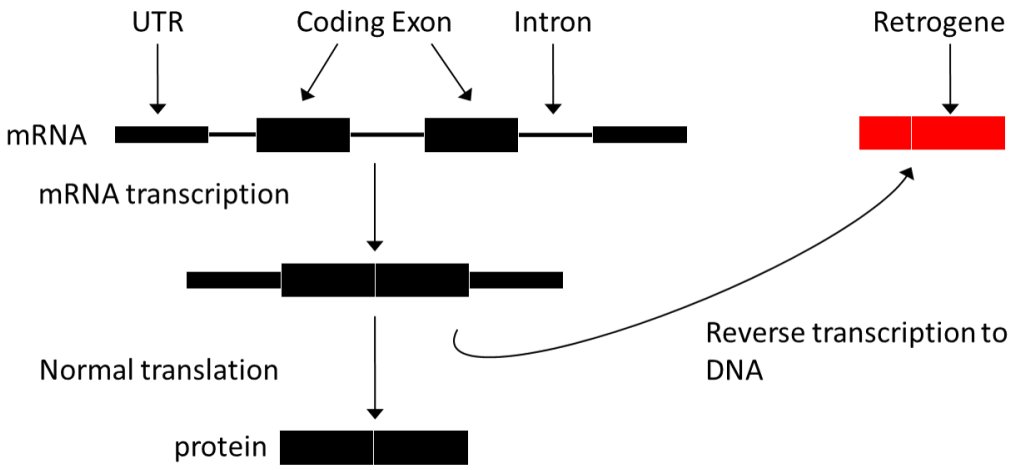

Figure 7: A retrogene formed when as a processed (introless) transcript was reverse transcribed and reinserted back to the genome

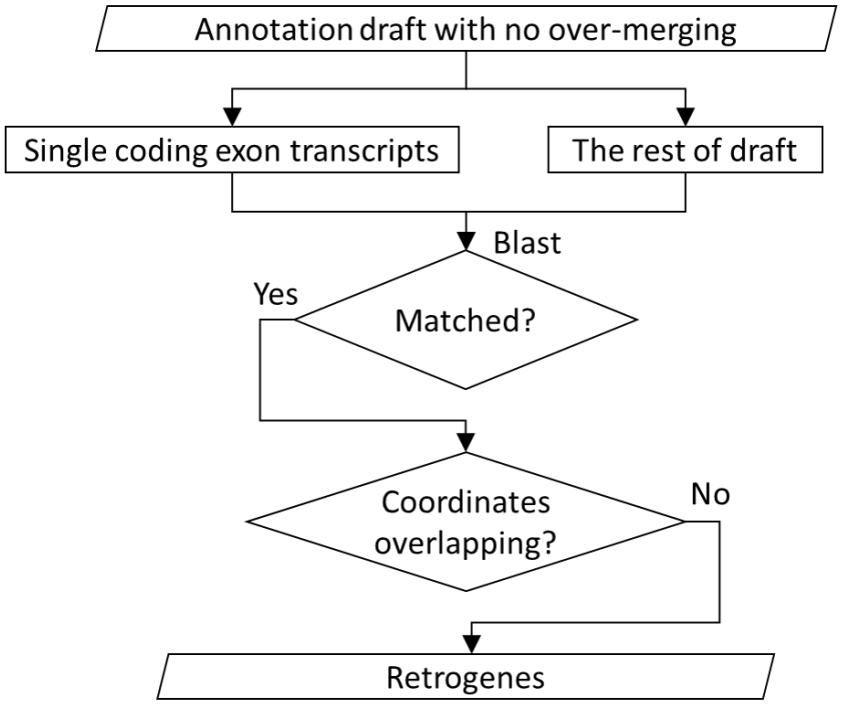

Figure 8: Retrogene identification pipeline,

Under-merging discovery

Under-merging is to describe the transcripts that are not complete transcribe or byproduct of incomplete splicing, or due to other reasons. It was a subjective concept. In this study, Ensembl and XenoRefGene were selected reference.

To rule out the overweight of over-merging, over-merging transcripts were excluded out in comparison. The transcripts from same gene were compared with each other by exon number and coding region length. To be identified as under-merging, transcripts had to meet following criteria [**Figure 9**]:

- The over-merging transcripts were ruled out during comparison.
- The exon number was under 30% of the median of exon number of transcripts with same gene name, on same chromosome and same strand.
- The length of transcript was under 30% of the median of length of transcripts with same gene name, on same chromosome and same strand.
- The exon number was under 30% of the median of exon number of transcripts on Ensembl with same gene name, on same chromosome and same strand.
- The length of transcript was under 30% of the median of length of transcripts on Ensembl with same gene name, on same chromosome and same strand.
- The exon number was under 30% of the median of exon number of transcripts on XenoRefGene with same gene name, on same chromosome and same strand.
- The length of transcript was under 30% of the median of length of transcripts on XenoRefGene with same gene name, on same chromosome and same strand.

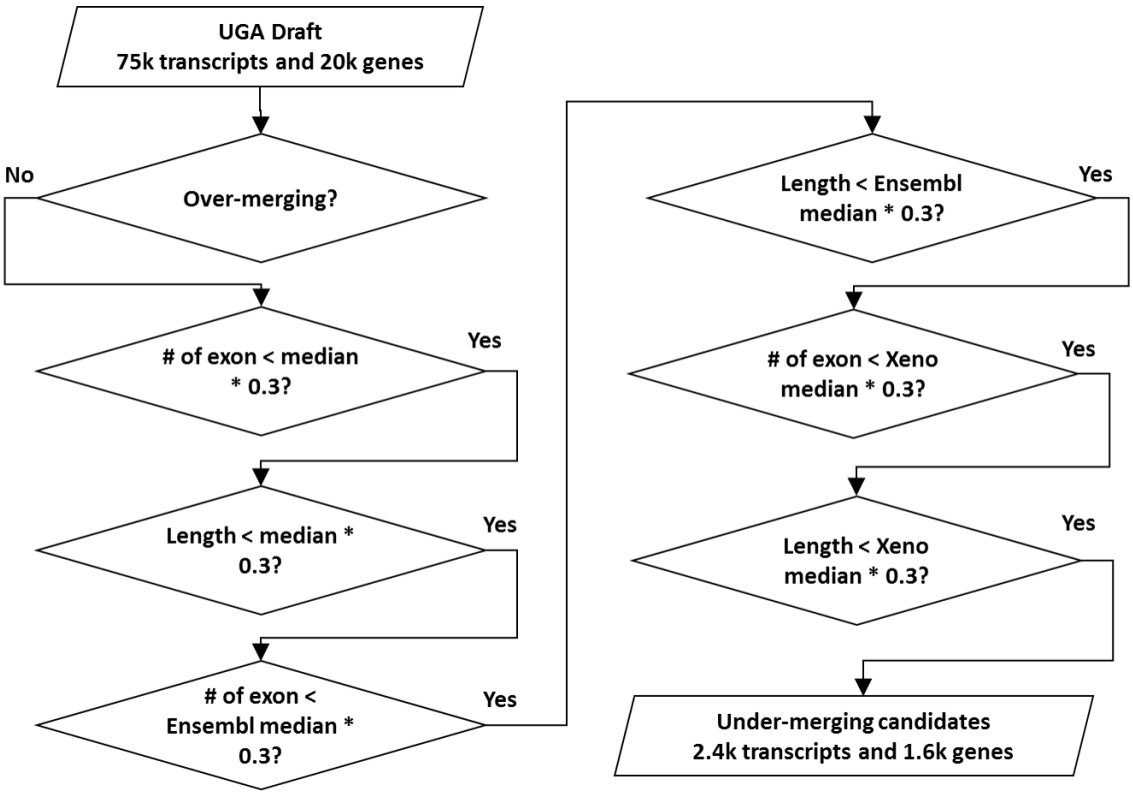

Figure 9: Under-merging identification pipeline

Human and dog alternative splicing comparison

To compare the abundance of alternative splicing forms between dog and human, synteny data of human GRCh38/hg38[32] to the dog genome were obtained from UCSC table canFam3.chainHg38[33]. The counts of transcripts and genes in the human genomic regions syntenic to each canine chromosome were obtained from UCSC hg38. wgEncodeGencodeCompV2[34]

CHAPTER 3 RESULTS

We identified several significant issues in current canine gene annotations.

