CHROMOSOMAL ASSEMBLY AND COMPARATIVE ANALYSIS OF THE RED FOX (*VULPES VULPES*) GENOME

BY

HALIE MARIE RANDO

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Bioinformatics
with a concentration in Animal Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Master's Committee:

Assistant Professor Anna Kukekova, Chair
Professor Jonathan Beever
Professor Sandra Rodriguez-Zas

ABSTRACT

In the early days of genomics, the development of a reference genome was an expensive, collaborative undertaking reserved only for traditional and popular model organisms; however, in a theoretical shift highlighted most clearly by the goals of the Genome 10K Project, the advent of next-generation sequencing (NGS) technology has resulted in a shift of focus towards the development of reference genomes for a variety of species less commonly studied. One non-traditional model organism selected as a priority species for the Genome 10K Project is the red fox (*Vulpes vulpes*), and specifically a fox from an experimental breeding project in which silver foxes (a melanistic variant of the red fox) have been selected over the past several decades to exhibit extreme behavioral phenotypes. The population consists of a strain of hyper-aggressive foxes and a strain of hyper-docile foxes, offering a model system through which the genetic underpinnings of behavior, as well as the genetic correlates of domestication, can be investigated.

The draft red fox genome, which was developed at BGI, has a sequence depth of 94x and is assembled into 676,878 scaffolds with an N50 of 11.80 Mbp. However, in order for the reference genome to be integrated with previous work in the model system, it is necessary to understand the relationship between the scaffolds and the chromosomes they comprise. Therefore, the primary goal of the present study was to assemble the fox chromosomes from the scaffolds of the draft red fox genome assembly.

The draft genome was first analyzed to detect bioinformatic errors known to occur in NGS-assembled genomes that might influence the integrity of the chromosome assembly. Based on these findings, the 500 largest scaffolds were assembled into the 17 fox chromosomes (16 autosomes and the X) based both on nucleotide-level synteny among the fox, dog, and cat identified through pairwise alignment of the reference genomes and on interspecies synteny reported in previously developed comparative maps. The result of the current analysis is the development of a new version of the red fox reference genome that will serve as a valuable tool in ongoing research by increasing the resolution at which mapping studies can probe the genetic architecture of complex behavioral phenotypes in the domesticated fox system.

ACKNOWLEDGEMENTS

I would like to thank Dr. Anna Kukekova for her guidance and support as my advisor over the past two years and Dr. Jon Beever and Dr. Sandra Rodriguez-Zas for serving on my committee and supporting me during my graduate studies. Drs. Marta Farré Belmonte and Dennis Larkin provided invaluable support and advice in the development of the pairwise alignments, and Dr. Jaebum Kim's guidance made the RACA phase of the analysis possible. Christopher Fields, Kathleen Keating and Gloria Rendon, who are consultants at the HPC-Bio group based at the Carl R. Woese Institute for Genomic Biology at the University of Illinois at Urbana-Champaign, should also be thanked for their services in adapting the LASTZ pipeline to run on Biocluster. Xueyan Xiang and colleagues at BGI have been very helpful in answering questions about their fox genome assembly. I'm also very grateful to Dr. Julian Catchen, Dr. Vikram Agarwal, and Jennifer Johnson for advice about genomic research and to Kai Zhao and Michael Robson for computational guidance. Finally, I would like to acknowledge the National Science Foundation's IGERT in Vertically Integrated Genomics (grant number 1069157) overseen by Dr. Andy Suarez for making my graduate research possible and for training me to consider the integration of genomic resources with organismal biology.

TABLE OF CONTENTS

CHAPTER I
GENERAL INTRODUCTION

On April 14, 2003, the announcement that the human genome had been sequenced ushered in what is now known as the Genomic Era (Guttmacher & Collins, 2003). In the 30 years leading up to this announcement, genomics had occupied an ever-increasing importance in modern biology as biotechnological advances allowed increasingly complex genetic sequencing to be conducted. As recently as the early 1970s, individual genes were sequenced using laborious, chromatographic methods, and it was only in 1977 that the first whole genome was sequenced: a bacteriophage (Sanger et al., 1977) whose genome was 5,375 base pairs (bp) long. By the late 1990s, first-generation sequencing technology had improved sufficiently to allow for the sequencing of much larger genomes, including the 4-Mbp *E. coli* (Blattner, 1997) genome and the 97-Mbp genome of the nematode *C. elegans* (The C. elegans Sequencing Consortium, 1998). The release of the 3-Gbp human genome in the early 2000s (Lander et al., 2001; Venter et al., 2001) thereby signified a shift in the power and potential of genomics; technology had advanced to the point where massive mammalian genomes could be sequenced and assembled. The assembly came, however, not without significant investment, as both the time and money required to conduct sequencing with early technologies rendered the cost of these genomes prohibitive.

The earliest reference genomes assembled belonged to commonly-studied model organisms, such as *E. coli* (Blattner, 1997)*, C. elegans* (The C. elegans Sequencing Consortium, 1998), fruit flies (M. D. Adams et al., 2000), humans (Lander et al., 2001; Venter et al., 2001), and mice (Waterston et al., 2002). These species were studied by a huge number of researchers, justifying the high cost of genome assembly. However, over the past decade, significant advances in sequencing technology have reduced the cost and time required to sequence a genome, even genomes as large and complex as mammalian genomes. The development of next-generation sequencing (NGS) technologies has reduced the cost of sequencing dramatically, from $5,292.39 per Mbp in September 2001 to $0.05 per Mbp in July 2014 (Wetterstrand, 2014). The time-cost of sequencing has also decreased dramatically, with Illumina's newest machine, the HiSeq X Ten, reportedly producing 6 Tbp of data per day (Hayden, 2014), compared to the 12 Kbp produced daily with early Sanger sequencing (J. U. Adams, 2008). These technological developments have therefore made sequencing large genomes increasingly feasible.

The increased accessibility of sequencing technology has led to initiatives such as the 10K Genomes Project (Genome 10K Consortium of Scientists, 2009), which seeks to catalogue vertebrate diversity among species and to unlock the full potential of comparative genomics. In fact, when the 10K Genomes Project selected its priority species, the characteristics evaluated included "scientific value" and "phylogenetic diversity" in addition to metrics such as popularity (Genome 10K, 2009). The 10K Genomes Project clearly illustrates a shift in the field of genomics, where genome sequencing projects no

longer focus on traditional model organisms exclusively. As the Genomic Era takes hold, genome sequencing is becoming an avenue to understanding unusual and endangered species, including species with extreme phenotypes of interest (Wagman, 2010).

One genome recently sequenced and assembled as part of the Genome 10K Project is that of a silver fox, a melanistic variant of the red fox (*Vulpes vulpes*), from an experimentally bred population of foxes with extreme behavioral phenotypes. Maintained at the Institute for Cytology and Genetics in Novosibirsk, Russia, one strain of foxes has been bred to be hyper-aggressive, whereas the other has been bred to be docile and even friendly towards humans (Trut, 1999). This genome assembly is thus expected to supply a valuable resource for research into the genetic architecture of complex behavioral phenotypes and the genetic changes accompanying domestication; however, before the genome is widely adopted by fox researchers, it is important to consider the possible repercussions that the bioinformatic challenges encountered during the assembly of genomes from NGS reads can hold for genomic research in organismal biology.

Challenges of Genome Assembly

In its purpose, the Genome 10K project has been almost universally well-received: cataloguing the genomes of species, some of which might soon be extinct, from all branches of the vertebrate tree of life will provide an invaluable resource for comparative genomic studies and species-specific research. However, on a methodological level, the project met some critiques. Without the reduced sequencing costs brought on by NGS technologies, Genome 10K would not be feasible; however, differences in the quality of genomes assembled from NGS reads compared to previous genomes raises questions about what the quality criteria should be and whether NGS methodologies alone are able to meet these standards.

Technological Considerations

NGS technologies have reduced the cost and expedited the output of sequencing. While a number of different platforms have emerged, each with individual strengths and weaknesses, one of the most popular sequencing approached used in the Genome 10K assemblies is the sequencing-by-synthesis technology produced by Illumina. In genome sequencing using Illumina platforms, two types of libraries are typically generated: standard paired-end libraries, and mate-pair libraries.

Paired-end sequencing methods are often used in building a reference genome. During template preparation, DNA is fragmented, and fragments are selected based on the desired library sizes (Illumina Inc., 2011). While the fragment is only sequenced at its ends, usually for between 50 - 150 bases, the fact that the insert size is known provides information about the position of the reads in relation to each other.

Thus, paired-end sequenced produces pairs of short reads separated by an unknown sequence of a known length. This sequencing method can therefore facilitate the assembly of sequences over short distances.

Another library preparation strategy used in genome assembly is that of mate-pair libraries, which allow for the generation of libraries with insert sizes on the kilobase scale (Illumina Inc., 2012). In mate pair library preparation, much like in paired-end sequencing, DNA is fragmented, but here the ends of the fragments are repaired with biotinylated adaptors and the adapters are joined to circularize the fragment (Illumina Inc., 2012). The circularized sequences are then re-fragmented and purified to retrieve only the labeled fragment (Illumina Inc., 2015). This fragment can then be sequenced using the same workflow as standard paired-end sequencing (Illumina Inc., 2010). The most notable difference between these sequences and sequences obtained using standard paired-end sequencing is their orientation. Because in PE sequencing, the fragments are never circularized, the orientation of their sequencing is forward-reverse (FR). However, because of the circularization step in mate-pair library prep, the ends of the original fragment comprise the middle of the fragment actually sequenced, resulting in the reverse-forward (RF) orientation of sequences produced from mate pair libraries. Mate pair libraries are popular in genome assembly because the sequencing of longer insert size libraries facilitates long-range sequence assembly.

Genome assembly projects using NGS methods typically utilize a variety of library insert sizes in order to mitigate the potential for misassembly that arises due to the challenges associated with assigning a single position within a genome to a short sequencing read (Alkan, Sajjadian, & Eichler, 2011). The impact of one of the most significant challenges to assembly, the assembly of repetitive segments, is strongly influenced by decisions made during library preparation. NGS assemblers are liable to conflate multiple occurrences of a similar sequence wherever the repetitive regions are longer than the largest insert size of the libraries used (Figure 1). Henson et al. (2012) modeled this phenomenon in the human by fragmenting the human genome sequence *in silico* at the repetitive regions that would be likely to disrupt assembly so that all unique sequences remained intact, thereby producing an idealized estimate of the maximum lengths of sequence that could be assembled. They estimated that if the human genome were assembled *de novo* using 1000-bp reads, or reads of a similar length to those produced by Sanger sequencing, the maximum possible N50 of the assembled sequences would be 8.978 Mbp. However, if the read size were reduced to a length more characteristic of NGS, such as 50 or 100 bp, the maximum possible contig N50 would drop down to 3-32 Kbp, although the effects of a variety of insert sizes were not fully explored in this analysis. Regardless, the introduction of new computational challenges with the shift to NGS technology is evident. Although the inclusion of libraries with long insert sizes also has the potential to mitigate misassembly due to repetitive sequences, the potential for error when assembling a genome from short reads is substantial.

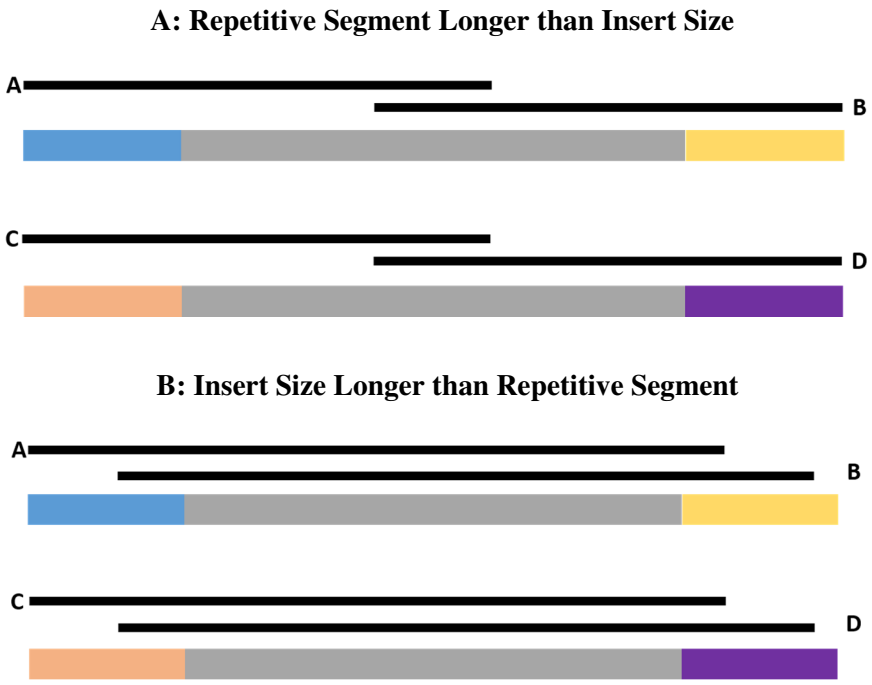


Figure 1: (A) Depiction of the ambiguity that arises when a repetitive segment (grey) is longer than the longest insert size. Sequencing reads are represented by the black bars labeled A, B, C, and D. Based on the reads, it is impossible to tell whether the blue sequence is adjacent to the yellow or purple sequence, as read A could be paired with either B or D. In (B), ambiguity is resolved because the insert size is long enough to cover the whole repetitive sequence, and it is clear that A pairs with B and that C pairs with D. Adapted from description by Henson et al. (2012).

Computational Considerations

The computational and algorithmic challenges of genome assembly have increased with the development of cheaper, faster technology because the short reads generated by NGS supply much less contextual information about each read. The older methods employed in early sequencing projects were more expensive and took more time but produced sequences that were easier to assemble. In the Human Genome Project, for example, a hierarchical shotgun sequencing strategy minimized the risk of long-range misassembly and reduced the risk of local misassembly because shotgun sequencing was undertaken only at the clone level, so that the fragments were known to derive from the same physical region of the genome (Lander et al., 2001). Even the competing Celera project, which utilized whole-genome random shotgun sequencing, did not face these challenges to the extent that NGS genome assembly projects do because Sanger sequencing generates long reads, thereby greatly facilitating the identification of overlapping regions (Henson et al., 2012). As a result, it has been necessary for the algorithmic strategies used in genome assembly to evolve to meet the challenges of genome assembly from NGS reads.

Current popular tools for the *de novo* assembly of NGS reads include ALLPATHS-LG (Gnerre et al., 2011; Ribeiro et al., 2012), Velvet (Zerbino & Birney, 2008) and SOAPdenovo2 (Luo et al., 2012), among others. These tools utilize algorithms that break reads into *k*-mers (fragments of length *k*), construct a de Bruijn graph from the fragments, and identify Eulerian cycles until they have constructed a path that traverses each edge (Compeau, Pevzner, & Tesler, 2011; Pevzner, Tang, & Waterman, 2001), from which the fragments can then be assembled into continuous sequences. However, there are some genomic features which present a significant challenge to assembly.

Repetitive elements and segmental duplications both often disrupt the assembly of continuous stretches of DNA. One analysis of second generation sequencing assemblers (Alkan et al., 2011) compared *de novo*-assembled human genomes from the original SOAPdenovo program (R. Li et al., 2010) against the human reference genome. They found, as anticipated, that repetitive sequences and segmental duplications were both underrepresented in the *de novo* assembly compared to NCBI build 36. These omissions, which occured when similar sequences were mistakenly collapsed together by the assembler, resulted in an assembly that was 16.2% shorter than the reference. They further found that these assembly issues affected not only non-coding regions but also genes, with 30.3% of human genes mapping to more than one scaffold. In a *de novo* genome, this lack of contiguous assembly of genes would be expected to introduce challenges later on in analyzing or predicting genes, given that many avenues of biological research hinge upon the assumption that the reference genome's structure is correct (Salzberg & Yorke, 2005). Additionally, many NGS assemblers have historically been liable to mistake unique repetitive segments for haplotype variation and to merge sequences that shouldn't be merged, resulting in the "orphaning" of unique sequences (Salzberg & Yorke, 2005). Despite the advances in assembly algorithms since the turn of the millennium, genome assembly from NGS reads still encounters a number of challenges in producing accurately assembled genomes. Furthermore, many of the metrics used to assess the quality of an assembly measure length (e.g. scaffold N50) without factoring in the structural integrity of the assembly (Salzberg & Yorke, 2005). Therefore, it cannot necessarily be taken for granted that an NGS-based genome assembly will accurately capture the real biological complexity of the organism sequenced, even when the traditional benchmarks of assembly seem strong.