We have identified a number of errors/issues in each of three canine annotation databases publically available, which include XenoRefGene, Ensembl and Broad as previously described. The first error is called "over-merging", resulting in a two distinct transcripts/genes fused into one transcript/gene. This is exemplified by as follows. Our analyses indicate that the chr1:112,621,382-112,645,509bp region encodes two distinct genes B9D2 and TGFB1, for which XenoRefGene has correctly annotated [**Figure 10**]. We have identified a number of errors/issues in each of the three canine annotation databases publically available, including XenoRefGene, Ensembl and Broad as previously described. The first error is called "over-merging", resulting in a two distinct transcripts/genes fused into one transcript/gene. This is exemplified by **Figure 10** as follows. Our analyses indicate that the chr1:112,621,382-112,645,509bp region encodes two distinct genes B9D2 and TGFB1, for which XenoRefGene has correctly annotated [**Figure 10 A**]. However, the two genes are mistakenly merged in both Ensembl and Broad annotations, yielding over-merged transcripts [**Figure 10 B** and **Figure 10 C**]. As a result, all transcripts in this region are incorrectly named as B2D9 in Ensembl [**Figure 10 B**]. or Q95N80_CANFA, the Uniprot protein ID for the canine TGFB1 [**Figure 10 C**]. Over-merging is not a rare event. We have identified a total of over-merged transcripts of 323 (1.1%) for Ensembl, 2563 (3.0%) for Broad, and 401 (0.8%) for XenoRefGene [**Table 1**].

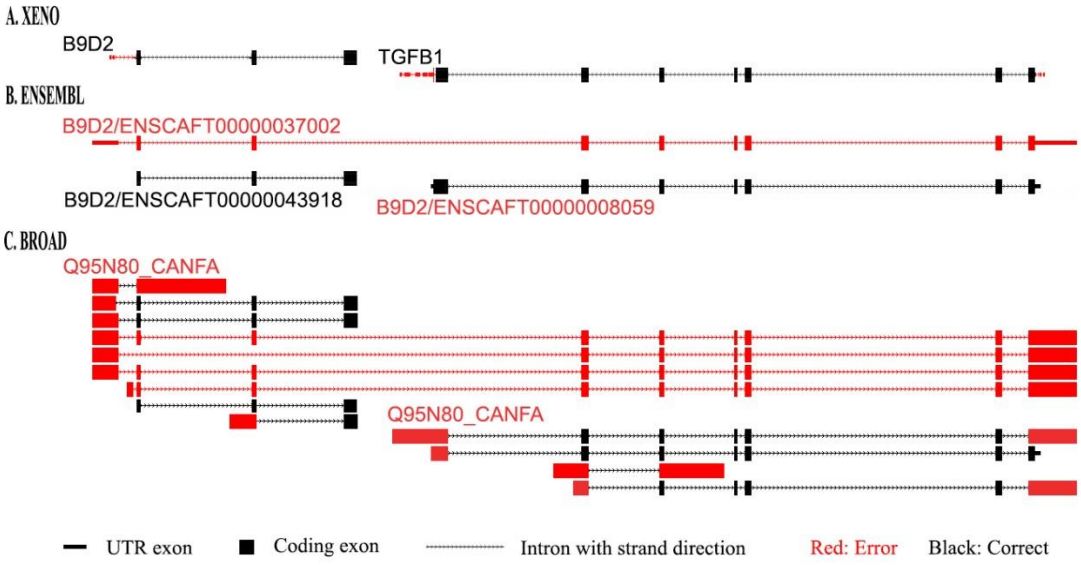

Figure 10: Examples of errors and issues identified in current canine annotation databases publically available.

Table 1: Summary of the issues identified in current canine gene annotations

	Count of transcripts							Count of UTR per transcript
	Transcript	**Gene**	**Transcript/gene**	**Un-annotated**	**Gene name error**	**CDS region error**	**Over-merging**	
UGA	70,511	17,361	4.14	-	-	-	3,950	0.38
Ensembl	29,881	24,580	1.27	6,557	547	5,657	323	0.41
Broad	85,168	16,265	5.24	2,897	15,244	46,308	2,563	0.92
XenoRefGene	53,824	21,619	2.49	-	45	1,073	401	9.78
Human	215,170	63,677	4.48	-	-	-	-	1.68
Mouse	104,129	43,629	2.38	-	-	-	-	0.78

* Human is GRCh38 **[32]**

** Mouse is GRCm38 [**35**]

The second error is genes being misnamed, some originated from over-merging as shown by Figure 1. In total, we identified how 547 (2%) Ensembl transcripts, 15,244 (18%) Broad transcripts, and 45 (0.1%) XenoRefGene transcripts having wrong gene names [**Table 1**].

The third error is mis-annotation of coding regions and untranslated regions (UTRs) of a transcript. This is especially evident in the Broad annotation, where many UTRs are labeled as coding regions, as exemplified by Figure 1. In total, we found 46,308 (55%) Broad transcripts, 5657 (19%) Ensembl transcripts, and 1073 (2%) XenoRefGene transcripts with incorrectly assigned coding regions [**Table 1**].

The forth error is fragmented UTRs, which is observed only in XenoRefGene transcripts [**Figure 10 A**]. This is caused by mapping human transcript sequences to the dog genome and UTRs are less conserved between the human and the dog, compared to coding regions. Because of the fragmentation, there are 9 UTR exons per transcript for XenoRefGene transcripts, compared to 0.4 UTR in Ensembl,0.9 in Broad, 1.67 for the human gene annotation and 0.78 for the mouse gene annotation.

The fifth issue is alternative splicing. While the Ensembl annotation has fewer transcripts with issues/errors described above, it has only1.02 transcripts per gene on average. This is significantly lower when compared to the human, with 4.48 transcripts per gene, and the mouse, with a 2.38 transcripts per gene [**Table 1**]. The Broad annotation has

increased the average transcript/gene ratio to 5.24 [**Table 1**]. Its total gene number is however only 16,265 [**Table 1**], with about 4000 genes missing in the Broad annotation.

The sixth issue is that a total of 6,557 (22%) Ensembl transcripts and 2,897 (3.4%) Broad transcripts are not annotated [**Table 1**]. These genes/transcripts are presented in the database with only Ensembl or Broad gene ID (e.g., transcript ENSCAFT00000043967 in Ensembl, which later is annotated as ENPP1 gene and CFRNASEQ_PROT_00011771, chr10:13,136,791-13,137,836), without official gene symbols identified.

We tried to build a more accurate canine gene annotation by addressing the issues/errors discussed above [**Figure 10** and **Table 1**]. To do this, we attempted first to build our own canine transcript assemblies, which then were compared with Ensembl and Broad transcripts to identify a set of transcripts with higher confidence [**Figure 1**]. Afterwards, we established an annotation pipeline that aimed to assign UTR/coding regions and gene symbol to each transcript accurately [**Figure 3**]. Finally, we attempted to identify and correct over-merged transcripts [**Figure 5**] and to discover retrogene [**Figure 8**] in each database. These steps are described below.