Importance of Accurate Genome Assembly

NGS technologies have allowed for the assembly of genomes for species outside of the cluster of traditional model organisms. However, the reduced investment in each individual reference genome assembly means that the assembly quality of the new, NGS-assembled genomes often varies significantly from that of reference assembles such as human or mouse. Whereas many older genomes are "finished", involving gap closing and chromosome assembly (Mardis, McPherson, Martienssen, Wilson, & McCombie, 2002), newer genomes are more likely to be assembled into scaffolds without further

refinement. Given the challenges to accurate assembly discussed above, the biological validity of certain aspects of an NGS assembly cannot necessarily be trusted. However, the biological accuracy of a reference genome is critical for downstream use by scientists in more organism-oriented fields within biology, so identifying and resolving possible areas of misassembly is an important step towards making a reference genome useful to the research it can benefit most.

Beyond issues of misassembly, however, many modern genomes are incompletely assembled: genomes are often left as scaffolds instead of being assembled into chromosomes. While scaffold assembly certainly provides a wealth of genomic information for analysis, a lack of information about the physical position of a scaffold within the genome can restrict the range of applications of the assembly. For example, Quantitative Trait Locus (QTL) studies facilitate the genetic analysis of complex traits by estimating association with a trait of interest in intervals along the chromosomes. When the genomic sequence for a chromosome is available, the sequence falling within a QTL interval of interest can be analyzed to identify the genetic features that might be associated with the trait of interest, strengthening the study's potential impact. Because QTL analyzes patterns of linkage at the chromosome level, however, genomic information stored in scaffolds cannot be integrated with these results unless the scaffolds have assigned to positions along the chromosomes. Knowledge about the sequence underlying the chromosomes might also offer insight into regions of cytogenetic interest, such as the site of a fusion event or the sequence surrounding an evolutionary breakpoint region (Murphy et al., 2005). Therefore, the assembly of scaffolds into chromosomes is an important step in creating a genome assembly that can be used across biological disciplines.

As useful as chromosomal assemblies can be, however, the assembly of scaffolds into chromosomes has been characterized as one of the most significant challenges in bioinformatics (Kim et al., 2013). Additionally, misassembly can complicate the relationship between scaffolds and the chromosomes they are supposed to represent, such as when the sequence within a scaffold collapses across a repetitive sequence originating from two separate chromosomes, resulting in a bioinformatic chimera. The challenges inherent to chromosome assembly can, however, be alleviated through the availability of phylogenomic information about the target species, especially if a genome assembly for a reference species is available (Flicek & Birney, 2009). In particular, a recently released tool known as Reference-Assisted Chromosome Assembly, or RACA (Kim et al., 2013), facilitates the assembly of NGS scaffolds into chromosomes by merging collinear alignments to identify syntenic blocks between target, reference, and outgroup species. The program then estimates the posterior probability of adjacency among syntenic fragments (SFs) to assemble the SFs into chromosomes. RACA thereby bypasses some issues of misassembly, such as chimerism, by breaking scaffolds into syntenic fragments and placing them only where alignment is collinear in both the target and reference species. Of course, the tool would

be best suited to cases where the karyotype is similar between the reference and target species, since segments along a scaffold can only be considered collinear if they align along a single chromosome in the reference. RACA has been used to assemble the chromosomes in the Tibetan Antelope genome assembly based on a cow reference (Ge et al., 2013) and provides a powerful tool for the refinement of other NGS assemblies.

Genomics of the Red Fox

The domesticated silver fox was selected as one of the 101 priority species for the Genome 10K project because of its extreme phenotype (Genome 10K, 2009; Wagman, 2010). Given the value of the fox model for research into the genetic correlates of complex behavioral traits, the genome has the potential to open new avenues of investigation into phenotypes of interest such as social behavior, domestication, and aggression, and the genome is likely to contribute significantly to efforts such as high-resolution mapping of regions of interest to behavioral phenotypes, such as previously identified selective sweeps (Kukekova et al., in preparation) and QTLs (Kukekova, Trut, et al., 2011). Additionally, although the species was selected because of its unique phenotype, a number of genetic resources have previously been developed for the fox that will allow for refinement of the draft genome. The newly sequenced red fox genome thus holds the potential to be a valuable tool for the study of complex behavioral phenotypes in a non-traditional model system, and to expand the previously developed resources for the study of foxes.

Development of Experimental Populations

How domesticated or tame behavior arises is a question that evolutionary biology has long sought to answer, especially because tame behavior tends to be accompanied by a number of morphological and physiological traits in what is known as “domestication syndrome” (e.g. Wilkins, Wrangham, & Fitch, 2014). For example, smaller body size (as measured osteologically) and, when applicable, horn size in the domesticated population compared to the wild progenitors is observed across nearly all early domestication events (Clutton-Brock, 1992). Similarly, it is common to find novel coat color variants among domesticated animals that are not commonly observed in the ancestral species, such as single-color coats (Clutton-Brock, 1992) or white spots, as summarized by Trut, Oskina, & Kharlamova (2009). Therefore, domestication is a topic of significant interest because of the complex patterns of the evolutionary change it has precipitated across many species.

Under a popular model of domestication known as the Self-Domestication Hypothesis (Hare, Wobber, & Wrangham, 2012), domestication syndromes arise during a phase of unconscious selection wherein human activities produce a niche where new resources will be available to animals that are able to tolerate increased proximity to humans (Morey, 1994; Trut, 1999). In the 1950s, Dmitry Belyaev and

colleagues as the Russian Institute of Cytology and Genetics sought to test whether an initial phase of selection for animals that could tolerate the stress of the domestication would be sufficient to produce the phenotypic changes characteristic of the domestication syndrome, including small body size and shifts in reproductive timing, as well as an increased tolerance of humans (Trut, 1999). Selection for behavior would likely act on neurotransmitters and hormones, which could in turn have downstream effects beyond those selected for. The hypothesis hinged on two assumptions: first, that genetics contribute to "tamability," or the ability to tolerate the domestication process, and second, that domestication introduces strong selection for this trait (Trut, 1999). The second assumption was already supported by the steep fitness gradient of animals in captivity (Trut, 1999), so the researchers sought to demonstrate the first experimentally by implementing a selective breeding protocol in *Vulpes vulpes*, the red fox.

Belyaev and colleagues acquired an initial population of 130 silver foxes, a melanistic variant of the red fox, from foxes kept at fur farms within the U.S.S.R. (Trut, 1999), the progenitors of which had been captured in Eastern Canada several decades before (Statham et al., 2011) . The fact that the foxes had been living in captivity meant that the initial, extreme phase of selection for tamability had already occurred; however, their "wild and vicious" behavior (Trut, 1980, p. 124) differed significantly from that of long-term domesticates such as goats, sheep, and especially dogs, suggesting that the initial phase of selection for the ability to tolerate proximity to humans was not complete. A scoring system was developed which quantified foxes in the range of -4 to 4 based on their defensive-aggressive response to human contact, incorporating both the distance at which a negative response was elicited, with a smaller distance corresponding to a more tamable animal, and the severity of the reaction (Kukekova, Trut, & Acland, 2014). Under this scoring system, negative scores were indicative of a negative response to human contact, whereas positive scores would have been indicative of a positive response to human contact. In each generation, only the 10% most tamable individuals were selected for breeding, thereby introducing strong selection pressure.

The effects of the breeding experiment on behavior were rapid. In the first generation, the average behavioral score jumped to 1.3 from an average score of -0.96 in the parental population, with a large majority of foxes scoring positively (Trut, 1980). This strong response to selection continued, and by the sixth generation, some kits began actively to seek human physical contact. By the 20th generation, 35% of the kits actively sought human contact, much like dogs (Trut et al., 2009; Trut, 1999). This proactively friendly behavior became increasingly common and by generation 50 was demonstrated by nearly every fox (Trut et al., 2009).

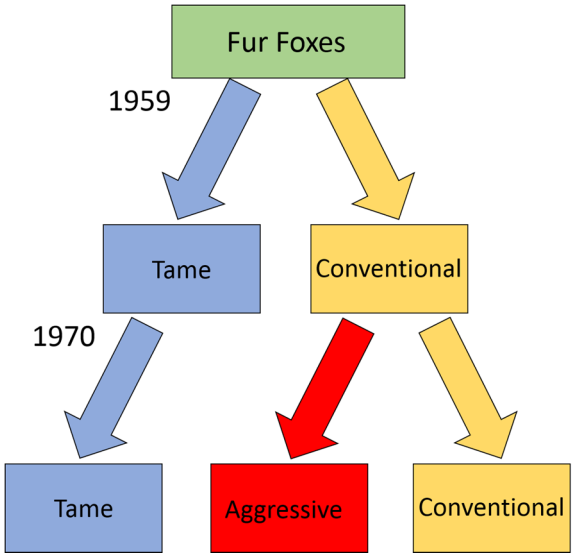


Figure 2: The progression of experimental populations developed at the ICG in Novosibirsk, Siberia during the 20th century. The founding population were captive fur foxes originally derived from eastern Canadian wild fox populations. Today, all three populations are still maintained.

A second experiment was begun in 1970 (Figure 2) in which the 10% of foxes with the most negative scores were bred together to create an "aggressive" line of foxes to contrast with the domesticated fox strain. These two populations demonstrated fixed patterns of behavior, consistently scoring at opposite ends of the behavioral scale (Trut, 1980), as is expected under divergent selection. The results of fifty-five years of experimental breeding suggest that selection for tamability produced a population of foxes with behavioral phenotypes similar to those of dogs, whereas selection against tamability produced a population of aggressive, fearful foxes. These unique phenotypes have rendered the silver fox a species of significant interest to the study of genetic effects on behavior, specifically related to stress, anxiety, and aggression.

Genetics of Tame and Aggressive Behavior

The experimentally bred fox populations have been studied extensively over the past half-century, with the topics studied evolving as new opportunities in biology emerged. Some of the earliest studies focused on quantifying the changes in phenotypes such as behavior (Belyaev, Plyusnina, & Trut, 1985; Belyaev, 1979; Trut, 1980), hormone levels (Belyaev, 1979), and morphology (summarized in Trut, 1999). More recently, the focus has shifted towards the identification of genetic loci producing these phenotypic shifts. Notably, microsatellite markers were developed for the fox (Kukekova et al., 2004) and used to build a meiotic linkage map linking the dog and fox genomes (Kukekova et al., 2007), gene expression profiling has revealed differences in the pre-frontal cortices of the strains (Kukekova, Johnson, et al., 2011), a genotyping-by-sequencing study identified candidate SNPs underlying population differences (Johnson, Wittgenstein, et al., 2015), and whole genome sequencing (WGS) was utilized to

identify a variant causing a coat color phenotype of interest (Johnson, Kozysa, et al., 2015). Thus, until recently, the methods used by the Fox Farm Experiment had shifted as technology has advanced; it is therefore not surprising that with the advent of affordable reference genome assembly, the assembly of a reference genome became the newest development towards the genetic mapping of the phenotypes of interest.

The Red Fox Genome Project

The red fox genome sequencing project began at BGI in 2012. An F1 hybrid (tame x aggressive) fox was selected from the population maintained at the ICG. Fifteen paired-end and mate-pair libraries (Table 1) were selected for sequencing on an Illumina HiSeq 2000 platform. The paired-end libraries were designed to have short insert sizes ranging from 170 to 800 bp and long read lengths of either 100 or 150 bp, and the mate-pair libraries to have long insert sizes ranging from 2,000 to 20,000 bp and short read lengths of 49 bp.

Library Name	Type	Insert Size	Read Length	High-Quality Sequence Depth
SZAXPI000586-5	Paired-End	170	100	25.43
SZAXPI000594-5	Paired-End	170	100	
SZAXPI008070-166	Paired-End	250	150	15.97
SZAXPI000585-11	Paired-End	500	100	15.21
SZAXPI000593-11	Paired-End	500	100	
SZAMPI008069-169	Paired-End	800	150	13.2
VULgnsDBDDWAAPEI-21	Mate Pair	2000	49	11.13
VULgnsDBEDWAAPEI-31	Mate Pair	2000	49	
VULgnsDBFDWAAPEI-16	Mate Pair	2000	49	
VULgnsDBDDLAAPEI-95	Mate Pair	5000	49	5.48
VULgnsDBFDLAAPEI-87	Mate Pair	5000	49	
VULgnsDBGDLAAPEI-34	Mate Pair	6000	49	2.26
VULgnsDBDDTAAPEI-95	Mate Pair	10000	49	4.54
VULgnsDBFDTAAPEI-35	Mate Pair	10000	49	
VULgnsDBEDUABPEI-17	Mate Pair	20000	49	0.69

Table 1: The fifteen paired-end libraries used in the fox genome assembly with their corresponding insert sizes and read lengths.

The sequencing project generated 366.87 Gbp of raw data, of which 225.39 Gbp were high quality. The average genome-wide sequencing depth was 152.86x, of which 93.91x was high-quality. Based on the assembly statistics, the genome size was estimated to be 2.29 Gb. Reads were assembled into contigs with SOAPdenovo2 (Luo et al., 2012) with a contig N50 of 20.12 Kb before the sequences were assembled into 676,878 scaffolds. The scaffold N50 was 11.80 Mb, corresponding to Scaffold58. The genome was also annotated using a combination of homology information, transcriptome data, and

gene prediction. Over 20,000 genes were identified, of which 98% were identified as orthologous to other mammalian genomes. This project has therefore produced an assembly which both adds to the resources available for the genetic study of the red fox and to the available resources for the genomic study of canids as a clade.

Comparative Genomic Analysis

For the past decade, studies in the fox have often relied heavily on the dog reference genome (Lindblad-Toh et al., 2005). Like the dog, the red fox is a member of the family *Canidae* within the order *Carnivora*, and the dog-fox divergence date (Figure 3) is estimated at 9-10 MYA (Wayne, 1993). Examples of the dog genome's uses for study of the fox include the development of fox microsatellite markers based on canine markers (Kukekova et al., 2004, 2007) and the design of primers in the dog for the amplification of fox sequences with PCR (Johnson, Kozysa, et al., 2015; Johnson, Wittgenstein, et al., 2015; Kukekova et al., 2007). Now that the fox reference genome has been assembled, it will be possible to compare the genomes of the two species at a higher resolution.

Karyotypes

Though the fox and dog are very closely related, their karyotypes have diverged dramatically: the fox karyotype is made up of $2n = 34$ chromosomes, all bi-armed, with 0-8 B-chromosomes depending on the individual fox (Becker et al., 2011; F. Yang et al., 1999). By comparison, the dog has 38 acrocentric autosomes in addition to its metacentric sex chromosomes (Becker et al., 2011), comprising $2n = 78$. Syntenic blocks between the dog and fox are highly conserved, and each fox autosome maps continuously to two to three dog chromosomes. This synteny can likely be explained by the fact that, though the dog karyotype has been characterized as one of the most rearranged within *Carnivora* (F. Yang et al., 1999), the ancestral canid is estimated to have had a karyotype very similar to that of the modern dog, with $2n = 82$ or more (Graphodatsky et al., 2008).

Outside of *Canidae*, however, carnivore chromosome numbers are much lower: for example, the cat, whose least common ancestor with the fox and dog was likely 50-60 million years ago (Murphy et al., 2007; Wayne, 1993) (Figure 3), has $2n = 38$ chromosomes. Sixteen of the cat autosomes as well as the sex chromosomes are bi-armed and two are single-armed. The ancestral carnivore is estimated to have had a karyotype very similar to the modern cat, with $2n = 42$ (Murphy, Stanyon, & O'Brien, 2001; Nash, Menninger, Wienberg, Padilla-Nash, & O'Brien, 2001). Despite the superficial similarity in the karyotypes of the cat and fox, syntenic blocks among *Canidae* and *Felidae* appear to have undergone substantially rearrangement (Davis et al., 2009; F. Yang et al., 2000).

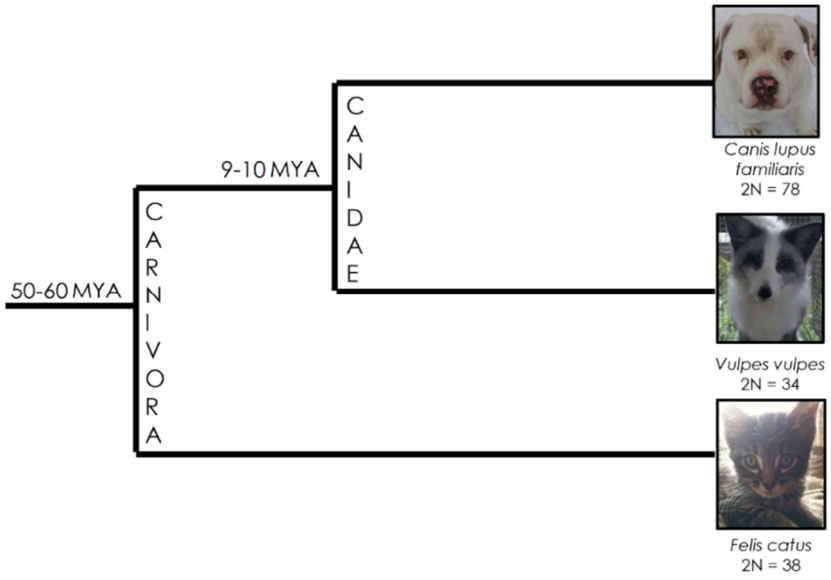


Figure 3: A cladogram, with approximate divergence dates (Wayne, 1993), of the fox, dog, and cat.