We established a set of transcripts of higher confidence through data integration.

We established a pipeline as outlined in **Figure 1** to build a set of canine transcripts with higher accuracy. To achieve this, we first assemble our own transcripts with our RNA-seq data from nine samples of normal canine skin and mammary gland [15, 36]. Each sample was sequenced by the Illumina platform to yield on average 56 million paired-end reads of 50bp [15, 36]. The sequences are all high quality based on the assessment by

FastQC [**Figure 1**]. With these sequences, we performed two transcript assemblies, reference-based and reference-independent [**Figure 1**].

For reference-based assembly, we first placed the sequence reads of each sample to the canine reference genome[18] with TopHat[25], achieving a mapping rate of 95% and alignment rate on concordant pairs of 84% on average [**Table 2**]. Then, we assembled the transcripts by Cufflinks[22] with the EnsEmbl 3.1.81 annotation [37] as the transcript reference. This yielded an average of 55,081 transcripts per sample [**Table 3** and **Figure 1**]. Then, we merged transcripts from the 9 samples with Cuffcmp, yielding a total of 382,409 distinct transcripts [**Table 3**, **Table 4, Table 6** and **Figure 1**]

For reference-independent assembly, we performed de novo assembly for each of the 9 samples using Trinity[38]. This analysis yielded on average 60,648 transcripts per sample [**Table 3**]. The transcripts were mapped to the dog reference genome with Blat[28] to identify their intron/exon regions [**Figure 1**]. We then merged the 9 assemblies via Cuffcmp, generating 89,831 distinct transcripts[**Table 3** and **Figure 1**].

We then combined the transcripts from reference-dependent and -independent assemblies via Cuffcmp. This leads to total of 381,181 distinct transcripts for our final assembly [**Table 3**, **Table 4** and **Figure 1**]. Analogous to Ensembl and Broad, we refer our own data as "UGA" hereafter.

We further merged our UGA transcripts with those of Ensembl and Broad via Cuffcmp, yielding 412,152 distinct transcripts in total [**Table 5** and **Figure 1**]. To increase accuracy, we identified those UGA transcripts with each share identical intron regions

with either an Ensembl or Broad transcript. A total of 79,887 such transcripts were found, with 15,441 with Ensembl/Broad, 56,455 with Broad and the rest, 7,991 with Ensembl [**Figure 11**]. These transcripts a higher confidence level and were subjected our annotation analysis described below.

Table 2: Tophat results summary

Tissue	Sample ID	Input reads	Aligned_pairs	multiple_alignments	discordant_alignments	concordant
Mammary	M-115N	27,597,117	23383202	2,387,341	1,132,799	80.60%
	M-31005N	21,738,806	19613143	2,286,047	1,012,215	85.60%
	M-32510N	28,106,024	25699198	3,674,402	1,421,487	86.40%
	M-401188N	32,599,348	28219439	4,484,354	1,825,093	81.00%
	M-406434N	23,925,378	21823117	2,184,139	852,634	87.60%
	M-76N	27,693,870	25083912	2,960,038	1,198,325	86.20%
Skin	S-240N	28,905,563	25895175	3,338,923	1,169,722	85.50%
	S-251N	31,045,511	28986137	4,263,531	1,473,895	88.60%
	S-465N	31,064,431	29078764	3,668,167	1,212,695	89.70%

Table 3: Results of Cufflink, trinity assemblies

Assembly	Tissue	Sample ID	# of Transcript	# of Exon
Reference	Mammary	115N	55,776	443,791
		31005N	55,140	444,126
		32510N	56,537	453,680
		401188N	57,670	430,548
		406434N	54,569	451,366
		76N	56,223	444,858
	Skin	240N	50,560	415,515
		251N	54,726	434,680
		465N	54,525	445,646
	Average		55,081	440,468
	All Combined		382,409	2,889,887
De novo	Mammary	115N	66,299	518,495
		31005N	63,344	450,594
		32510N	62,270	509,143
		401188N	61,662	485,378
		406434N	62,454	483,186
		76N	64,197	490,317
	Skin	240N	50,772	379,288
		251N	55,606	447,171
		465N	59,229	487,747
	Average		60,648	472,369
	All Combined		89,831	890,794

Table 4: Assembly results in categories from Cufflinks and Trinity

Class	# of Transcript	# of Exon
=	26,334	236,362
J	62,091	756,327
C	4,558	23,001
.	21,534	36,129
E	1,689	1,689
I	14,813	20,654
O	167,786	1,457,804
P	1,563	1,563
S	28,387	203,622
U	23,057	47,174
X	29,369	100,593
= and j	88,425	992,689
All Combined	381,181	2,884,918

* # Merged transcripts is less than total of cufflink and trinity result, due to that one result transcript could contain multiple original ones.

** In the assembly stage, gene ID is generated computationally. There are no tight biology senses on these gene ID numbers.

Table 5: Assembly result of combining UGA, Ensembl and Broad in category

Category	# of Transcript		
Class	UGA & Ensembl	UGA & Broad V1	UGA, Ensembl and Broad V1
=	24,396	76,651	**79,887**
J	106,820	5,259	
C	3,589	3,422	
.	673		
E	3,373	18,169	
I	13,028	49,158	
O	168,553	162,553	
P	1,439	781	
S	23,039	23,407	
U	38,337	30,084	
X	28,905	36,832	
= and j	131,216	81,910	
All Combined	412,152	406,316	

Table 6: Category explanation

Category details		
Priority	Class Code	Description
1	=	Complete match of intron chain
2	C	Contained
3	J	Potentially novel isoform (fragment): at least one splice junction is shared with a reference transcript
4	E	Single exon transfrag overlapping a reference exon and at least 10 bp of a reference intron, indicating a possible pre-mRNA fragment.
5	I	A transfrag falling entirely within a reference intron
6	O	Generic exonic overlap with a reference transcript
7	P	Possible polymerase run-on fragment (within 2Kbases of a reference transcript)
8	R	Repeat. Currently determined by looking at the soft-masked reference sequence and applied to transcripts where at least 50% of the bases are lower case
9	U	Unknown, intergenic transcript
10	X	Exonic overlap with reference on the opposite strand
11	S	An intron of the transfrag overlaps a reference intron on the opposite strand (likely due to read mapping errors)

12 . (.tracking file only, indicates multiple classifications)

* Categories definition is copied from Cufflink manual [22]

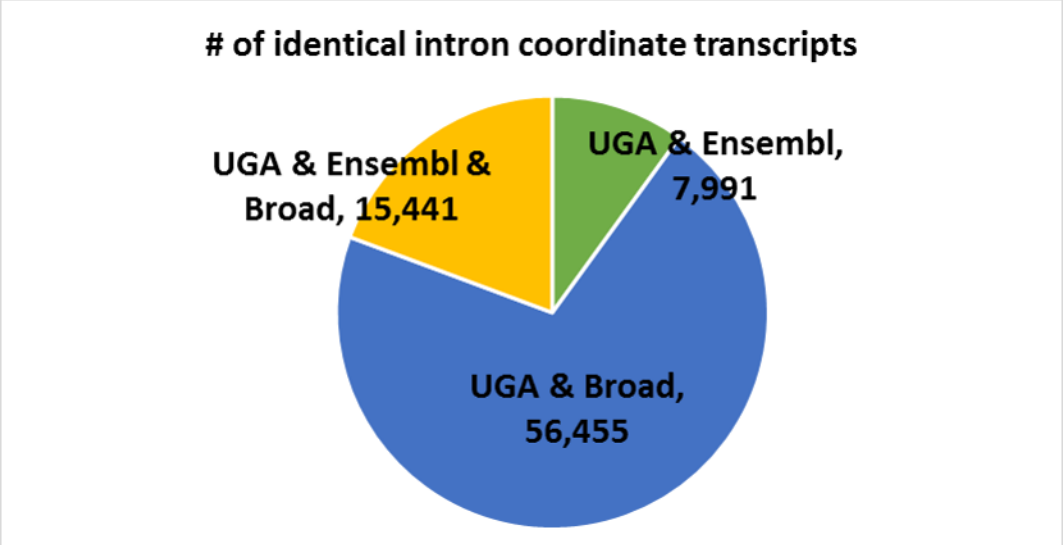

Figure 11: High confident transcripts consist of those UGA transcripts with each sharing identical intron regions with an Ensembl and/or Broad transcript.