The syntenic blocks among the species were originally characterized with chromosome painting to identify regions of interspecies chromosomal homology. In chromosome painting, probes designed from the genetic material of one species are hybridized onto another species to identify regions of synteny at a resolution of about 5-10 Mbp (Becker et al., 2011). Canid karyotypes and interspecies synteny have been explored through a number of studies, with several comparative karyotypes developed that include, among other species, the red fox and the dog (Becker et al., 2011; Graphodatsky et al., 2000; F. Yang et al., 1999). Comparative karyotypes have also been developed for the dog and cat (F. Yang et al., 2000). Thus, a significant body of research is available allowing for the identification of chromosomal rearrangements within the evolutionary history of these species.

These studies suggest, as suggested by divergence time, that synteny is highly conserved between dog and fox, and much less so between dog and cat. For example, the red fox karyotype could emerge from the dog karyotype through 26 fusions and 4 fissions (F. Yang et al., 1999), with full acrocentric dog chromosomes often corresponding to arms of the metacentric fox chromosomes. The comparative karyotype therefore identifies 42 conserved syntenic blocks between the two species. Between the dog and the cat, on the other hand, many more rearrangements would be required to estimate the evolutionary steps leading to the two species' modern karyotypes: painting of the dog chromosomes with probes from cat autosomes identified 65 conserved segments among the species, while painting in the reverse direction identified 68 (Yang et al., 2000). A fox-cat comparative karyotype has not currently been experimentally developed.

Marker-Based Comparative Maps

In addition to the chromosome-level comparative maps developed with chromosome painting at the turn of the millennium, a number of studies have sought to characterize interspecies homology at a finer resolution. In the case of fox-dog comparisons, the major resource is a meiotic linkage map developed for the fox chromosomes using canine microsatellite markers (Kukekova et al., 2007). In this study, 320 canine microsatellite markers were assembled into 17 linkage groups corresponding to the sixteen fox autosomes and the X-chromosome. The average inter-marker distance was reported as approximately 7.5 cM. This map was considered a significant advancement in the tools available for mapping the loci associated with domestication (Spady & Ostrander, 2007) and has been used in subsequent research, such as Johnson, Wittgenstein, et al. (2015).

However, though all of the markers that mapped both to the dog and fox were uniquely assigned, some of the markers were too closely linked to be assigned a specific position with high confidence (Kukekova et al., 2007). Therefore, there is a possibility that higher-resolution analysis of dog-fox alignment could yield refinements to the current dog-fox syntenic map through the identification of rearrangements. Additionally, breakpoints could not be resolved at very high resolution in the four regions where synteny between the dog and fox was disrupted, which presents another opportunity for the dog-fox map to be refined through further analysis. The meiotic linkage map thus provides a high-quality map of synteny between the dog and fox which, though consistent with the findings from chromosome painting, offers increased resolution.

A syntenic map has also been developed for the dog and cat at a resolution of 939 Kbp through radiation-hybrid (RH) mapping (Davis et al., 2009). This tool was developed to assist with the assembly of the feline genome and to offer insight into chromosome evolution in cats with respect to other species with reference genomes. In this study, 2,662 markers were placed with roughly uniform density along the eighteen cat autosomes and X-chromosome. This analysis was complemented by a previously-developed meiotic linkage map for the cat (Menotti-Raymond et al., 2009), with which it was 94% consistent. This RH-map thus offers improvement in the resolution of cat-dog synteny from earlier comparative karyotypes built using chromosome painting (F. Yang et al., 2000).

Genome-by-Genome Comparisons

Whereas chromosome painting and linkage mapping are useful tools that have offered valuable insight about syntenic relationships among species for decades, the advent of the genomic era offered an opportunity for finer resolution analysis through pairwise genome alignment. Pairwise genome alignment is the genome-scale comparison of two genomes using an alignment program, often LASTZ (Harris, 2007). Unlike the draft red fox genome, long-running genome projects have produced and refined reference genome assemblies for both the dog (Lindblad-Toh et al., 2005) and the cat (Montague et al.,

2014; Pontius et al., 2007). The dog genome, which was published a decade ago, used the whole-genome shotgun approach with much longer Sanger sequencing reads. Both versions of the cat genome utilized a whole-genome shotgun approach with long sequencing reads. Additionally, both the cat and dog genomes have been assembled into chromosomes, with these assemblies publicly available through the UCSC Genome Browser (Kent et al., 2002; Rosenbloom et al., 2014). Therefore, the status of these two genomes present an opportunity for comparative genomic analysis of other species within the *Carnivora* family, such as the fox, via pairwise genome alignment.

Current Objectives

The experimentally bred silver fox system presents a unique opportunity to study the genetic architecture of behavior and the genetic correlates of domestication. While a number of tools have been developed for the study of the fox, the development of genomic resources are the next step towards utilizing this model at its full potential. The genome has been newly sequenced and assembled by BGI as part of the Genome 10K initiative; however, this assembly is still in draft form. The current research thus seeks to take the first steps towards refining the fox genome assembly for use by the larger biological community.

The present study focuses on addressing two issues characteristic of *de novo* genome assembly with NGS. This first area considered is the quality of the assembly beyond statistics such as N50 and coverage, focusing on evaluating the extent to which issues of misassembly such as genome shortening, loss of repetitive sequences, non-contiguity of genes, and bioinformatic chimerism can be expected to affect the genome in its current form. The second phase of analysis focuses on the assembly of the fox chromosomes from the scaffolds. These analyses facilitate the integration of the genome with previously developed tools in order to strengthen research into the genetic architecture of behavior in the fox and also produce deliverables that will strengthen the fox genome as a tool for use by organismal biologists.

CHAPTER II
NGS ASSEMBLY OF THE RED FOX GENOME

Overview/Motivation

The typical benchmarks of *de novo* genome assembly speak to the quality of the draft red fox genome, which has a contig N50 of 20.12 Kb, a scaffold N50 of 11.80 Mb, and average genome-wide coverage of 93.91x. However, these statistics speak mainly to the length of the assembly contigs and scaffolds without considering the structural integrity and accuracy of the assembly (Salzberg & Yorke, 2005). The draft genome is expected to provide a valuable tool for genetic and genomic study of the silver fox, but misassembly could complicate the use of this assembly in the anticipated behavioral genetic studies. Therefore, the draft genome was evaluated for assembly quality through a set of experiments designed to address assembly issues characteristic of *de novo* genome assemblies sequenced with NGS.

As discussed above, a number of admonitions have been put forth about the quality of *de novo*-assembled genomes (Alkan et al., 2011; Eichler, 2001; Henson et al., 2012; Salzberg & Yorke, 2005). Many of these studies identified pitfalls of NGS assembly by assembling human or mouse sequences *de novo* and comparing them against the organism's accepted reference genome. For the fox genome assembly, however, this type of comparison is not feasible. While the high-quality dog reference can be used for comparison, several millions of years of divergence between the two species prevents direct assessment of the fox based on the dog; however, the dog genome can serve as a guide in some respects.

In order to analyze the biological integrity of the draft red fox genome assembly, several features were examined, including the size of the draft genome and its composite scaffolds, the rate at which repetitive segments were incorporated into scaffolds, the fraction of short scaffolds containing unique sequence, whether scaffolds had been misassembled such that they were chimeric (i.e. contained sequence from multiple, distinct chromosomes), and whether genes had been assembled continuously or were split up across multiple scaffolds.

Misassembly would be expected to have widespread potential impacts on future studies, especially if misassembled sequences were incorporated into chromosome assemblies. More immediately, however, it was necessary to evaluate the scaffolds to determine whether or not they appeared to accurately recapitulate the sequences that comprise the fox chromosomes. Typical problems with using NGS for genome assemblies would suggest the risk that many of the small scaffolds might contain sequences such as repetitive segments or heterozygous regions that had not been correctly incorporated into the assembly of larger chromosomes. Additionally, characteristics such as a loss of sequence in the assembly relative to the length expected would suggest that the mapping of the scaffolds to the chromosomes would be incomplete. Chimerism is a particularly challenging problem for the assembly of scaffolds into chromosomes, since a scaffold cannot be assigned to a unique location in the genome if it

contains sequence from multiple regions of the genome. The shortcomings of the draft genome needed to be analyzed to determine the strategy needed to build a version of the genome that would most accurately capture the biology of the fox.

Methods

Library Composition

Each of the 15 libraries of sequencing reads was aligned against the 500 largest fox scaffolds using BWA (H. Li & Durbin, 2009). Because of the differences in read length between the paired-end and mate-pair libraries (Table 1), two different BWA alignment programs were used for the two categories of libraries: BWA-backtrack for the mate pair libraries and BWA-MEM for the paired-end libraries. BWA was selected for alignment after Bowtie2 (Langmead & Salzberg, 2012) was found to produce very low (<= 0.5%) rates of alignment for the mate-pair libraries because it is not designed for the alignment of very short reads.

Size metrics were then estimated for the mapping of each library against the fox genome using Picard (Broad Institute, 2014), a suite of tools developed by the Broad Institute for the manipulation and analysis of SAM and BAM files. Observed insert size distributions were computed for each library and compared against the estimated insert sizes reported by BGI to ensure that mapping to the fox genome accurately reflected the size distributions selected during template preparation. Finally, insert size distributions were visualized for each library using Picard's histogram function to facilitate visual inspection of the results.

Some challenges were encountered in estimating the insert size distributions of the mate pair libraries. By default, BWA assumes reads are in the "forward-reverse" (FR) orientation as is typical in single-end and paired-end sequencing, but mate-pair sequences are in the "reverse-forward" (RF) orientation (Illumina Inc., 2012). Therefore, in the mate-pair libraries, the RF reads were incorrectly flagged by BWA as non-concordant whereas the FR reads were incorrectly flagged as concordant, leading to erroneous estimates of the insert sizes consistent with analyzing only sequences inadvertently sequenced due to incomplete enrichment during sequencing (Illumina Inc., 2012). Once the BAM files were computationally filtered to remove any reads flagged as concordant using the parameter -F 0x2, the insert sizes for the correct alignments were obtained and analyzed.

Genome Size and Scaffold Size Distribution

The observed length of the fox genome sequence was compared against a previously reported value obtained from the Animal Genome Size Database (Gregory, 2015) that was based on analysis of DNA content using flow cytometry (Wurster-Hill et al., 1988). The C-value reported for the red fox is 2.85 pg, which was converted to a length in bp using the formula: $G = (0.978 \times 10^9) \times C$ (Dolezel, Bartos,

Voglmayr, & Greilhuber, 2003), where *C* is the C-value in pg and *G* is the size of the genome in bp. The percent differences in the two sizes was then calculated and compared against the estimated percentage of sequence lost during *de novo* NGS assembly reported by Alkan et al. (2011).

Then, the length of each scaffold was measured by opening the corresponding FASTA file in Python and calculating the word count minus the header using Python's len() function. These results were then written to a .CSV file and their distribution examined both as raw and log-transformed values in order to elucidate the scaffold lengths included in the assembly.

Analysis of the Small Scaffolds

The major challenges in sequence assembly from NGS reads are repetitive elements and orthologous regions, including haplotype diversity. These features are liable to result in bioinformatic errors. In order to evaluate the effect of these genomic features in the red fox genome assembly, the 676,878 scaffolds were screened to identify any scaffolds composed entirely of repetitive elements. This was achieved by soft-masking the scaffolds using the command-line release of RepeatMasker 4.0 (Smit, Hubley, & Green, 2013) with the "species" parameter set to dog and the "xsmall" parameter specified for soft-masking. Soft-masking generated versions of each scaffold where all repetitive sequences had been transformed to lower-case. Each soft-masked scaffold was then evaluated in Python using (any(x.isupper() for x in seq)), where seq was the complete sequence of the soft-masked scaffold, in order to determine whether any of the bases remained unmasked, or in other words, to identify sequences composed entirely of repetitive elements.

In order to assess whether the repeat-only scaffolds contained sequences that had been incorporated into larger scaffolds, the unmasked version of each scaffold was then mapped against the 500 largest scaffolds in the fox genome using LAST (Frith, Hamada, & Horton, 2010). Mapping a repetitive scaffold against the largest scaffolds revealed whether or not it had been incorporated into a long (defined as 50-Kbp or longer) stretch of sequence. In order to be considered a match, 95% of the total length of the repeat-only scaffold was required to map to a single large scaffold. The number of times a scaffold mapped to the large scaffolds was then evaluated across all repeat-only scaffolds.

A second analysis was undertaken using the small scaffolds that were not repeat-only in order to determine the extent to which the small scaffolds recapitulated "unique" (i.e. non-repetitive) sequences that were also represented in the 500 largest scaffolds. First, the 500 largest scaffolds were aligned against all larger scaffolds (i.e. Scaffold19 would be aligned against scaffolds 1-18) using LAST to ensure that each contained unique information. Then, a random set of 1000 scaffolds in the range between Scaffold501 and Scaffold676878 was selected with pseudorandom sampling in Python. The randomly selected scaffolds were screened to ensure that they were not repeat-only scaffolds, as described above. While it would have been ideal to map all scaffolds against all other scaffolds, due to the excessive

computational challenge presented by working with 676,878 scaffolds, each of the 1000 randomly sampled scaffolds was instead mapped back against the 500 largest fox scaffolds. A Python script was written to identify scaffolds that mapped completely (defined as within 3bp of the total length) to one of the large scaffolds with an alignment score of at least 80% of the match length.

Assembly and Misassembly of Genes

In order to estimate the extent to which the genes in the draft red fox genome had been misassembled, a subset of transcripts in the dog transcriptome published by the National Center for Biotechnology Information (NCBI)[1] was used as a proxy. The transcriptome contains a large number of predicted genes, but the subset selected for analysis was exclusively the 1,334 Known RefSeq genes (those belonging to the category "NM"), which have been experimentally validated in the dog. Only one transcript of each gene was included in the analysis. The sequences were mapped against the full masked fox genome assembly using *blastn*, and only hits with an e-value less than 1.5 were considered. This liberal threshold was selected to minimize any the effects of sequence divergence.

Blast results were returned in a tabular format (using the -m 8 parameter), and the results were analyzed with a Python script written for this analysis. Specifically, the script scanned the output to identify the scaffold with which the gene had an alignment with the lowest e-value, corresponding to the highest probable match. It then evaluated whether the cDNA (query) mapped contiguously against that scaffold (target). Because introns were expected to be represented in the genome target but not in the cDNA query, the alignment was required to map continuously on the query, with gaps of no more than 10% of the total length. The gene was assumed to map continuously to the scaffold only if at least 90% of the transcript by length had mapped.

Those transcripts which did not map continuously to a single scaffold were then scanned to determine whether gaps at the beginning or end of the alignment of the transcript to the genome could be filled in by including matches to another scaffold. If a segment of a transcript mapped to a second scaffold, the transcript was identified as chimerically assembled in the fox. Any transcripts that remained uncharacterized after both rounds of analysis were examined manually. During manual inspection, the gap criterion was relaxed so that if the beginning and end of a transcript mapped to a single scaffold, it was assigned to that position.

Estimating Chimerism

When sequences are misassembled, they can be sampled from biologically disparate regions of the genome. The identification of two non-contiguous segments of sequence as adjacent is known as chimerism. The identification of chimeric scaffolds in a *de novo* assembly can be facilitated by a

[1] Version GCF_000002285, downloaded April 21, 2015 from http://www.ncbi.nlm.nih.gov/genome?term=canis%20lupus%20familiaris

reference genome if a syntenic map between the two species is available. In the case of the draft red fox genome, identification of chimeras can therefore be facilitated by comparing scaffolds against the dog genome and against the patterns of dog-fox synteny identified in the dog-fox comparative karyotype (F. Yang et al., 1999) and meiotic linkage map (Kukekova et al., 2007).

Mapping of the largest 500 fox scaffolds against the dog genome was conducted with LAST (Frith et al., 2010). The dog genome used in the present analysis consisted of the 38 autosomes, mtDNA, and X-chromosome of CanFam3.1, and, additionally, the dog Y-chromosome recently assembled by Gang Li and William Murphy (G. Li, personal communication, October 14, 2014). For each scaffold, the dog chromosome(s) to which it was most likely orthologous was/were identified. Every scaffold mapped to many chromosomes, so it was necessary to identify the chromosome(s) with which it was most likely to be syntenic. Thus, for each scaffold, the maximum LAST score corresponding to each chromosome was identified, and its z-score relative to the maximum LAST scores of all other chromosomes to which the scaffold mapped was computed with the formula $Z = \frac{x-\mu}{\delta}$, where, for each scaffold, x is the maximum score of the chromosome of interest, μ is the average maximum score across all chromosomes, and δ is the standard deviation of maximum scores across all chromosomes. Example distributions are provided in Appendix A. Chromosomes with scores that were large enough to be significant at $p<0.05$ were considered significant hits and identified as syntenic to at least a portion of the scaffold.