Our annotation more accurately assigns coding regions and UTRs.

Following our annotation pipeline outlined in **Figure 3**, we first identified open read frames (ORFs) in each of the transcripts of high confidence with Transdecoder [29], a widely used tool. As a result, 66,447 transcripts were classified as the coding transcripts, with both coding DNA sequence (CDS) region and UTRs assigned [**Table 7** and **Figure 3**]. This analysis corrected many CDS miss-assignment in Broad transcripts [**Table 1**]. Among 62,605 coding transcripts that share identical intron regions with Broad transcripts, only 16,328 (26%) share identical CDS regions[**Table 7**], leading to a discrepancy rate of >74%. We investigated the 46,308 Broad transcripts with different CDS regions from ours. We found that 43,998 (95%) of them have their entire exonic

regions simply labeled as CDS, apparently without considering coding frame [**Table 7**]. This is significantly higher when compared to our UGA or Ensembl annotation, with 23,607 (51%) 4,409 (78%) such transcripts respectively [**Table 7**]. In Broad overall, there is total 63,113 (74%) transcripts with no UTR, which is also much higher than other parties.

Our CDS identification agrees better with the Ensembl and XenoRefGene annotation, with a discrepancy rate of 34% and 43% respectively [**Table 7**]. Finally, the same as both Ensembl and Broad annotations that are RNA-seq based[20], our analysis has effectively addressed the UTR fragmentation issue of XenoRefGene transcripts [**Figure 10**], with the UTR-exon/CDS-exon ratio decreasing from 10 to about 0.4 [**Table 8**].

Table 7: CDS region comparison

	Count of all transcript	Count of coding transcript				
	Total	**Total**	**Identical intron regions**	**Identical CDS intron regions**	**Different CDS intron regions**	**Different CDS intron regions with no UTR**
Broad V1	85,168	85,168	62,630	16,332	46,308	43,998
UGA	70,511	66,447	62,605	16,328	46,291	23,607
Ensembl	29,881	25,157	16,626	10,969	5,657	4,409
UGA	70,511	66,447	16,626	10,969	5,657	3,155
XenoRefGene	53,824	46,212	2,494	1,424	1,073	234
UGA	70,511	66,447	2,610	1,526	1,103	217

Table 8: Count of UTRs in transcripts

	Count of UTRs in all transcripts		
	Total	**transcript with no UTR**	**Average Count of UTR**
Broad V1	85,168	63,113	0.38
Ensembl	29,881	11,771	0.41
XenoRefGene	53,824	4,695	0.92
UGA	70,511	4,356	9.78
Human	100,778	4,927	1.68
Mouse	53,819	4,765	0.78

Our gene name and transcript annotation are standardized and more accurate.

We annotate the gene name of each coding transcript by blasting its CDS against the UniProtKB/Swiss-Prot database[30], which contains 552,000 manually annotated non-redundant proteins from various species [**Figure 3**] . For each transcript, we selected the best match that met our cutoff (E< 1-E5). Then, we assign the gene name from the match to each transcript. To maximally standardize the gene names, we first prioritize the use of official gene symbols approved by Human Genome Organization (HUGO) Gene Nomenclature Committee [39] [**Figure 3**] . Then, for those transcripts without official symbol assignment, we adopt the names available in the order other gene name alias, UniProt IDs, the Enseml gene IDs, and Broad gene IDs [**Figure 3**].

We were able to assign 22,465 gene names to 66,447 transcripts [**Figure 12**]. Among these gene names, 13,066 (~64%) are official gene symbol[39].

Our annotation contains 3,443 gene names missing in the Ensembl database, and 6,022 gene names missing in the Broad database [**Table 10**]. The difference may arise from annotation errors or different naming scheme. This illustrated by Figure 1, which indicates that the gene name "TGFB1" is missing in Ensembl because of over-merge and in Broad because of the use of the UniPort identifier (i.e., Q95N80_CANFA), instead of official gene symbol (i.e., TFGB1) [**Table 1**] .

Conversely, we found that 5,769 gene names in Ensembl are missing in our annotation [**Table 10**]. Among these gene names, 4,372 are simply Ensembl gene IDs, all beginning with ENSCAFG (e.g., *ENSCAFG00000000002*). As described a later section, a

substantial fraction (at least 4097) of these genes are retrogenes. The remaining gene names missing our database include those like *c1orf27* (330 in total), which are under more constant updating with more meaningful gene names. Members from complex gene families, e.g., olfactory receptor (OR) and zinc finger genes also explain the difference [**Table 11**] These genes also explains a large portion of gene names missing in our annotation, when compared to Broad and XenoRefGene databases.

Our transcripts are named in the pattern of gene-name_chromosome_strand_duplication, such as transcripts of SCUBE2 at chromosome 21, forward strand would be named as *SCUBE2_21_+_dup_0 and SCUBE2_21_+_dup_1* etc.. This naming convention provides more information by the name of transcript.

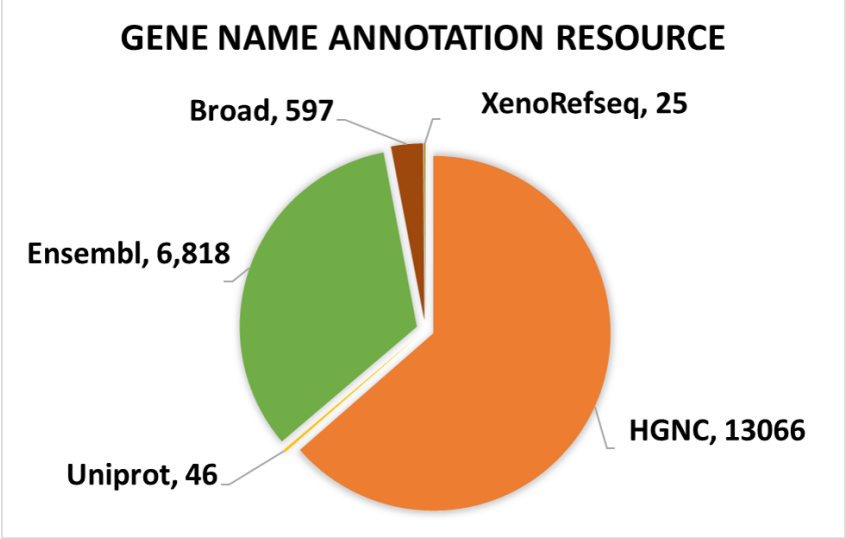

Figure 12: Gene name annotation resource

Table 9: Categories of gene name errors/flaws in Broad and Ensembl

Resource	Category	# of Transcript	# of Gene Name
Broad	Wrong name	11,739	3,001
	(human)	299	290
	_CANFS	3206	652
	CFRNASEQ_PROT	903	903
Ensembl	Wrong name	547	477
	ENSCAF	6,557	5,980

Table 10: Coding gene name comparison

	Count of Coding Gene Name					
	Total	**Not in UGA**	**Not in Ensembl**	**Not in Broad**	**Not In XenoRefGene**	**Not in Human**
UGA	17,526	-	3,443	6,022	2,047	1,971
Ensembl	19,852	5,769	-	6,239	4,583	4,475
Broad V1	18,562	4,692	2,583	-	3,955	3,799
XenoRefGene	16,266	3,083	3,293	6,321	-	221
Human	20,326	5,769	4,949	7,929	1,985	-

Table 11: Missed gene name compared in naming pattern

	Counts of gene names					
Pattern ***	**in Ensembl not UGA ********	**in UGA not Ensembl**	**in XenoRefGene not UGA**	**in UGA not XenoRefGene**	**in Broad not UGA**	**in UGA not Broad**
*** ENSCAF**	4372	-	-	-	1994	-
**** C.orf**	330	220	419	220	410	220
ZNF	-	-	184	-	-	109
OR	20	-	141	-	-	239
KRTAP	83	-	83	-	-	38
HIST	12	-	55	-	-	-
FAM	-	-	44	-	-	-
LOC	-	-	38	-	-	-
USP	-	-	31	-	-	-
PCD	-	-	36	-	-	-
SLC	-	-	-	-	-	83
OLFR	-	14	-	14	-	14
***** CFRNASEQ_PROT**	-	-	-	-	903	-

* ENSCAF	Ensembl in-house named
** C.orf	C[digit]orf
*** CFRNASEQ_PROT	Broad V1 in-house named

**** in column title, E stand for Ensembl, U stand for UGA, B stand for Broad, e.g. in E not U, in Ensembl, but not in UGA

***** Pattern: Gene Name prefix or gene name pattern

We developed a pipeline to identify over-merged transcripts.

As previously defined and illustrated in **Figure 4**, an over-merged transcript is a fusion product of two transcripts from two neighboring but distinct genes. Based on this, we developed a pipeline as outlined in **Figure 5** for over-merged transcript identification. Briefly, we first identified those transcripts that span two neighboring genes by comparing transcript among those that are on the same chromosome, in the same strand, and overlapped by at least 1bp [**Figure 5**]. As a result, a total of 5913 transcripts in our annotation, 597 transcripts in Ensembl, 4,533 transcripts in Broad, and 5626 transcripts in XenoRefGene were classified as over-merged candidates [**Table 12**]. We performed several analyses to reduce false positives as described below.

We first identified those over-merged candidates whose child transcript(s) is (are) functionally not annotated [**Figure 5**]. "Functionally unannotated" here refers to those transcripts whose gene names are merely Ensembl or Broad IDs starting with ENSCAFG or CFRNASEQ_PROT, with no official gene symbols or other biological more meaningful names associated. These transcripts have a higher probability to be pseudo-genes (e.g., retrogenes described later), we hence excluded over-merged candidates involving these transcripts from out list [**Figure 5** and **Table 12**].

We then determined how many are already known "read-through transcripts". Read-through transcripts are produced during transcription by combining exonic regions of two distinct neighboring genes encoded in the same strand[40]. Some read-through transcripts have already been experimentally validated[40]. Among our putative over-

merged transcripts, a total of 74 in our annotation, 54 in XenoRefGene, one in Broad, and zero in Ensembl were found to be known read-through transcripts [**Table 12**]. Because read-through transcripts are considered truly transcribed products, we excluded them from our over-merged transcript candidates [**Figure 5**]. The list of read-through genes are presented at **Table 13**.

We studied the remaining over-merged transcript candidates, and found many involve large gene families such as ZNF and OR [**Table 12**]. Because these gene families are still evolving, many members share high sequence identities [**Figure 5**]. As a result, it is difficult to accurately assign gene names to those members as our method is sequence identity based. To avoid ambiguity and reduce false results, we excluded over-merged candidates involving these gene families [**Table 14**]. As a consequence, a total transcripts of 1651 for our annotation, 172 for Ensembl, 762 for Broad, and 2967 for XenoRefGene were removed [**Table 12**]. XenoRefGene has the largest number (2967) because it contains the most ZNF genes.

To better understand transcript over-merging, we examined the genomic distance or gap, represented by delta [**Figure 6**], between the genes of the two child transcripts. A negative delta indicates the two genes overlapping in their exons, while a positive delta value indicates otherwise. We found that after complex gene family removal as described above [**Figure 5**], the over-merged candidates with delta < 0 in total are 238 for our annotation, 16 for Ensembl, 62 for Broad, and 38 for XenoRefGene [**Table 12**].

Thus, a majority of over-merged candidates are from two neighboring but non-overlapping genes [**Figure 13**]. For over-merged candidates involving complex gene families, however, the delta values are mostly negative [**Figure 6**]. Compared to other annotation databases, XenoRefGene has more positive data values, possibly because the UTR exons of XenoRefGene are significantly shorter [**Figure 10**]. Because of these observations, we removed those with delta < 0 from our over-merged transcript candidate lists to reduce false positives.

In summary, excluding known read-through transcripts and those involving complex gene families or with delta < 0, we discovered 3950 potential over-merged transcripts in our annotation. Similarly, we identified such transcripts of 323 for Ensembl, 2563 for Broad, and 401 for XenoRefGene [**Table 12**]. As a result, we corrected the mis-annotated gene names of 50 Ensembl transcripts, 1851 Broad transcripts, and 1 XenoRefGene [**Table 12**]. These mis-annotations originated from transcript over-merging, as exemplified in **Figure 10**.

Table 12: Over-merging filters with number of transcripts at each step

				Complex						
Annotation	**Cross**	**Un-annotated**	**Read-through**	**CEACAM**	**OR**	**Zinc_ Finger**	**other_complex**	**Complex Subtotal**	**Negative delta**	**Over-merging**
UGA	5,913	-	74	29	30	464	1,128	1,651	238	3,950
Ensembl	597	86	NA	8	3	43	118	172	16	323
Broad V1	4,533	1,145	1	16	18	293	435	762	62	2,563
XenoRefSeq	5,626	2,166	54	54	174	1,894	845	2,967	38	401

Table 13: List of read through genes

Read-through gene name
ANKHD1-EIF4EBP3
ARPC4-TTLL3
ATP5J2-PTCD1
BCL2L2-PABPN1
BIVM-ERCC5
C7orf55-LUC7L2
CCDC169-SOHLH2
CHURC1-FNTB
CKLF-CMTM1
COMMD3-BMI1
CORO7-PAM16
FPGT-TNNI3K
FXYD6-FXYD2
HSPE1-MOB4
IQCJ-SCHIP1
ISY1-RAB43
JMJD7-PLA2G4B
LY75-CD302
MSANTD3-TMEFF1
NEDD8-MDP1
NME1-NME2
NT5C1B-RDH14
PALM2-AKAP2
PMF1-BGLAP
PPAN-P2RY11
PRR5-ARHGAP8
PTGES3L-AARSD1
RBM14-RBM4
RNF103-CHMP3
RPS10-NUDT3
SAA2-SAA4
STON1-GTF2A1L
SYNJ2BP-COX16
SYS1-DBNDD2
TGIF2-C20orf24
TMED7-TICAM2
TMEM189-UBE2V1
TNFSF12-TNFSF13
TRIM39-RPP21

TRIM6-TRIM34

Table 14: List of complex gene name prefix or patterns

Gene name pattern
ADGRE2
ALPK3
ATP6V0E2
C1*orf*
CEACAM1
DRD4
DUT
ECSIT
ENSCAFG*
FBXL19
FIZ1
GCKR
GIMAP8
HKR1
KRBOX*
MAN2B1
MZF1
OR10A4
PSG*
Q004B0_CANFA
SAMD1
SQLE
SRCAP
TRIM*
VN1R*
ZFP*
ZIK1
ZIM3
ZKSCAN3
ZNF*
ZSCAN12