Analysis using a second approach complemented the results of the maximum LAST score analysis. Here, a Python script was written to scan the LAST mapping results nucleotide by nucleotide and to identify, at each position along the scaffold, the mapping hit with the highest score. The genome position in the dog of this best-mapped hit was then graphed at each position along the scaffold. This method produced an alternative analysis that allows for comparison of the best hit at each nucleotide to the scaffold-level syntenic analysis using z-scores.

Results/Discussion

Library Metrics

The average observed insert sizes in each of the libraries when mapped to the Fox500 were similar to the predicted insert sizes (Table 2; Appendix B). While most of the libraries show smooth, approximately normal distributions, the distribution of the estimated insert sizes for the library with the largest insert size, VULgnsDBEDUABPEI-17, is much less smooth than the others (see Appendix B for visualization). This phenomenon could potentially be caused by the fact that this library has much lower coverage than the others, at only 0.69X (Table 1) and that there would be fewer opportunities for the read-pairs of a library with an insert size of 20 Kbp to map concordantly onto the scaffolds, given that so many of the scaffolds are small: Scaffold500, for example, at 47,686-bp long, is only 2.4 times the length of this

library. It has also been noted that errors are more common for longer insert sizes (Henson et al., 2012), though no trend of smoothness decreasing with insert size is evident in the other libraries (see Appendix B). Despite the roughness of this single graph, however, as a whole, the statistics calculated from the mapping of the libraries to the genome closely approximate the predicted sizes (Table 2).

Library Name	Expected Insert Size	Read Length	Mean Insert Size	Standard Deviation	Median Insert Size
SZAXPI000586-5	170	100	155.23	15.72	157
SZAXPI000594-5	170	100	152.39	13.40	153
SZAXPI008070-166	250	150	220.70	12.33	222
SZAXPI000585-11	500	100	464.24	18.39	465
SZAXPI000593-11	500	100	461.36	18.25	462
SZAMPI008069-169	800	150	785.00	32.75	787
VULgnsDBDDWAAPEI-21	2000	49	2362.02	178.33	2364
VULgnsDBEDWAAPEI-31	2000	49	2045.96	188.07	2044
VULgnsDBFDWAAPEI-16	2000	49	2265.11	171.51	2265
VULgnsDBDDLAAPEI-95	5000	49	4967.83	318.65	4965
VULgnsDBFDLAAPEI-87	5000	49	5520.72	351.67	5525
VULgnsDBGDLAAPEI-34	6000	49	5942.20	388.64	5939
VULgnsDBDDTAAPEI-95	10000	49	10764.81	1267.83	10967
VULgnsDBFDTAAPEI-35	10000	49	10905.67	1377.52	11133
VULgnsDBEDUABPEI-17	20000	49	21551.75	5462.30	23277

Table 2: Insert size metrics for each of the 15 libraries. Mean, standard deviation, and median insert sizes were estimated with Picard.

The similarity of the measured insert sizes to the expected values supports the integrity of the assembly. The reads would map onto the assembled genome only if it had faithfully recapitulated the sequence from which they were derived. While this does not preclude misassembly, it does speak to the relative absence of major structural errors among the 500 largest scaffolds. The shorter scaffolds, which would be expected to provide less information for long-range assembly, were excluded from the present analysis; therefore, the integrity of the small scaffolds that would be useful only for short-range assembly has not been evaluated. In the future, it may be interesting to use the paired-end libraries to evaluate the structural support for the small scaffolds, but based on the present analysis, the estimated library metrics largely support the integrity of the draft genome assembly.

Genome Size Comparison and Trends in Scaffold Size

The conversion of the C-value size from the Animal Genome Size Database produced an estimate for the size of the fox genome of 2.79 Gbp. The size of the fox genome assembly estimated by BGI was 2.29 Gbp, meaning that the BGI estimation is only 82% of the size estimated with flow cytometry. This

discrepancy is larger than the 16.2% reduction in genome size due to misassembly reported of *de novo* NGS assemblies by Alkan et al. (2011).

However, the human genome is currently estimated at 3.23 Gbp (Genome Resource Consortium, 2013), but the flow cytometric estimate for the size of the human genome produced by the same authors as the estimate for the fox (Wurster-Hill et al., 1988) suggests an approximate human genome size of 3.42 Gbp. This discrepancy suggests that the C-value estimates may be slightly inflated, given that the genome size estimation for the finished human genome is only 94.4% of the size estimated with flow cytometry. Scaling the C-value estimated for the fox to account for this discrepancy would give an estimate of 2.63 Gbp, corresponding to a discrepancy of about 13% with BGI's estimate. These results therefore suggest that the fox draft genome assembly is likely to be missing at least several hundred megabases of genetic material.

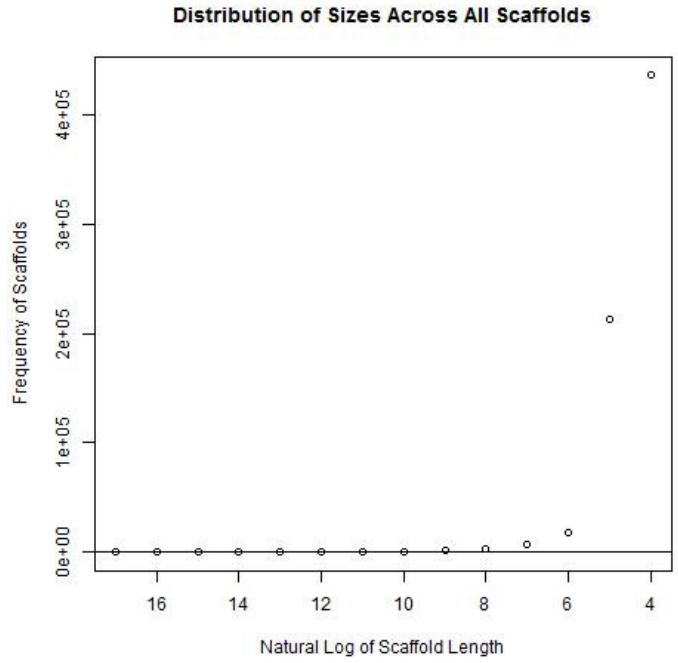


Figure 4: Frequency of scaffold sizes in the fox genome, with size log-transformed and binned. As scaffold size decreases, frequency of occurrence among the scaffolds increases, meaning that there are a lot of very small scaffolds.

Additionally, the analysis of the length distribution among the scaffolds suggests that the majority of the genomic information in the fox genome assembly has been incorporated into a few large scaffolds, while a huge number of very small scaffolds supply very little additional data. The largest scaffold is 55,683,013-bp long, and the smallest scaffold runs only 100 bp. The length of 100 bp appears, however, to be a specified minimum cutoff, and there are 39,843 100-bp scaffolds in the full assembly, and nothing smaller. Across the genome as a whole, most of the scaffolds are very small. Figure 4 shows the

distribution of scaffold sizes when the lengths are transformed with a natural log. In total, 64.4% of the 676,878 scaffolds fell into the smallest bin, corresponding to lengths less than 149 bp or to ln(length) less than five. Only 1.69% of scaffolds were longer than 1097 bp (i.e. had a ln(length) of seven or higher). Though the distribution is not normal, it is interesting to note that the median scaffold length for the genome as a whole is 126 bp and the average is 243 bp, suggesting that a few very large scaffolds may be skewing the distribution to make the average larger than the mean.

In order to determine the effect that the largest scaffolds might have on the genome-wide length statistics, the distribution of lengths within the largest 500 scaffolds was examined to see whether it followed the same trends as the distribution in the genome as a whole. The smallest scaffold included in this analysis was Scaffold500, which is 47,686-bp long. Though of course the distribution is not normal (Figure 5), it's interesting to note that the average scaffold size among the largest 500 scaffolds is 4,698,760 bp, with a standard deviation of 7,639,151, whereas the median is 1,524,697 bp. The trend here therefore recapitulates the genome-wide trend, with a few scaffolds containing most of the genomic information. These 500 scaffolds were calculated to contain 94% of the entire sequence of the draft fox genome by length.

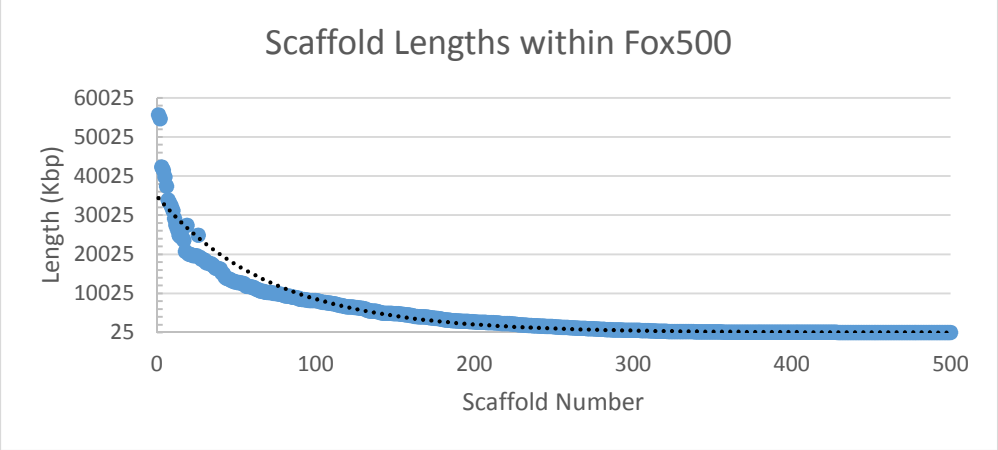


Figure 5: The lengths of the scaffolds in Fox500 are each represented by a blue dot. The maximum size was 55.7 Mbp and the minimum size was 47.7 Kbp. Scaffolds 19 and 26 are slightly longer than their ordinal position would indicate by several Mbp, but in general the scaffold sizes decay in a roughly exponential trend, as indicated by the black dotted line.

These trends in scaffold length suggest that the elimination of some scaffolds, either by selecting a cut-off or by eliminating scaffolds that are unlikely to contain unique sequences (discussed in the next section), could potentially produce a more computationally usable draft genome without eliminating any unique information. The presence of so many small scaffolds suggests that assembly failed at a number of points, "orphaning" many sequences. The very small scaffolds, which in many cases are shorter than the smallest library size and are therefore likely to represent isolated sequencing reads, are unlikely to contain any information that will be useful in constructing the genome at long- or even short-range.

Without further analysis of the composition of these small scaffolds, it was impossible to determine why they were not incorporated into larger sequence assemblies. Assembly would be more likely to fail at genomic features such as repetitive segments and segmental duplications, and heterozygous sequences are known to introduce a number of assembly errors when short reads are used (Alkan et al., 2011; Henson et al., 2012) because the nearly-identical sequences cannot be merged (Salzberg & Yorke, 2005). Additional analysis was therefore necessary to determine whether any of these common bioinformatic challenges could explain the large number of very small scaffolds in the assembly.

Analysis of the Small Scaffolds

After soft-masking, 107,185 scaffolds were found to be composed entirely of repetitive sequences, corresponding to 15.8% of all scaffolds by count and comprising 12.5 Mbp of sequence data. The largest scaffold identified in this analysis was Scaffold40760, which is 273 bp long. When the unmasked all-repeat scaffolds were mapped back onto the 500 largest fox scaffolds, 747 (0.69%) were not represented at all, mapping zero times (Figure 6). An additional 4461 (4.61%) were represented only once. Most scaffolds, however, were repeated between one and 500 times in the largest scaffolds, with Scaffold45466 mapping back onto the largest scaffolds 1,538 times.

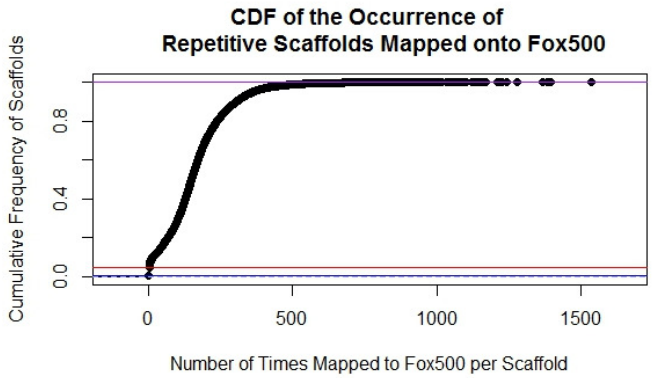


Figure 6: Cumulative Distribution Function indicating the number of times repeat-only scaffolds mapped back onto the largest scaffolds. The black line's intersection with the blue line indicates the percentage of scaffolds mapping 0 times (0.69%) and with the red line, scaffolds mapping 1 or fewer times (4.86%). The purple line demarcates the full set of scaffolds. Almost all repetitive scaffolds mapped to the largest scaffolds between 1 and 500 times.

It is theoretically impossible to assign a location in the genome to any scaffold that contains only repetitive elements because by definition it has no single genomic position, rendering these scaffolds unusable in further assembly. Furthermore, canid genomes are known to contain a large number of SINE elements (Bentolila et al., 1999), whose transposition is a major source of canine diversity (Wang & Kirkness, 2005). Therefore, the position of these elements within the genome would be expected to be informative to both fox genomic diversity and fox-dog divergence. Identification of the scaffolds containing only repetitive sequences with RepeatMasker and their corresponding representation among

the largest stretches of assembled sequence served to underscore that in many cases, information about the positions of these sequences in the genome has likely been lost during assembly.

Mapping the repeat-only scaffolds against the 500 largest fox scaffolds suggested that 95% of these small scaffolds had been incorporated into the genome in at least two locations. However, the number of times each repetitive scaffold truly occurs in the fox genome remains unknown, so it is not currently possible to determine how closely this representation would be expected to recapitulate fox biology. It is evident that a number of sequences were never incorporated into stretches of unique assembled sequence, and that 4.61% are very likely to be underrepresented once the scaffolds are assembled into chromosomes. Therefore, the challenge of incorporating repetitive elements in *de novo* assembly appear to affect the fox genome and will cause the loss of some of this information once the scaffolds are assembled into chromosomes.

The second set of analyses on the small scaffolds sought to estimate the uniqueness of the small scaffolds by mapping them against larger scaffolds. As expected, none of the scaffolds ~50 Kbp or larger were subsumed by a larger scaffold, and thus were treated as unique sequence assemblies in the present analysis. Sizes among the 1,000 randomly sampled small scaffolds ranged from 100 to 9,086 bp and represented to 185,851 bp of sequence. Of the scaffolds, 3.8% did not align anywhere in the largest scaffolds, suggesting that they contain unique sequences (Table 3). For an additional 0.3%, partial matches were found, suggesting similar motifs but unique sequence.

Match Type	Frequency	Sequence Length (bp)
None	3.8%	40,102
One Location	56.0%	89,923
Multiple Locations	39.9%	52,385
Incomplete	0.3%	3,441

Table 3: Mapping of the 1000 randomly selected small scaffolds onto the largest fox scaffolds. Total sequence length queried was 185,851 bp. “Incomplete” matches did not meet the length threshold qualification, which required matches to be no more than 3 bp shorter than the scaffold itself.

The general mapping trends of the small, unique scaffolds against the largest scaffolds suggested that assembly errors significantly influenced the composition of the small scaffolds. The majority of the scaffolds analyzed (56.0%) mapped to a single location in the large, masked scaffolds. The fact that these sequences were not incorporated into the genome would seem likely to have been caused by duplications or by heterozygosity. However, the fact that so many of the sequences (39.9%) mapped to multiple locations within the largest fox scaffolds suggests that these small scaffolds are not all the product of heterozygosity, but rather may also represent prolific duplication events or possibly novel fox-specific repetitive sequences that were not detected by RepeatMasker. Future analysis could assess the content of

these scaffolds in order to better estimate the bioinformatic or genomic features that have resulted in such high identity between small scaffolds and regions of the larger scaffolds.

Assembly and Misassembly of Genes

Another potential implication of misassembly for downstream analysis is the misassembly of genes, which can result from the disruption of assembly due to duplications or to repetitive intronic sequences (Alkan et al., 2011). If a gene is split into multiple scaffolds, it will be detected in pieces during gene prediction and may present problems during common uses of a reference genome, such as primer design. To assess whether gene misassembly was likely to affect a large number of genes in the fox genome, transcripts from the NM category of the NCBI's dog RefSeq transcriptome were mapped against the full fox genome. The NM category was selected because predicted genes may contain errors, and therefore the use of transcripts from only experimentally validated genes offered a more appropriate standard by which to evaluate the fox assembly.

Category	Number of Transcripts	Percent of NM Transcripts
Number Mapped to One Scaffold	1247	93.50%
Number Mapping Chimerically	54	4.00%
Number Unplaced	33	2.50%
Total	1334	--

Table 4: Known RefSeq Genes mapped back onto the fox genome and assigned to one or more scaffolds.

Of the 1,334 transcripts in dog NM database, 1,247 could be assigned to a position on a single scaffold in the fox genome (Table 4). An additional 53 genes were identified as mapping chimerically to two or more scaffolds, and 33 could not be assigned to a position in the draft fox genome because a large segment of the transcript did not align anywhere in the fox genome. This analysis suggests that a majority of known dog genes are represented by continuous orthologous sequences in the red fox assembly, but that errors may have been introduced during the assembly of approximately 6.5% of genes. One interesting phenomenon of note was that out of all 1,247 transcripts assigned to a single location in the fox genome, all but one were placed in scaffolds in the range of Scaffold1 to Scaffold449. This trend in assignment suggests that, much like genomic information, most of the genetic information in the fox assembly is also contained within a small number of large scaffolds.

The fact that 33 transcripts did not map in their entirety to the draft fox genome could be influenced by a number of factors. For instance, if these genes have undergone significant sequence evolution between the fox and the dog, certain exons may not have aligned due to a loss of homology, though a relaxed e-value of 1.5 ($p \leq 0.777$) was used to try to mitigate the potential for sequence evolution to disrupt alignment. More likely, these transcripts did not align because of gaps or errors in the

fox genome assembly. Further investigation into the genes that could not be aligned in this analysis could potentially reveal insight into either errors in the genome assembly or into fox-dog evolutionary differences. However, the fact that 93.5% of the transcripts mapped to a single scaffold in the draft fox genome suggests that genes in this assembly are much less fragmented than the numbers reported for other *de novo* assemblies, such as the 69.7% of genes mapping to a single scaffold reported by Alkan et al. (2011), though some bias could be introduced by the selection of only a subset of transcripts.

Estimating Chimerism

The alignment of the 500 largest fox scaffolds against the dog genome and subsequent analysis of the z-scores generated a list of the dog chromosomes significantly syntenic to each scaffold (Appendix C). Most of the scaffolds (84%) mapped to a single chromosome, while 10% were spilt among two and 2% among three chromosomes (Figure 7). Five scaffolds (1%), the longest of which was approximately 252 Kb long (Scaffold337) could not be placed because they mapped with similar affinity to each chromosome, and 13 scaffolds (3%), all of which were shorter than 253 Kb, were identified as mapping to 4, 5 or 6 different chromosomes.

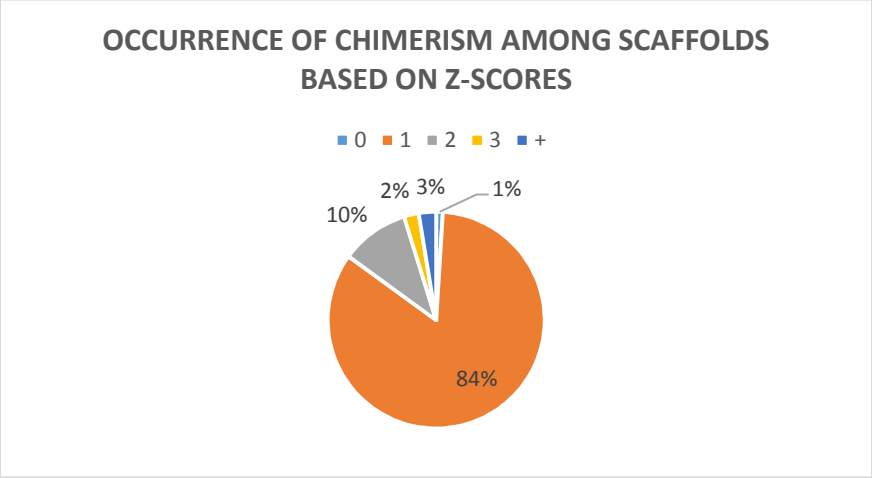


Figure 7: Percent of scaffolds (of the 500 largest scaffolds) mapping to zero (light blue), one (orange), two (grey), three (yellow) or more (dark blue) dog chromosomes, as estimated by calculating the z-scores of the maximum alignment score against each chromosome.

For the scaffolds that could not be assigned a position in the dog genome, the mapping results were visually inspected to determine whether any syntenic relationship to the dog could be discerned. The raw mapping results from three scaffolds (348, 378, and 384) suggested synteny with a single chromosome, though none of the others could be assigned to a position in the dog genome. Nonassignability could occur for a number of reasons, including the presence of a large number of long repetitive segments or synteny to an uncharacterized region of the dog, as the uncharacterized regions were excluded from the current analysis.

This second approach, where each nucleotide along the fox scaffolds was assigned to a position in the dog genome, generated visual images depicting the relationship between the scaffold and the dog genome. These figures allowed for validation of the results generated based on the z-scores. Visualizations of three example scaffolds are provided in Figure 8.

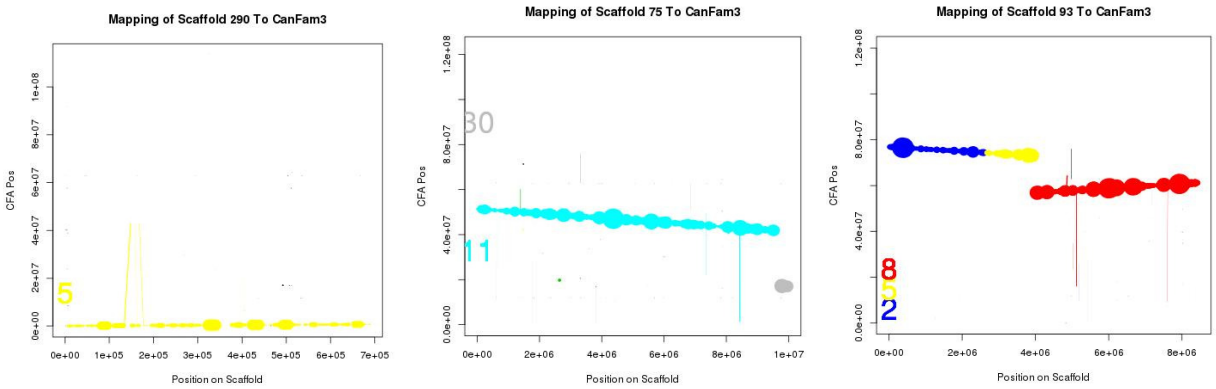


Figure 8: Example of nucleotide-level best-hit mapping indicating synteny with one (Scaffold 290, furthest left), two (Scaffold 75, center) or three (Scaffold 93, furthest right) dog chromosomes. Scaffold position is represented along the x-axis, while chromosome position corresponds to the y-axis. Dog chromosomes are color-coded, with the chromosome number in the corresponding color printed next to the y-axis, and point size is scaled by LAST score.

Though dog and fox are very closely related at the nucleotide level, their karyotypes, as discussed previously, have been rearranged extensively, with most fox chromosomes comprising 2-3 dog chromosomes. Therefore, it was necessary to evaluate all "chimeric" scaffolds to determine whether they were likely to indicate bioinformatic errors, or whether they encompassed regions of the genome where syntenic blocks distinct in the dog genome lie adjacent in the fox genome. Of the chimeric scaffolds, five mapped to dog chromosomes that are known to map to a single chromosome in the fox. The details of these five scaffolds are provided in Table 5. Additional analysis will be necessary to validate these scaffolds, and to determine whether any of the apparent bioinformatic chimeras correspond to small, previously unidentified, inter-chromosomal rearrangements between fox and dog.

In summary, 15% of the 500 largest fox scaffolds are expected to contain sequences from disparate regions of the genome that have been erroneously assembled. This number may not be precise: it's possible that some of the scaffolds that contain sequences from a single dog chromosome or from different dog chromosomes that map adjacently in the fox genome also represent bioinformatic errors in assembly. Intrachromosomal misassembly will be difficult to identify, but the probability of a chimeric scaffold containing "fused" dog chromosomes by chance alone can be modeled very simply.

Bioinformatic chimeras are generated due to sequence similarity among different chromosomes. Assuming that sequence similarity between a chromosome and all other chromosomes is randomly distributed, 5.6% of chimeric scaffolds would be expected to include sequences from chromosomes

mapping to the same fox chromosome by chance alone, based on the known pattern of synteny between the dog and fox chromosomes. Because there were 88 cases where two chromosomes mapped to the same scaffold, we would expect 5 of these 88 cases to include two chromosomes that are fused in the fox simply due to chance. Therefore, whether the chimeras in Table 5 are real or bioinformatic will need to be further analyzed.

SCAFFOLD	*DOG CHROMOSOME*	*CORRESPONDING LOCATION IN FOX*
1	CFA 28	VVU 15
	CFA 30	VVU 15
7	CFA 33	VVU 1
	CFA 12	VVU 1
9	CFA 18	VVU 5
	CFA 38	VVU 5
	CFA 20	VVU 9
239	CFA 35	VVU 12
	CFA 5	VVU 12
414	CFA 19	Both VVU 4 and VVU 5
	CFA 18	VVU 5
	CFA 8	VVU 6

Table 5: Five fox scaffolds appear to be chimeric based on mapping to the dog, but in fact map to regions of the fox where the syntenic blocks corresponding to different dog chromosomes lie adjacent. Scaffolds 9 and 414 map to three dog chromosomes, and in both cases one of those chromosomes is not expected to lie adjacent to the others in the fox (i.e. is predicted to be a bioinformatic chimera). The syntenic blocks comprising CFA 19 are split apart in the fox karyotype, with one composite syntenic block located on VVU 4 and another on VVU 5.

Conclusion

The results of the current analysis indicate that the bioinformatic errors frequently found in *de novo* genome assembly using NGS technology do affect the draft red fox genome assembly. First of all, the length of the assembly as reported by BGI, 2.29 Gbp, is between 13-18% shorter than flow cytometric estimates of the fox genome size, making the loss of genomic material in the red fox assembly consistent with the 16% reduction in length during *de novo* assembly estimated by Alkan et al. (2011). Second, the genome is comprised of a large number of very short scaffolds, and analysis of these scaffolds suggests that many of them will complicate assembly of the genome into chromosomes because they are comprised of repetitive elements (15.8% of the scaffolds in the genome) or are very similar to other sequences in the genome (an estimated 80.7% of the scaffolds in the genome). Finally, even among the largest assembled scaffolds, the large-scale structural integrity of approximately 15% is questionable, as they appear to be chimeric. These types of errors are expected in a *de novo* assembly, but suggest that the loss genetic information and the misassembly of sequences must be considered in downstream analysis including chromosome assembly.

In some respects, however, the genome assembly performed better than other *de novo* assemblies. The successful mapping of the sequenced libraries back onto the fox genome clearly indicates that the assembly has, on the large-scale, recapitulated the genomic biology of the fox. Additionally, only a small number of dog transcripts were found to be fragmented among multiple scaffolds (4.0%) or incompletely sequenced (2.5%), numbers much lower than those reported in previous research (Alkan et al., 2011), based on the alignment of experimentally validated dog transcripts. These results suggest that the draft genome holds the potential to, with refinement, offer significant potential for genetic and genomic study of the fox.

CHAPTER III
CHROMOSOME ASSEMBLY OF THE RED FOX GENOME

Overview/Motivation

The primary objective of the current analysis is to develop a chromosomal assembly for the red fox genome. Assembling genomic sequences into chromosomes will facilitate the use of the genome assembly in addressing questions of interest to biologists studying the fox, such as the investigation of regions of interest in complex behavioral traits (Johnson, Wittgenstein, et al., 2015; Kukekova, Trut, et al., 2011) or of regions containing interesting cytogenetic features, such as interstitial telomeric sequences (Becker et al., 2011). The current study approaches the challenge of revising a draft genome comprised by 676,878 scaffolds into seventeen chromosomes by identifying sequence-level synteny between the fox and the dog genomes.

In order to identify syntenic sequences, the fox genome was aligned to the dog genome using LASTZ (Harris, 2007) and the results were analyzed with the Kent utilities (Kent, Baertsch, Hinrichs, Miller, & Haussler, 2003) to produce syntenic nets. This alignment and analysis was repeated to develop a syntenic net for an outgroup genome, the cat, aligned against the dog. The chains and nets were then analyzed with the program Reference Assisted Chromosome Assembly, or RACA (Kim et al., 2013), which assist with the construction of reference-based chromosomes in *de novo* assemblies by constructing syntenic fragments (SFs) from syntenic nets. SFs are sequence fragments that map to unique, continuous locations in both the dog genome and fox scaffolds, and may also map to one or more positions in the cat. These SFs served as the basis for the assembly of the chromosomes.

The SFs were first ordered along each dog chromosome and then evaluated in the context of known fox-dog synteny (Kukekova et al., 2007; F. Yang et al., 1999). Synteny between the fox and dog has been mapped through a reciprocal chromosome painting study (F. Yang et al., 1999) and through the construction of a fox meiotic linkage map using dog-derived microsatellite markers (Kukekova et al., 2007). Therefore, once the dog chromosomes were constructed from the SFs, it was possible to assign the SFs to positions on the fox chromosomes based on the known sytenic relationships between the dog and fox chromosomes. Additionally, the corresponding cat chromosomes were also identified and their order along the dog chromosomes compared against known dog-cat synteny (Davis et al., 2009; F. Yang et al., 2000) to determine the extent to which the current assembly recapitulated previous results. This analysis allowed for each SF to be ordered along the dog, cat, and fox chromosomes. Additionally, whereas the fox and cat chromosomes have not previously been compared with chromosome painting or meiotic linkage, this assembly results in a cat-fox syntenic map, albeit one based on bioinformatic rather than experimental mapping results.

The secondary goal of the analysis was to refine the fox-dog syntenic map and the scaffolds of the draft fox assembly. One such refinement was the reduction of gaps in the dog-fox comparative map. Though the fox and dog karyotypes are extensively rearranged, almost all of the fox chromosomes can be broken into 2-3 syntenic blocks that correspond to whole dog chromosomes. However, there are four dog chromosomes that map to two different syntenic regions each in the fox genome. Previous studies (Becker et al., 2011; Kukekova et al., 2007) have sought to identify the breakpoints, but their studies identified breakpoint regions at the megabase scale. In the present study, these breakpoint regions are significantly reduced through the placement of the SFs and the identification of corresponding cat-dog-fox synteny.

Another outcome was the refinement of the chimeric scaffolds. Because the SFs identify precise locations in the dog that align to precise locations in the fox, it was possible to determine where the previously developed dog-derived microsatellite markers (Kukekova et al., 2004, 2007) would align in the fox scaffolds. Because these markers have been used previously in the development of the fox meiotic linkage map (Kukekova et al., 2007), their relative positions along the fox chromosomes are known. Therefore, it was possible to determine whether markers mapping to a single scaffold were known to map to different fox chromosomes, allowing for some scaffolds to be identified as bioinformatically chimeric.

Through this analysis, synteny between the fox and the dog and between the cat and the dog is refined to the nucleotide level through pairwise mapping of the genomes, and the resulting syntenic fragments are used to assemble fox chromosomes. The analysis produces a number of deliverables that will be useful to canine, feline, and vulpine genetic researchers, including the fox chromosome assembly and a cat-fox comparative alignment.

Methods

Syntenic Chains/Nets

Synteny between the dog and fox genomes is well established at the chromosome level, in terms of chromosomal segments identified with chromosome painting (F. Yang et al., 1999) and centimorgan-scale estimations of their sizes and positions identified with meiotic linkage mapping (Kukekova et al., 2007). Therefore, synteny identified at the nucleotide level between the fox scaffolds and the dog chromosomes could be integrated with the results of physical mapping studies to reveal the relationship between the fox scaffolds and the fox chromosomes. This process required the construction of a dog-fox pairwise genomic alignment comprised of chains and nets, which are two bioinformatic constructs used frequently in inter-species genomic comparisons.

Chains are ungapped collinear alignments of sequence between two species. To make a chain, pieces of the target genome must be aligned to the reference using a program such as LASTZ (Harris, 2007), and then these alignments are merged where they overlap and pruned by score (Figure 9). A net, on the other hand, is an assembly of chains which can be strung together through the addition of gaps. In

the present analysis, the target genome (the 500 largest scaffolds of the draft fox genome, referred to as vv2) and outgroup genome (felCat5) (Montague et al., 2014; Pontius et al., 2007) were each aligned against the reference genome (canFam3.1) (Lindblad-Toh et al., 2005) using LASTZ. Because rearrangements are expected over the course of millions of years of evolution, chromosomes and scaffolds are partitioned into smaller "chunks" prior to alignment to make the alignment more resilient against rearrangements. This partitioning and subsequent alignment was conducted using a set of scripts provided by Marta Farré Belmonte and Denis Larkin at the Royal Veterinary College in London, U.K. that allowed for the LASTZ alignments to be conducted in parallel on a computing cluster (M. Farré Belmonte, personal correspondence, February 19, 2015). The dog chromosomes were partitioned into pieces of 40,010 Kbp that overlapped each other by 10 KBp, whereas the fox scaffolds and cat chromosomes were partitioned into 20,000 Kbp chunks prior to being aligned against the dog.