*is the wild card for the number or letters

Table 15: gene name errors affected by over-merging

	# of Transcript	# of Gene
Ensembl	50	44
Broad	1851	471
XenoRefGene	1	1

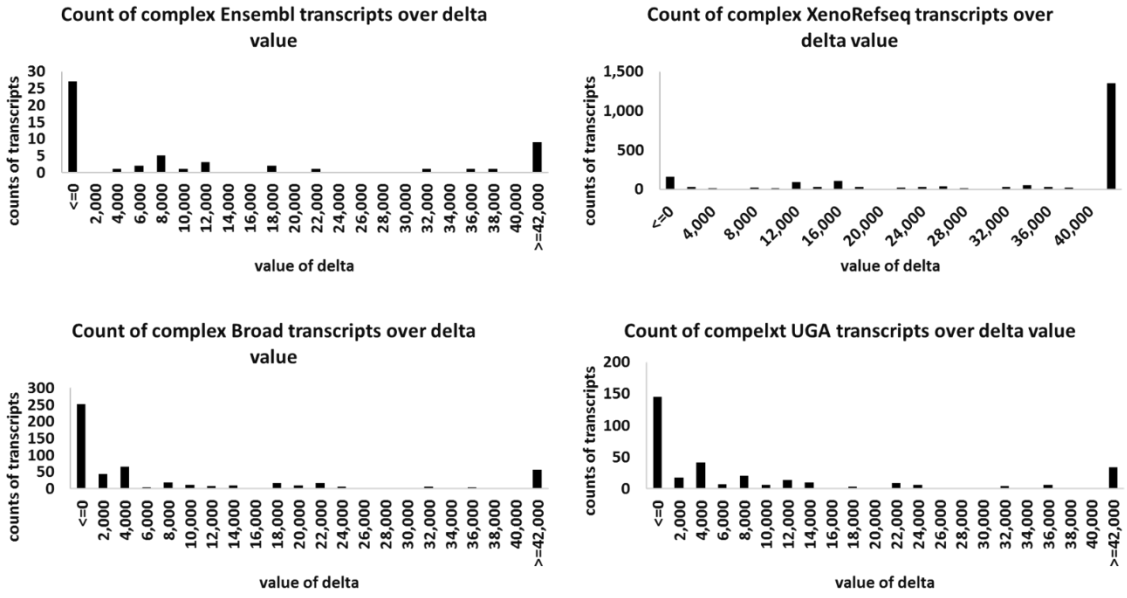

Figure 13: Transcript density over delta, value of delta 0 includes the count of transcripts with negative delta, max value of delta includes counts of transcripts more than max value

We established a pipeline to identify retrogenes.

After removing over-merged transcripts described above, our annotation contains 77645 transcripts from 22703 genes. Among these transcripts, about 10% (7663 in total) are from single exon genes, which is significantly higher when compared to the human genome gene annotation (4%). To determine if some of the single exon transcripts are

from retrogenes, defined as processed (intronless) transcripts being reverse transcribed to DNA and reinserted back to the genome [**Figure 7**], we developed a pipeline as outlined in **Figure 8**. Briefly, we searched each single exon transcript against each multi-exon transcript in the database. If a match was found (see Methods) and the single exon transcript and the multi-exon transcript differ in genomic location, the single exon transcript will be classified as representing a retrogene [**Figure 7**].

Through this analysis, we classified 2984 transcripts as retrogene candidates in our annotation. Similarly, we classified 4290 Ensembl as retrogenes, 1650 Broad as retrogenes for Broad, and 1965 XenoRefGene transcripts retrogenes [**Table 16**]. The retrogene exon size is 5-6 times larger than the overall gene exon size on average [**Figure 15** and **Table 17**]. Furthermore, among the 4299 Ensembl retrogene candidates identified, 4097 (95.5%) have no official gene symbols and are only assigned an Ensembl ID starting with ENSCAF, as previously described [**Figure 14**]. These observations support that our pipeline [**Figure 8**] is valid.

After removing the identified retrogene candidates, the single exon gene fraction decreased from 21% to 9% for Ensembl. The modified numbers better match that of the human annotation (GRCh38), which contain about 4% single exon transcripts [**Table 18**].

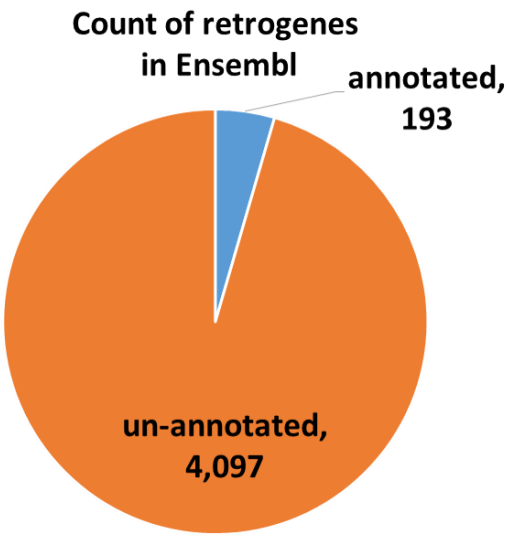

Figure 14: Ensembl retrogene composition, annotated and un-annotated

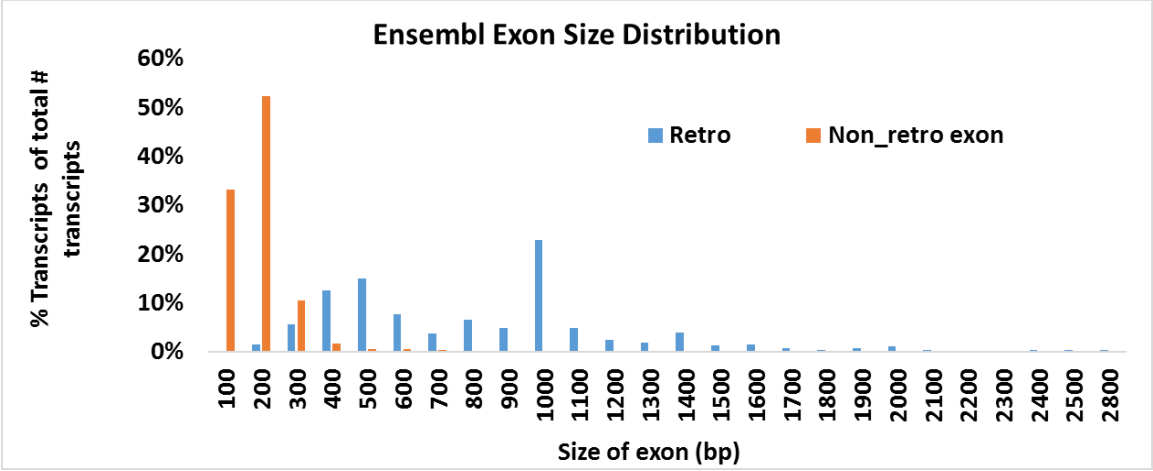

Figure 15: exon size distribution of retrogene and non-retro gene

Table 16: Count of retrogene

Count of retrogene	
UGA	2,984
Ensembl	4,290
Broad V1	1,650
XenoRefSeq	1,965

Table 17: Comparison of exon size and retrogene size

	Ensembl		Broad		Xeno		UGA	
	Overall	**Retrogene**	**Overall**	**Retrogene**	**Overall**	**Retrogene**	**Overall**	**Retrogene**
Average exon size(bp)	166.56	806.88	385.40	2,447.60	149.34	660.37	165.16	1,028.33

Table 18: Single exons and retrogenes

	Total #	Single Exon #	Single Exon / Total %	Total #	Single Exon #	Single Exon / Total %
UGA	74,024	3,513	4.75%	70,511	529	0.75%
Ensembl	29,881	6,419	21.48%	23,324	2,129	9.13%
Broad V1	85,168	1,950	2.29%	81,071	300	0.37%
XenoRefSeq	53,824	2,632	4.89%	51,364	667	1.30%
Human	215,170	7,892	3.67%	215,170	7,892	3.67%

We tried to identify potential under-merging transcripts

Similar to over-merging described previously, under-merging transcripts (i.e., missing certain exons) are also possible for reasons such as inadequate sequencing of lowly transcribed genes in RNA-seq analysis. However, under-merging discovery is not as straight forward as over-merging, because it is difficult to distinguish short but real alternative splicing forms from under-merged transcripts. We attempted to reply on exon number and length cutoffs, as outlined in **Figure 9**, for this task. With this pipeline, we discovered 2,442 transcripts that are possibly under-merged in our database. We also identified such transcripts of 20 for Ensembl, of 1506 for Broad, and of 113 for XenoRefGene [**Table 19** and **Figure 9**].