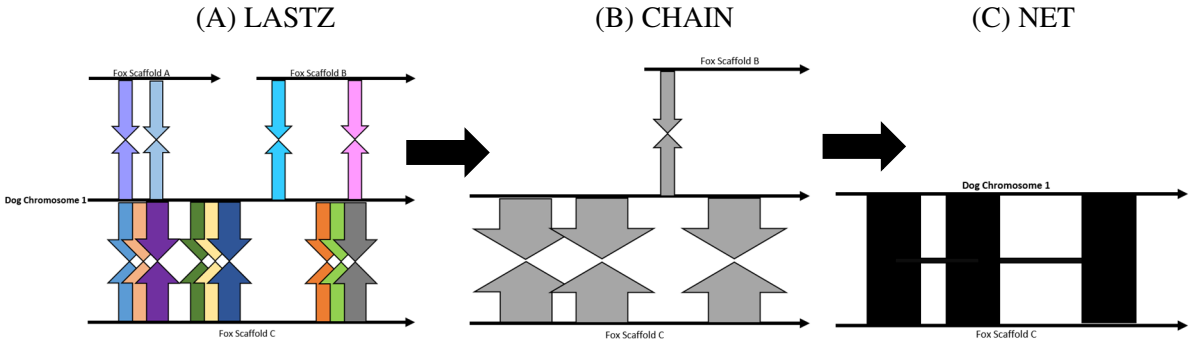


Figure 9: An illustration of the chaining and netting of LASTZ alignment data. Many individual LASTZ alignments (A), highlighted by the plethora of colorful arrows whose widths represent score, will be produced between the dog chromosome (middle line) and different fox chromosomes (top and bottom lines). Arrows of the same color pointing at each other represent corresponding alignments in the dog and fox genomes. In the chaining step (B), these alignments are pruned to include only the highest-scoring alignments and are merged (grey arrows) based on overlap. No gaps are allowed in the chaining phase. The chains are then netted together with gaps allowed (C).

LASTZ takes several parameters that influence the scoring schema, including the cost of opening or extending a gap and a score threshold below which alignments should be discarded. UCSC has published the parameters they recommend for building chains and nets for a number of interspecies alignments. Because of the similarity of the divergence time between the fox and dog and that of human and chimp, UCSC's human-chimp alignment parameters were selected for the fox-dog alignment (available online at http://hgdownload.cse.ucsc.edu/goldenpath/hg19/vsPanTro3/). The LASTZ parameters included a gap opening penalty (O) of 600, a gap extension penalty (E) of 150, a minimum score threshold for inclusion of an alignment on the first pass (K) of 4500 and on the second pass (L) of 2200, a minimum score threshold for interpolation (H) of 2000, and the default LASTZ scoring matrix.

The chaining step of the analysis also takes several parameters, including the minScore cut-off parameter, which here was set to 5000, and the linearGap parameter, which was set to "medium." However, because of the extensive karyotype rearrangement between dog and fox despite the short

divergence time, the alignment was also tested on CFA 24 with the linearGap parameter set to "loose", as recommended by collaborators on the original RACA publication (M. Farré Belmonte, personal communication, April 17, 2015). Based on these results, the fox-dog alignment proceeded with the UCSC-recommended human-chimp alignment parameters and linearGap set to medium.

As for the cat-dog alignment, a previous version of the dog genome (canFam2) had been aligned to a previous version of the cat genome (felCat3), and the parameters for this alignment are available online at http://hgdownload.cse.ucsc.edu/goldenPath/canFam2/vsFelCat3/. These same parameters were used in the current analysis, with the LASTZ parameters set to use the default scoring matrix, O=400, E=30, K=3000, L=2200, H=2000 and M=50 (which dynamically masked dog sequences that appeared more than 50 times). The chain parameters used were a score threshold of 3000 and a linearGap of medium.

Because this alignment was a large-scale computational procedure requiring over 40,000 individual LASTZ alignments for the fox-to-dog comparison and over 10,500 for the cat-to-dog, High-Performance Computing resources, namely the Carl R. Woese Institute for Genomic Biology's Biocluster, were utilized. The Biocluster has over 2824 cores available across 35 nodes and 5 queues (Carl R. Woese Institute for Genomic Biology at the University of Illinois at Urbana-Champaign, 2015). The scripts, which were written by Marta Belmonte Farré for Sun Grid Engine (a propriety queue management system released by Oracle), were altered by consultants hired from the HPCBio group associated with the Roy J. Carver Biotechnology Center and the Carl R. Woese Institute for Genomic Biology at the University of Illinois at Urbana-Champaign to be compatible with the open-source Terascale Open-source Resource and QUEue Manager (TORQUE) used by Biocluster. The HPCBio consultants also altered the submission structure of the script so that it would work given the restrictions on the Biocluster's Highthroughput queue, which limits job submissions to 550 per user at any time. Using these altered scripts, jobs were manually queued in batches of 550, which allowed for each multi-species alignment to be run in under a week. For each of the fox and cat genomes mapped against an individual dog chromosome, three output files were produced: a .chain file, a .net file, and a multiple alignment .psl file. The results of the alignment were visualized as a custom track on UCSC Genome Browser against canFam3.1 for visual evaluation.

Reference Assisted Chromosome Assembly (RACA)

The software Reference Assisted Chromosome Assembly (Kim et al., 2013), or RACA, was used to build syntenic fragments (SFs) corresponding to unique, continuous sequences in both the dog and fox genomes and, where possible, in the cat genome as well. RACA constructs SFs by comparing the alignments of the target and outgroup species' genomes to the reference genome using an algorithm developed to identify and reconstruct regions of a genome that are likely to have been contiguous in an

ancestral species (Ma et al., 2006). In order to construct SFs, RACA merges collinear alignments of syntenic nets, which are described above. RACA was instructed to identify SFs of no less than 150 Kbp in size using the "resolution" parameter. While RACA can provide additional analysis, such as estimating the probability of adjacency among two SFs, these analyses were not necessary for the current study because of the pre-existing physical maps. RACA requires as input a number of files providing information about the assembly and the target, reference, and outgroup species, which are discussed in detail below. Data used exclusively in the estimation of adjacency probabilities is not discussed, because the program was not used for that purpose in the current analysis.

Scaffold Sizes

RACA requires that the length of each scaffold be provided in a tab-delimited .txt file. The measurement of scaffold sizes was previously discussed in Chapter II. The measured lengths of the scaffolds were converted to a tab-delimited format using Python and provided as input to RACA.

Estimating Dog/Fox Divergence

A Newick tree indicating the branch lengths between the target, reference, and outgroup species must be provided to RACA as input. In order to estimate the phylogenetic relationship between the target (fox), reference (dog) and outgroup (cat) species' genomes, the .net files generated during the creation of the chain and net files was converted to .maf format using netToAxt and axtToMaf from the Kent Utilities (Kent et al., 2003), then concatenated, and finally analyzed using phyloFit (Hubisz, Pollard, & Siepel, 2011) to produce a Newick tree. Nucleotide substitution rates were estimated under the reversible nucleotide substitution model (Tavare, 1986), as recommended by Yang (1994). The Newick tree was then visualized using Polydendron (Gilbert, 1999).

Construction of Fox Chromosomes

The output from RACA was used in all further analysis, including the construction of the fox chromosomes. This output was a list of syntenic fragments that identified a position in the dog genome, the corresponding position on a fox scaffold, and, when available, the corresponding position in the cat genome. The primary goal of the current study was to assemble these SFs into the fox's sixteen autosomes and X-chromosome. Thus, the syntenic fragments were transformed into a .csv format using Python and sorted according to their order on the dog chromosomes. For each fox chromosome, the SFs mapping to the corresponding dog chromosomes were selected and assigned an order based on the order and direction of the dog chromosomes on the fox chromosome, as identified by the meiotic linkage map (Kukekova et al., 2007). Through these steps, an Excel worksheet was produced for each fox chromosome that contained a list of all SFs assigned to that chromosome, their order on the fox

chromosome, their position in the dog genome, their position on the fox scaffolds, and their position in the cat genome (where available).

The order of the SFs was then visually inspected to ensure that their ordering was consistent with known dog-cat and dog-fox synteny. First, the chromosomes were scanned to identify any fox scaffolds that contained multiple SFs whose order in the dog genome was out of order or in an inconsistent orientation. For example, if the first five megabases of a scaffold mapped in a positive orientation to the chromosome and the next ten mapped in a negative orientation, this shift in orientation would suggest either a break in synteny between dog and fox or an error in scaffold assembly. Where available, the relationship between the cat and the dog was considered to see if the discordance could be attributed to either a dog-specific or fox-specific rearrangement. Next, for each dog chromosome, the SFs were arranged according to their order in the cat genome in order to see whether any scaffolds overlapped cat sequences out of order. This method had the potential to reveal differences in the cat compared to the canid genomes, and could also potentially identify rearrangements in the dog relative to the ancestral state. Rearrangements shorter than 100 bp were ignored.

These results were also compared against previous work using chromosome painting in the dog and cat (Davis et al., 2009; F. Yang et al., 2000) in order to identify whether the fox-based synteny between dog and cat diverged from direct comparisons among those two species. This analysis involved re-sorting the SFs into their order within the cat genome. Any anomalies were noted for future analysis, as possible areas of dissolution of cat-canid synteny.

Marker Analysis

One of the major potential problems identified with the assembly in Chapter II was the presence of chimeric scaffolds. However, because the SFs link positions in the scaffolds to positions in the dog genome, it is possible to assign microsatellite markers to positions in the fox scaffolds based on their known positions in the dog genome. Because the physical order of the markers mapping to the fox genome is known (Kukekova et al., 2007), mapping the markers onto the scaffolds can provide insight into whether SFs within a single scaffold belong to the same linkage group within the fox genome.

Thus, 411 markers previously used in linkage mapping of the fox (Kukekova et al., 2004, 2007; Kukekova, Temnykh, Johnson, Trut, & Acland, 2012; Kukekova, Trut, et al., 2011) were assigned locations in the SFs based on their positions in the dog genome. Because the SFs identify specific, collinear regions of the dog and fox genomes, it was possible to map each marker to an SF, and thus also to a scaffold, by writing a simple Python script. The scaffolds were analyzed to build a list of all scaffolds whose composite SFs did not unanimously identify synteny with a single dog chromosome; these were a subset of the chimeras identified previously with other methods (Chapter II). If the scaffold contained markers in regions that were assigned to two different dog chromosomes, the positions of the markers in

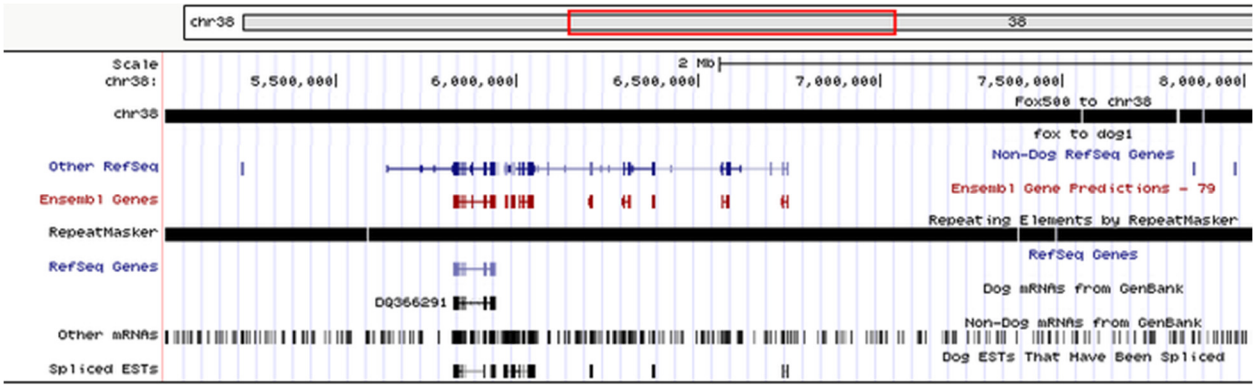
the fox genome could be examined to determine whether the scaffold was indeed chimeric. However, if markers could not be assigned to one or more of the regions, then further evaluation would be necessary to determine whether the chimerism was bioinformatic or whether a previously undetected rearrangement could potentially produce the observed chimerism.

Results/Discussion

Syntenic Chains/Nets

Visualizing the netted alignment as a custom track on the UCSC Genome Browser revealed that coverage of the dog genome by the nets assembled from the fox-dog genome alignment was high. The alignment of the fox to the dog had fewer gaps than the cat-dog alignment (Figure 10), as would be expected based on evolutionary divergence. These visualizations suggested that the nets constructed using the Kent Utilities had identified fox-dog and cat-dog synteny genome wide.

(A) Fox-Dog Net, CFA 38



(B) Cat-Dog Net, CFA 38

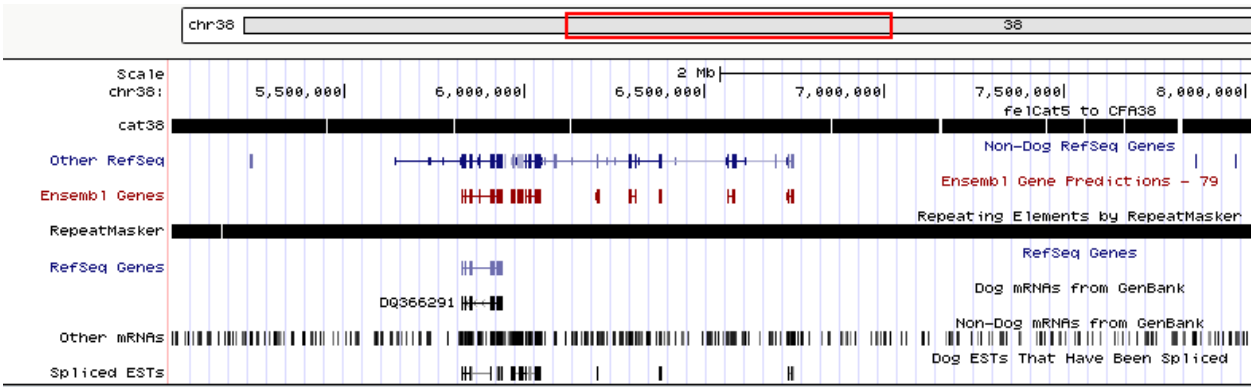


Figure 10: UCSC Genome Browser visualization for a ~3Mbp region on CFA 38 of (A) the fox-dog net and (B) the cat-dog net. Very few gaps (white breaks in the black bar labeled “chr38” or “cat38” in A and B, respectively) are apparent. More gaps are apparent in the cat-dog alignment than in the fox-dog alignment.

Divergence Estimates

PhyloFit estimated the following Newick tree for the three species: ((canFam3: 0.0156945, vv2: 0.00985954): 0.0890152, felCat5: 0.0890152). This tree is visualized as Figure 11. The branch lengths in the figure and in the nested parenthetical format (above) represent substitution rates. Interestingly, although the analysis is outside the scope of the present study, these results suggest that the dog lineage may be slightly more diverged from cat than the fox is, potentially consistent with findings of moderately accelerated evolution in some dog gene families (Lindblad-Toh et al., 2005), though any conclusions from this data would require further analysis such as that suggested by Tajima (1992).

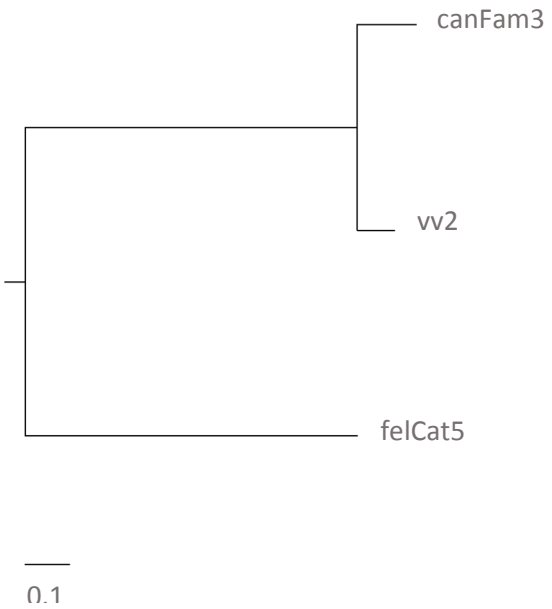


Figure 11. Phenogram constructed from the Newick tree and visualized with Phylodendron. Branch lengths to scale. For the branch names, canFam3 refers to the published dog reference genome, vv2 to the 500 largest fox scaffolds, and felCat5 to published cat reference genome.

Chromosome Assembly

The assembly of the scaffolds of the draft fox genome into chromosomes was the primary goal of this study. The analysis was approached from multiple angles to assess synteny among all the three species. The results used in the assembly, and the results obtained from the assembly itself, are discussed below.

Assignment of Syntenic Fragments

RACA identified 428 syntenic fragments (SFs) conserved between the dog and the fox. The SFs were ordered according to their position in the dog chromosome, so that all SFs on a single dog chromosome were numbered continuously. For each SF, the program identified the position on the dog chromosome and the corresponding position on the fox scaffold. If corresponding sequence was identified in the cat, the SF would also include this information, but the fragment was not required to map

continuously in the cat genome. For example raw SF data, see Appendix D. The ranges of SFs assigned to each fox chromosome and to each composite dog chromosome are indicated in Table 6.

The SFs were ordered along each chromosome based on their position within the dog, and then this order was shifted to reflect any regions where the fox showed greater synteny to the cat than the dog. The final results from this analysis are included in Appendix D.

Fox Chromosome	Total Number SFs	Dog Chromosome	SF Numbers Assigned	Total Number of SFs
1	27	1	409-428	20
		33	199	1
		12	17-22	6
2	27	9	313-324	12
		13	27-31	5
		2	32-41	10
3	24	36	251-254	4
		34	232-239	8
		6	276-287	12
4	23	19	104-108	5
		32	192-198	7
		4	221-231	11
5	29	19	109-117	9
		1	405-408	4
		18 (P)	98-103	6
		38	257-259	3
		18 (D)	91-97	7
6	27	22	139-150	12
		8	298-312	15
7	26	16	62-75	14
		14	42-53	12
8	19	27	161-164	4
		17	76-90	15
9	17	25	214-220	7
		20	118-127	10
10	22	26	391-404	14
		15	54-61	8

Table 6: The assignment of SFs to each fox chromosome, and to each composite dog chromosome. Information about the individual SFs indicated in the "SF Numbers Assigned" column are included in Appendix D. (P)=proximal, (D) = distal, based on fox position. Continues on next page.

11	21	21	128-138	11
		23	151-160	10
12	43	11	1-16	16
		35	240-250	11
		5	260-275	16
13	15	13	23-26	4
		29	171	1
		7	288-297	10
14	14	24	206-213	8
		3	200-205	6
15	26	31	177-191	15
		30	172-176	5
		28	165-170	6
16	9	37	255-256	2
		10	325-331	7
X	59	X	332-390	59

Table 6 (cont.): The assignment of SFs to each fox chromosome, and to each composite dog chromosome.

Chromosome Synteny

Because the fox chromosomes were assembled from SFs that were built based on synteny to the dog and cat genomes, comparative genomic analysis allowed for the construction of syntenic maps of each fox chromosome (Figure 12) compared to the dog and cat chromosomes. For the most part, synteny between the fox, dog, and cat was consistent with previously published results, although this is the first explicit analysis of fox-cat chromosomal synteny.

In two cases, the fox karyotype appears to be more similar to that of the cat than the dog: on VVU 4 (fox chromosome 4), the segment mapping continuously to FCA B1 (cat chromosome B1) is split into two chromosomes in the dog, CFA 19 (dog chromosome 19) and CFA 32 and on VVU 13, the segment mapping continuously to FCA F2 is split into two chromosomes in the dog (CFA 13 and 29). Another region of potential evolutionary interest is on VVU 6, where the breakpoints between CFA 22 and 8 and between FCA A1 and B3 appear to be separated by 485 Kbp, a feature which suggests possible breakpoint reuse, though this region was not identified in an analysis of re-used breakpoints within *Canidae* (Becker et al., 2011). In general, however, the assembly closely adhered to the syntenic regions identified for dog and fox by linkage mapping, with only six novel rearrangements between dog and fox, all smaller than 10Mbp, identified.

While small rearrangements of sequence in the cat compared to the fox were more common, only a few syntenic regions unidentified between cat and dog in the RH-map (Davis et al., 2009) were found in

the current nucleotide-level analysis, and all are shorter than 600 Kbp. For example, the region of FCA B1 estimated as syntenic to VVU 10 and CFA 15 is larger in the present analysis than was reported in the dog-cat RH map (Davis et al., 2009), and small segments (230 and 221 Kbp, respectively) of FCA B3 are assigned by this analysis to CFA 13/VVU 2 and CFA 10/ VVU 16 that were previously unidentified. Similarly sized refinements can be made for synteny of FCA B4 and CFA 3/VVU 14, FCA C2 and CFA 33/VVU 1, and FCA D3 and CFA 1/VVU 5.

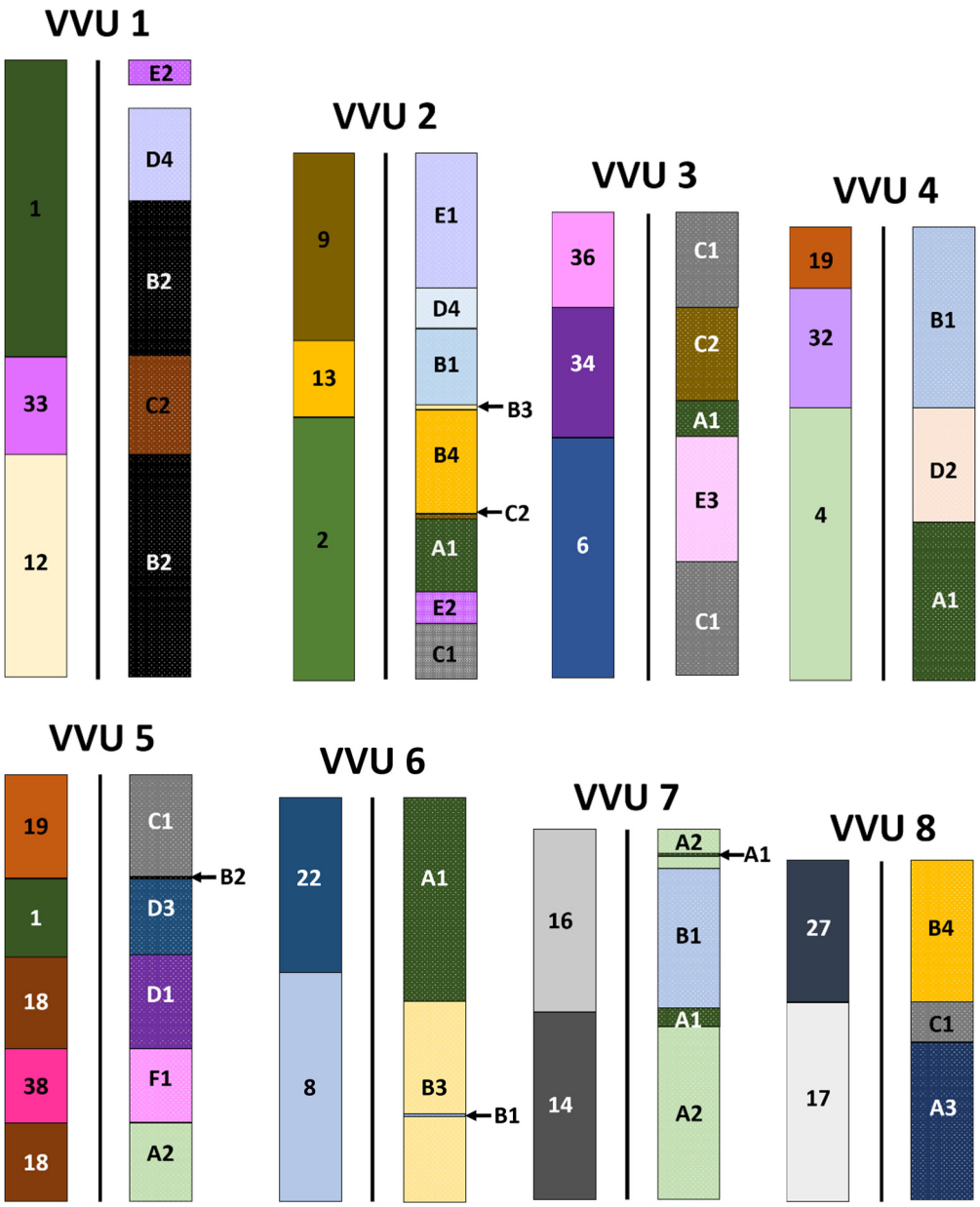


Figure 12: Synteny between the fox chromosomes (black bars) and the dog (indicated to the left) and the cat (indicated to the right). Comparative maps were constructed based on synteny detected through pariwise genome alignment. Continues on next two pages.

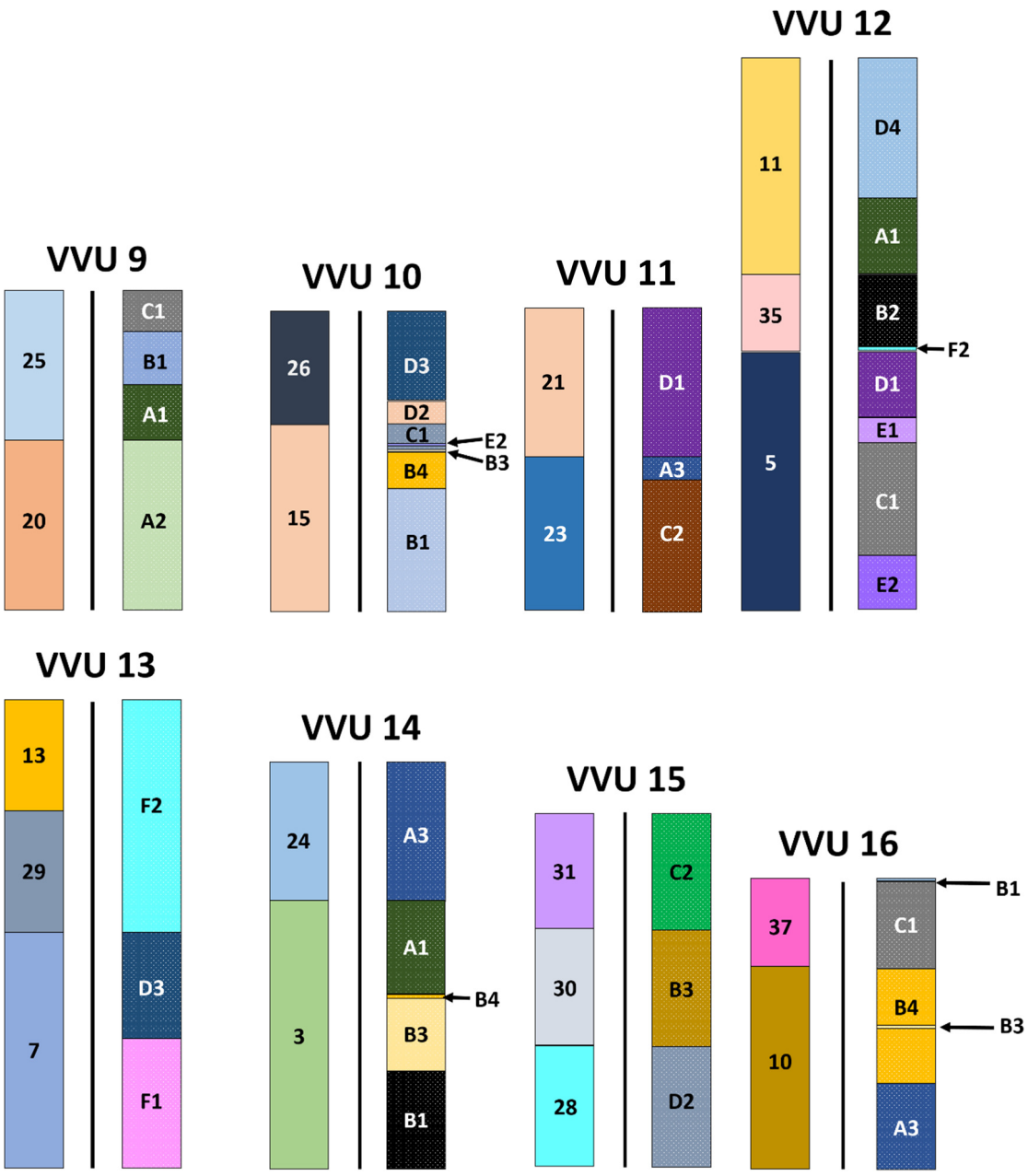


Figure 12 (cont.): Comparative chromosomal maps for fox (black bars), dog (to the left of the bars) and cat (to the right of the bars).

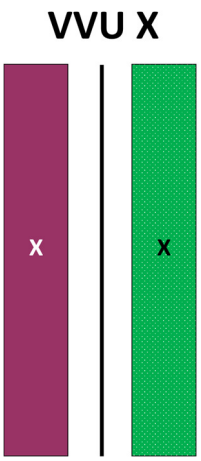


Figure 12 (cont.): Comparative chromosomal maps for fox (black bars), dog (to the left of the bars) and cat (to the right of the bars).

Estimation of Gaps

In total, 98.95% of canFam3.1 by nucleotide count was covered by the 428 SFs. The percentage of each dog chromosome missing from the chromosome assemblies varied by chromosome, and is indicated in Table 7. Gaps include the spaces between adjacent SFs and between terminal SFs and the ends of the chromosomes. By definition, SFs do not overlap. At least 97% of the sequence of each chromosome was captured in the SFs, with >99% of most chromosomes included in the assembly. The best-assembled chromosome was CFA 33, which maps to VVU 1. This chromosome was contained within a single SF and is missing only 291 bp of sequence, located within the telomeres. This result suggests that the assembly in this region of VVU 1 has been largely successful. The dog chromosome with the largest percentage of its nucleotides not covered in the assembly is CFA 31, which is located on VVU 15. For the majority of dog chromosomes, at least 99% of the sequence by length is represented in the fox chromosomal assembly.

Refinement of Known Gaps

Four dog chromosomes do not map in conserved syntenic blocks to the fox, with CFA 1, 13, 18 and 19 each mapping to two regions of the fox chromosomes. The specific regions of each dog chromosome belonging to each syntenic block were previously unknown. Based on the nucleotide-scale syntenic blocks identified by RACA, synteny in these four regions was analyzed to determine whether the size of the gaps could be reduced. SFs were analyzed primarily in terms of whether they fell into the gap identified with the meiotic linkage map, as coordinates for these gaps were available in canFam3.

CFA #	CFA Length	Number of Gaps	Total Len. of Gaps (bp)	% CFA Falling in Gaps	Average Gap Len.	Standard Deviation	Median Len. of Gaps
1	122,678,785	25	1,708,319	1.39%	68,332	143,224.6	6,030
2	85,426,708	11	2,085,949	2.44%	189,631	441,482.0	100
3	91,889,043	7	310,162	0.34%	44,308	83,284.1	7,045
4	88,276,631	12	105,118	0.12%	8,759	18,855.7	104
5	88,915,250	17	408,190	0.46%	24,011	42,856.3	35
6	77,573,801	13	116,739	0.15%	8,979	13,823.1	5,992
7	80,974,532	11	741,271	0.92%	67,388	135,264.1	1,800
8	74,330,416	16	240,585	0.32%	15,036	29,654.6	573
9	61,074,082	13	1,444,122	2.36%	111,086	180,706.1	37,045
10	69,331,447	8	215,319	0.31%	26,914	42,287.9	5,877
11	74,389,097	17	1,442,100	1.94%	84,829	178,429.0	3,791
12	72,498,081	7	49,466	0.07%	7,066	9,099.5	4,587
13	63,241,923	10	492,454	0.78%	49,245	84,574.8	12,019
14	60,966,679	13	490,777	0.80%	37,752	86,930.7	0
15	64,190,966	9	475,590	0.74%	52,843	94,885.8	30
16	59,632,846	15	1,670,793	2.80%	111,386	186,527.6	18,549
17	64,289,059	16	522,726	0.81%	32,670	60,803.4	6,826
18	55,844,845	14	902,357	1.62%	64,454	156,703.6	2
19	53,741,614	15	1,630,679	3.03%	108,711	187,715.2	688
20	58,134,056	11	326,019	0.56%	29,638	52,894.4	1,121
21	50,858,623	12	148,543	0.29%	12,378	26,341.0	1,734
22	61,439,934	13	645,208	1.05%	49,631	83,999.7	2,919
23	52,294,480	11	20,231	0.04%	1,839	4,223.0	0
24	47,698,779	9	124,183	0.26%	13,798	25,051.57	591
25	51,628,933	8	610,630	1.18%	76,328	122,925.8	4
27	45,876,710	5	157,895	0.34%	31,579	35,725.3	20,967
28	41,182,112	7	43,578	0.11%	6,225	14,928.9	74
29	41,845,238	2	122,422	0.29%	61,211	70,908.7	61,211
30	40,214,260	6	182,986	0.46%	30,497	56,847.6	7,835
31	39,895,921	16	1,676,572	4.20%	104,785	235,452.2	300
32	38,810,281	8	444,453	1.15%	55,556	83,468.1	5,236
33	31,377,067	2	291	0.00%	145	64.3	145
34	42,124,431	9	391,128	0.93%	43,458	68,216.0	10,388
35	26,524,999	12	91,765	0.35%	7,647	17,481.4	189
36	30,810,995	5	5,180	0.02%	1,036	1,597.2	267
37	30,902,991	3	146,493	0.47%	48,831	84,576.9	1
38	23,914,537	4	43,253	0.18%	10,813	10,079.6	10,216
X	123,869,142	60	3,866,474	3.12%	64,441	128,038.3	14,283

Table 7: Because SFs were defined primarily in terms of their position on the dog chromosomes, this table indicates putative gaps in the fox assembly according to the regions covered in the dog.