Table 19: Count of under-merging transcript and gene

	count of under-merging	
	# of Transcript	**# of Gene**
UGA	2,442	1,580
Ensembl	20	20
Broad V1	1,506	1,208
XenoRefGene_hg	113	93

We provided an improved canine gene annotation to the public.

After addressing issues described above, our annotation contains 70,511 transcripts from 17,361 genes, among which 66,447 transcripts from 16062 gene are protein coding [**Table 20**]. As previously described, each transcript shares identical intron/exon junction an Ensembl or Broad transcript [**Figure 11**]. Our annotation contains 12 exons per transcript and 4 transcripts per gene on average, with more alterative splicing forms compared to the Ensembl database. We studied the transcript/gene ratio distribution in

the canine genome and found that the distribution resembles that of the human transcript/gene ratio in the syntenic human genomic regions [**Figure 16**]. This support the overall accuracy of our alternative splicing form identification.

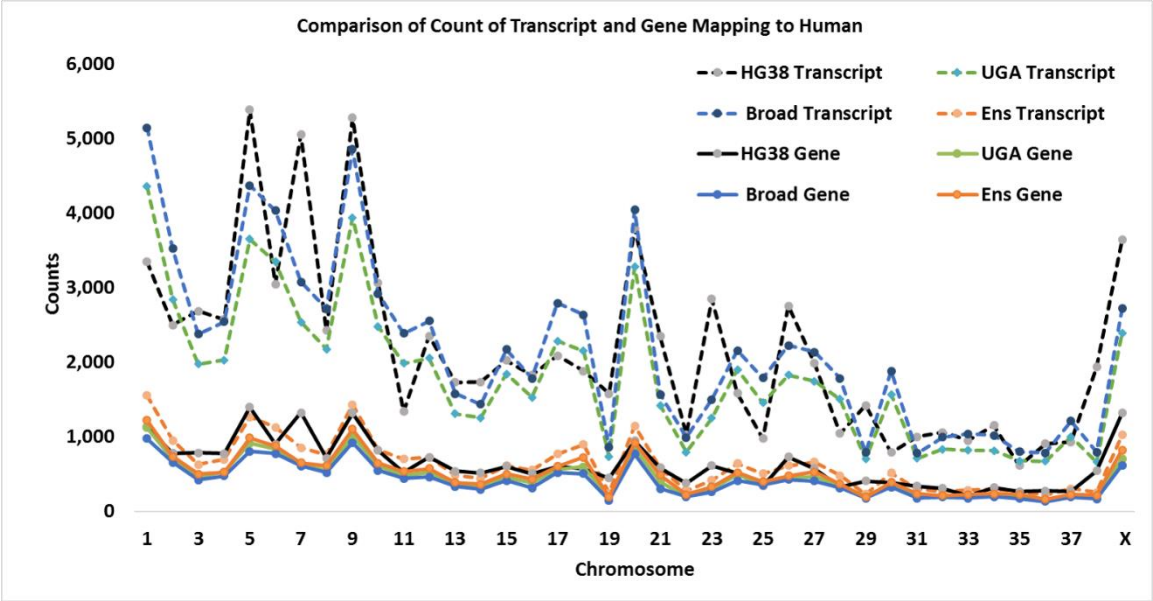

Figure 16: Transcript and gene distribution over chromosome

Data access

Our annotation, along with the identified over-merging genes and retrogenes in each database are publically available at the UCSC genome website.

http://genome.ucsc.edu/cgi-bin/hgTracks?hgS_doOtherUser=submit&hgS_otherUserName=yatchin&hgS_otherUserSessionName=N9_v2_omf_ready%2D0

Table 20: Transcript over gene ratio

	All			CDS		
	Transcript #	**Gene #**	**Transcript/gene**	**CDS transcript**	**CDS gene**	**Transcript/gene**
UGA	70,511	17,361	4.06	66,447	16,062	4.14
Ensembl	29,881	24,580	1.22	25,157	19,856	1.27
Broad	85,168	16,265	5.24	85,168	16,265	5.24
XenoRefSeq	53,824	21,619	2.49	46,212	18,561	2.49
Human *	215,170	63,677	3.38	104,763	23,393	4.48
Mouse **	104,129	43,629	2.39	53,819	22,621	2.38

* Human is GRCh38

** Mouse is GRCm38

CHAPTER 4 DISCUSSION

We have improved canine gene annotation by addressing issues in current databases.

We have diligently tried to address the issues, which we identified in current publically available canine gene annotation databases. As a result, we have built an annotation, which we believe is significantly improved in several aspects. First, to increase the accuracy of canine transcript assemblies, we performed reference-dependent and independent transcript assembly with our high quality RNA-seq data[15, 16] and only selected those that share identical intron/exon junctions with existing Ensembl or Broad transcripts. Second, we developed a pipeline that aims to accurately assign coding regions and an official gene symbol to each transcript. In this process, we corrected hundreds to thousands of misnamed transcripts or transcripts named not using official gene symbols. Third, we developed a strategy that aims to identify over-merged transcripts. For RNA-seq based annotation, over-merging happens because all transcript assembly programs are sequence match-based and RNA-seq reads are still relatively short (<150bp). Thus, if the end of the last exon of one gene share high sequence identify with the beginning of the first exon of a neighboring gene, an over-merged transcript could arise. For XenoRefGene, the identification of gene in the canine genome is also sequence-match based, and the same issue could arise. Our study corrected hundreds to thousands of over-merged transcripts in each annotation database, and as well as the misnamed genes originated from over-merging. Lastly, we developed a strategy that discovered thousands of retrogenes in each database.

Our annotation is improved. Compared to the Broad annotation[20], we have corrected transcripts with incorrectly labeled CDS regions, over-merged, and/or misnamed. Compared to the Ensembl annotation, we have classified thousands of transcripts as retrogenes and increased the transcript/gene ratios from 1 to 4. Compared the XenoRefGene annotation, we corrected the UTR regions of nearly every coding transcripts and increase the accuracy of CDS regions of many transcripts. We have released this improved database to the public.

Our pipelines need to be improved.

First, for over-merging discovery, our current pipeline only considers those neighboring but non-overlapping transcripts to avoid read-through transcripts. However, over-merging could happen with overlapping genes. We hope that with further understanding of over-merging; we could use more accurate criteria to distinguish between read-through and over-merging transcripts.

Second, we currently only focus on the transcripts with identical intron region with existing Ensembl or Broad transcripts. There are >342,000 transcripts not included in our analysis. Granted, many of them are artifacts, e.g., transcript fragments or under-merged transcripts (for discussion below). Some, however, could be novel alternative splicing forms or novel genes. Further efforts are required to develop strategy to classify these transcripts.

Several critical issues are not considered by our pipelines.

Although being improved, our canine annotation still lags behind in accuracy and comprehensiveness, when compared to the human genome annotation. Several critical issues are not addressed in our current pipelines.

First, like over-merging, under-merging transcripts are possible and hence the transcript/gene ratio of four in our databases may be inflated. On contrary to over-merging, under-merging transcripts miss certain exons). They occur for reasons such as inadequate sequencing of lowly transcribed genes in RNA-seq analysis. However, under-merging discovery is not as straight forward as over-merging, because it is difficult to distinguish short but real alternative splicing forms from under-merged transcripts. Future studies are need to establish efficient strategy for under-merged transcript discovery.

Second, tissue specific alternative splicing is nearly unknown for the dog. As more dog tissues and cells being sequenced in the future, this situation will be improved.

Third, the dog genome is in a draft state, contain assembly, and sequence errors. This in turn could result in gene mis-annotation, e.g., missing exons, CDS region mis-assignment. However, this however requires manual curations in many cases. An automatic pipeline can only be established once we have a better understanding of the errors and RNA-seq data and software.

Fourth, regulatory regions (e.g., enhancers) are less conserved between human and dogs. Hence, these regions need more species-specific efforts, e.g., ChIP-seq with makers specific to enhancers. Once published, we will incorporate these studies for regularity regions findings. For noncoding RNA genes fortunately, other groups have taken efforts for their discovery[41].

Our pipelines should be useful to other mammalian species.

Except for the human, the mouse and perhaps the rat, the other mammalian species share many of the same gene annotation issues as the dog. Also similar to the dog, RNA-seq data are being generated for these species. Hence, our pipelines developed are species independent and should be useful to these species.

APPENDICES

1. Khanna, C., et al., *The dog as a cancer model.* Nat Biotechnol, 2006. **24**(9): p. 1065-6.
2. Neff, M.W. and J. Rine, *A fetching model organism.* Cell, 2006. **124**(2): p. 229-31.
3. Parker, H.G., A.L. Shearin, and E.A. Ostrander, *Man's best friend becomes biology's best in show: genome analyses in the domestic dog.* Annu Rev Genet, 2010. **44**: p. 309-36.
4. Rowell, J.L., D.O. McCarthy, and C.E. Alvarez, *Dog models of naturally occurring cancer.* Trends Mol Med, 2011. **17**(7): p. 380-8.
5. Boyko, A.R., *The domestic dog: man's best friend in the genomic era.* Genome Biol, 2011. **12**(2): p. 216.
6. Meuten, D.J., *Tumors in domestic animals*. 4th ed. 2002, Ames, Iowa: Iowa State University Press. xii, 788 p.
7. Paoloni, M. and C. Khanna, *Translation of new cancer treatments from pet dogs to humans.* Nat Rev Cancer, 2008. **8**(2): p. 147-56.
8. Gordon, I., et al., *The Comparative Oncology Trials Consortium: using spontaneously occurring cancers in dogs to inform the cancer drug development pathway.* PLoS Med, 2009. **6**(10): p. e1000161.
9. Tang, J., et al., *Copy number abnormalities in sporadic canine colorectal cancers.* Genome Res, 2010. **20**(3): p. 341-50.

10. Youmans, L., et al., *Frequent alteration of the tumor suppressor gene APC in sporadic canine colorectal tumors.* PLoS One, 2012. **7**(12): p. e50813.

11. Tang, J., et al., *Cancer driver-passenger distinction via sporadic human and dog cancer comparison: a proof-of-principle study with colorectal cancer.* Oncogene, 2013.

12. Nasir, L., et al., *Telomere lengths and telomerase activity in dog tissues: a potential model system to study human telomere and telomerase biology.* Neoplasia, 2001. **3**(4): p. 351-9.

13. Rangarajan, A. and R.A. Weinberg, *Opinion: Comparative biology of mouse versus human cells: modelling human cancer in mice.* Nat Rev Cancer, 2003. **3**(12): p. 952-9.

14. Li, Y., et al., *Cancer driver candidate genes AVL9, DENND5A and NUPL1 contribute to MDCK cystogenesis.* Oncoscience, 2014. **1**(12): p. 854-65.

15. Liu, D., et al., *Canine spontaneous head and neck squamous cell carcinomas represent their human counterparts at the molecular level.* PLoS Genet, 2015. **11**(6): p. e1005277.

16. Liu, D., et al., *Molecular homology and difference between spontaneous canine mammary cancer and human breast cancer*. Cancer Res, 2014.

17. Tang, J., et al., *Cancer driver–passenger distinction via sporadic human and dog cancer comparison: a proof-of-principle study with colorectal cancer.* Oncogene, 2014. **33**(7): p. 814-822.

18. Lindblad-Toh, K., et al., *Genome sequence, comparative analysis and haplotype structure of the domestic dog.* Nature, 2005. **438**(7069): p. 803-19.

19. Derrien, T., et al., *Annotation of the domestic dog genome sequence: finding the missing genes.* Mammalian Genome, 2012. **23**(1-2): p. 124-131.

20. Hoeppner, M.P., et al., *An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts.* PLoS One, 2014. **9**(3): p. e91172.

21. Venter, J.C., et al., *The sequence of the human genome.* science, 2001. **291**(5507): p. 1304-1351.

22. Trapnell, C., et al., *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nat Biotechnol, 2010. **28**(5): p. 511-5.

23. Andrews, S., *FastQC: a quality control tool for high throughput sequence data*. 2010.

24. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.

25. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.

26. Roberts, A., et al., *Improving RNA-Seq expression estimates by correcting for fragment bias.* Genome Biol, 2011. **12**(3): p. R22.

27. Grabherr, M.G., et al., *Full-length transcriptome assembly from RNA-Seq data without a reference genome.* Nat Biotechnol, 2011. **29**(7): p. 644-52.

28. Kent, W.J., *BLAT--the BLAST-like alignment tool.* Genome Res, 2002. **12**(4): p. 656-64.

29. Haas, B.J., et al., *De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis.* Nat Protoc, 2013. **8**(8): p. 1494-512.

30. UniProt, C., *UniProt: a hub for protein information.* Nucleic Acids Res, 2015. **43**(Database issue): p. D204-12.

31. Camacho, C., et al., *BLAST+: architecture and applications.* BMC Bioinformatics, 2009. **10**: p. 421.

32. Ensembl, *Human assembly and gene annotation GRCh38*. 2013.

33. UCSC, *CanFam.1 chainHg19*. 2016.

34. UCSC, *hg38. wgEncodeGencodeCompV2* 2016.

35. Ensembl, *Mouse assembly and gene annotation GRCm38*. 2013.

36. Liu, D., et al., *Molecular homology and difference between spontaneous canine mammary cancer and human breast cancer.* Cancer research, 2014. **74**(18): p. 5045-5056.

37. Ensembl, *Dog assembly and gene annotation CanFam3.1 3.81*. 2013.

38. Grabherr, M.G., et al., *Full-length transcriptome assembly from RNA-Seq data without a reference genome.* Nature Biotechnology, 2011. **29**(7): p. 644-U130.

39. Gray, K.A., et al., *Genenames.org: the HGNC resources in 2015.* Nucleic Acids Res, 2015. **43**(Database issue): p. D1079-85.

40. Varley, K.E., et al., *Recurrent read-through fusion transcripts in breast cancer.* Breast Cancer Res Treat, 2014. **146**(2): p. 287-97.

41. Penso-Dolfin, L., et al., *An Improved microRNA Annotation of the Canine Genome.* PLoS One, 2016. **11**(4): p. e0153453.