Significant reductions in size were made in the unmapped regions between syntenic blocks on CFA 1, 13, 18 and 19. Both continuity within the scaffolds and synteny to the cat were considered in refinement. The reduction in size for each gap is summarized in Table 8. These revised breakpoints fall within the breakpoint regions identified in both Becker et al. (2011) and the meiotic linkage map (Kukekova et al., 2007). As is evident in Table 8, the genome-by-genome alignment conducted in the present study has significantly reduced the size of the unassigned region on each dog chromosome.

Dog Chr.	Gap (Becker)	Gap Size (Becker) (Mbp)	Gap (Kukekova)	Gap Size (Kukekova) (Mbp)	Gap (revised)	Gap Size (revised)
CFA1	24,600,000 to 25,700,000	1.1	21,746,216 to 27,254,893	5.51	25,534,824 to 25,579,247	44,423 bp
CFA13	37,800,000 to 38,600,000	0.8	25,138,447 to 42,432,978	17.3	38,258,211 to 38,277,954	19,743 bp
CFA18	24,400,000 to 26,000,000	1.6	21,513,345 to 29,484,473	7.97	25,259,332 to 25,332,083	72,751 bp
CFA19	18,800,000 to 22,100,000	3.3	17,989,216 to 27,017,104	9.03	19,878,341 to 20,333,685	455,344 bp

Table 8: Gaps in the assignment of dog chromosomes to syntenic regions on the fox chromosomes. Conversions of the positions in the Becker paper, which uses canFam2, were done using UCSC Genome Browser. Previous gap estimates from Becker = Becker et al. (2011); Kukekova = Kukekova et al. (2007) and subsequent work.

Marker Analysis

The mapping of markers to SFs was undertaken in the hopes that the marker placement could confirm whether or not SFs mapping to chimeric scaffolds belonged to a single linkage group as identified by Kukekova et al. (2007). If markers could be placed in SFs that mapped to each of the unique dog chromosomes assigned to that scaffold, then the scaffold's chimerism could be determined to be either bioinformatic or biological depending on the marker linkage.

Of the 411 markers mapped to the syntenic fragments in this analysis, 409 could be assigned to a location within a syntenic fragment. Table 9 lists all scaffolds identified as chimeric in the present analysis (i.e. all scaffolds that contained SFs mapping to multiple different dog chromosomes). These scaffolds are a subset of those discussed in Chapter II. Eleven of these scaffolds contained markers that allowed for analysis of whether the observed chimerism was caused by bioinformatic errors in assembly.

The markers allowed for the disambiguation of the 11 chimeras: the chimerism of scaffolds 13, 18, 22, 28, 41, 60, 75 and 93 was likely to be bioinformatic and in scaffolds 1 and 7 was likely to be biologically valid. Scaffold 9 contained two chimeras, one of which appeared to be biological and one of which appeared to bioinformatic. The chimeras for which informative markers were not available might still represent small, previously-unknown rearrangements; the development of a new set of markers would be one way to evaluate whether this is indeed the case.

SCAFFOLD	DOG CHR. COVERED	BIOLOGICALLY PREDICTED?	INFORMATIVE MARKERS?	VERDICT
1	28, 30	Yes	Yes	Biological
5	2, 6			
7	12, 33	Yes	Yes	Biological
9	18, 38, 20	Yes	Yes (all)	Biological (18&38), Bioinformatic (20)
12	32,3			
13	14, 21, 5		Yes (all)	Bioinformatic
18	11, 13		Yes	Bioinformatic
21	9, X			
22	6, 8		Yes	Bioinformatic
28	3, 32		Yes	Bioinformatic
29	30, 24			
35	2, 23			
41	24, 1		Yes	Bioinformatic
57	18, 23			
60	7, X		Yes	Bioinformatic
75	11, 30		Yes	Bioinformatic
93	2, 5, 8		Yes (all)	Bioinformatic
100	4, 36			
101	2, 23			
148	15, 1			
174	19, 35			
195	10, 1			
239	35, 5	Yes		
255	18, 27			

Table 9: Chimeric scaffolds were evaluated to determine whether microsatellite markers could be placed in the segments that mapped to different dog chromosomes. Where this was the case, it was possible to determine whether the chimera represented a biological departure in dog-fox synteny, or whether the scaffold was chimeric because it had been misassembled.

Conclusion

The primary purpose of this study was to assemble the scaffolds of the draft fox genome into the fox chromosomes for use in genetic mapping studies. The alignment of the fox genome against the dog genome and subsequent analysis with the Kent utilities for chaining and netting and with RACA for identification of syntenic blocks has produced the results needed for chromosomal assembly. The syntenic blocks were assembled into the fox chromosomes based on their position in the dog genome and on previously identified fox-dog synteny. Comparison of this map to the previously established pattern of cat-dog synteny supported the alignment and also allowed for the refinement of the assembly in locations where the dog has diverged from the other carnivores, and for the development of a fox-cat syntenic map.

Over 98% of the dog genome by length was placed by the SFs, suggesting that the chromosome assemblies are likely to contain a large percentage of the fox chromosome sequence as well.

The assembly offers refinements both to the preexisting gaps between syntenic blocks in the dog-fox linkage map and to the draft genome assembly by highlighting nine artificially chimeric scaffolds. Because SFs were mapped into the regions currently unassigned to a syntenic block in the dog, synteny with the cat could be used to determine which syntenic block the SFs were most likely to belong to. As for the chimeras, mapping of previously-assigned meiotic linkage markers to the scaffolds indicated that pieces of the scaffolds belong to different linkage groups, suggesting that they had been erroneously assembled. Thus, the fox chromosome assembly has provided information that can be used to mitigate some of the assembly problems in the draft fox genome.

This study is expected to produce a number of deliverables. First, the fox-dog pairwise alignment (based on the chain and net) can be uploaded to UCSC Genome Browser with the release of the draft fox genome to facilitate use of the new genome by fox researchers globally who currently rely on the dog genome. Second, no pairwise alignment has yet been released for canFam3/felCat5, though they were previously available for canFam2/felCat3, so one by-product of the current study was to produce an updated tool for the cat and dog genomes. Third, the alignment of the cat and fox chromosomes depicted in Figure 12 is, to the author's knowledge, the first estimated cat-fox syntenic map. Finally, once the red fox chromosome assembly has been transformed from a list of syntenic blocks to the corresponding .fasta files, versions of these files will be prepared for upload to UCSC Genome Browser as a new track; they are anticipated to comprise the second version of the red fox genome following the release of the draft fox genome. The chromosomal assembly of the fox genome is anticipated to provide a valuable reference for genetic and genomic studies using the red fox as a model. In summary, this study has produced a number of tools to facilitate the analysis of vulpine genomics and to integrate the new fox genome with the analysis of related species.

REFERENCES

Adams, J. U. (2008). DNA sequencing technologies. *Nature Education*, *1*, 193.

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., … Venter, J. C. (2000). The genome sequence of Drosophila melanogaster. *Science*, *287*, 2185–2195.

Alkan, C., Sajjadian, S., & Eichler, E. E. (2011). Limitations of next-generation genome sequence assembly. *Nature Methods*, *8*, 61–65.

Becker, S. E. D., Thomas, R., Trifonov, V. A., Wayne, R. K., Graphodatsky, A. S., & Breen, M. (2011). Anchoring the dog to its relatives reveals new evolutionary breakpoints across 11 species of the Canidae and provides new clues for the role of B chromosomes. *Chromosome Research*, *19*, 685–708.

Belyaev, D. K. (1979). Destabilizing selection as a factor in domestication. *Journal of Heredity*, *70*, 301–308.

Belyaev, D. K., Plyusnina, I. Z., & Trut, L. N. (1985). Domestication in the silver fox (Vulpes fulvus Desm): Changes in physiological boundaries of the sensitive period of primary socialization. *Applied Animal Behaviour Science*, *13*, 359–370.

Bentolila, S., Bach, J. M., Kessler, J. L., Bordelais, I., Cruaud, C., Weissenbach, J., & Panthier, J. J. (1999). Analysis of major repetitive DNA sequences in the dog (Canis familiaris) genome. *Mammalian Genome*, *10*, 699–705.

Blattner, F. R. (1997). The complete genome sequence of Escherichia coli K-12. *Science*, *277*, 1453–1462.

Broad Institute. (2014). Picard. Retrieved from http://broadinstitute.github.io/picard/

The C. elegans Sequencing Consortium. (1998). Genome sequence of the nematode C. elegans: A platform for investigating biology. *Science*, *282*, 2012–2018.

Carl R. Woese Institute for Genomic Biology at the University of Illinois at Urbana-Champaign. (2015). Biocluster. Retrieved May 27, 2015, from http://help.igb.illinois.edu/Biocluster

Clutton-Brock, J. (1992). The process of domestication. *Mammal Review*, *22*, 79–85.

Compeau, P. E. C., Pevzner, P. A., & Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, *29*, 987–991.

Davis, B. W., Raudsepp, T., Pearks Wilkerson, A. J., Agarwala, R., Schäffer, A. A., Houck, M., … Murphy, W. J. (2009). A high-resolution cat radiation hybrid and integrated FISH mapping resource for phylogenomic studies across Felidae. *Genomics*, *93*, 299–304.

Dolezel, J., Bartos, J., Voglmayr, H., & Greilhuber, J. (2003). Nuclear DNA content and genome size of trout and human. *Cytometry. Part A : The Journal of the International Society for Analytical Cytology*, *51*, 127–128.

Eichler, E. E. (2001). Segmental duplications: What's missing, misassigned, and misassembled — and should we care? *Genome Research*, *11*, 653–656.

Flicek, P., & Birney, E. (2009). Sense from sequence reads: methods for alignment and assembly. *Nature Methods*, *6*, S6–S12.

Frith, M. C., Hamada, M., & Horton, P. (2010). Parameters for accurate genome alignment. *BMC Bioinformatics*, *11*, 80.

Ge, R.-L., Cai, Q., Shen, Y.-Y., San, A., Ma, L., Zhang, Y., … Wang, J. (2013). Draft genome sequence of the Tibetan antelope. *Nature Communications*, *4*, 1858.

Genome 10K. (2009). Species List. Retrieved June 30, 2015, from https://genome10k.soe.ucsc.edu/species

Genome 10K Consortium of Scientists. (2009). Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *The Journal of Heredity*, *100*, 659–674.

Genome Reference Consortium. (2013). Human Genome Assembly Data. Retrieved from http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/data/

Gilbert, D. G. (1999). Polydendron. Retrieved from http://iubio.bio.indiana.edu/treeapp/treeprint-form.html

Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., … Jaffe, D. B. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 1513–1518.

Graphodatsky, A. S., Perelman, P. L., Sokolovskaya, N. V, Beklemisheva, V. R., Serdukova, N. A., Dobigny, G., … Yang, F. (2008). Phylogenomics of the dog and fox family (Canidae, Carnivora) revealed by chromosome painting. *Chromosome Research*, *16*, 129–143.

Graphodatsky, A. S., Yang, F., O'Brien, P. C., Serdukova, N. A., Milne, B., Trifonov, V., & Ferguson-Smith, M. A. (2000). A comparative chromosome map of the Arctic fox, red fox and dog defined by chromosome painting and high resolution G-banding. *Chromosome Research*, *8*, 253–263.

Gregory, T. R. (2015). Animal Genome Size Database. Retrieved from http://www.genomesize.com/

Guttmacher, A., & Collins, F. (2003). Welcome to the genomic era. *New England Journal of Medicine*, *394*, 996–998.

Hare, B., Wobber, V., & Wrangham, R. (2012). The self-domestication hypothesis: evolution of bonobo psychology is due to selection against aggression. *Animal Behaviour*, *83*, 573–585.

Harris, R. S. (2007). *Improved pairwise alignment of genomic DNA*. The Pennsylvania State University.

Hayden, E. C. (2014, January). Is the $1,000 genome for real? *Nature News & Comment*. Retrieved from http://www.nature.com/news/is-the-1-000-genome-for-real-1.14530

Henson, J., Tischler, G., & Ning, Z. (2012). Next-generation sequencing and large genome assemblies. *Pharmacogenomics*, *13*, 901–915.

Hubisz, M. J., Pollard, K. S., & Siepel, A. (2011). PHAST and RPHAST: Phylogenetic analysis with space/time models. *Briefings in Bioinformatics*, *12*, 41–51.

Illumina Inc. (2010). *Genomic Sequencing*. Retrieved from http://www.illumina.com/Documents/products/datasheets/datasheet_genomic_sequence.pdf

Illumina Inc. (2011). *Paired-End Sample Preparation Guide*. Retrieved from http://supportres.illumina.com/documents/myillumina/e5af4eb5-6742-40c8-bcb1-d8b350bcb964/paired-end_sampleprep_guide_1005063_e.pdf

Illumina Inc. (2012). *Data Processing of Nextera® Mate Pair Reads on Illumina Sequencing Platforms*. Retrieved from http://res.illumina.com/documents/products/datasheets/datasheet_nextera_mate_pair.pdf

Illumina Inc. (2015). *Mate Pair Sequencing*. Retrieved from http://www.illumina.com/technology/next-generation-sequencing/mate-pair-sequencing_assay.html

Johnson, J. L., Kozysa, A., Kharlamova, A., Gulevich, R. G., Perelman, P. L., Fong, H. W. F., … Kukekova, A. V. (2015). Platinum coat color in red fox (Vulpes vulpes) is caused by a mutation in an autosomal copy of KIT. *Animal Genetics*, *46*, 190–199.

Johnson, J. L., Wittgenstein, H., Mitchell, S. E., Hyma, K. E., Temnykh, S. V, Kharlamova, A. V., … Kukekova, A. V. (2015). Genotyping-by-sequencing (GBS) detects genetic structure and confirms behavioral QTL in tame and aggressive foxes (Vulpes vulpes). *PLoS ONE*, *10*, e0127013.

Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W., & Haussler, D. (2003). Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 11484–11489.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The Human Genome Browser at UCSC. *Genome Research*, *12*, 996–1006.

Kim, J., Larkin, D. M., Cai, Q., Asan, Zhang, Y., Ge, R.-L., … Ma, J. (2013). Reference-assisted chromosome assembly. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 1785–1790.

Kukekova, A. V., Johnson, J. L., Teiling, C., Li, L., Oskina, I. N., Kharlamova, A. V., … Acland, G. M. (2011). Sequence comparison of prefrontal cortical brain transcriptome from a tame and an aggressive silver fox (Vulpes vulpes). *BMC Genomics*, *12*, 482.

Kukekova, A. V., Temnykh, S. V, Johnson, J. L., Trut, L. N., & Acland, G. M. (2012). Genetics of behavior in the silver fox. *Mammalian Genome*, *23*, 164–177.

Kukekova, A. V., Trut, L. N., & Acland, G. M. (2014). Genetics of Domesticated Behavior in Dogs and Foxes. In *Genetics and the Behavior of Domestic Animals* (pp. 361–396).

Kukekova, A. V., Trut, L. N., Chase, K., Kharlamova, A. V., Johnson, J. L., Temnykh, S. V., … Lark, K. G. (2011). Mapping loci for fox domestication: deconstration/reconstruction of a behavioral phenotype. *Behavioral Genetics*, *41*, 593–606.

Kukekova, A. V., Trut, L. N., Oskina, I. N., Johnson, J. L., Temnykh, S. V, Kharlamova, A. V., … Acland, G. M. (2007). A meiotic linkage map of the silver fox, aligned and compared to the canine genome. *Genome Research*, *17*, 387–99.

Kukekova, A. V., Trut, L. N., Oskina, I. N., Kharlamova, A. V., Shikhevich, S. G., Kirkness, E. F., … Acland, G. M. (2004). A marker set for construction of a genetic map of the silver fox (Vulpes vulpes). *Journal of Heredity*, *95*, 185–194.

Lander, E. S., Linton, L. M., Birren, B. W., Nusbaum, C., Zody, M. C., Baldwin, J., … Chen, Y. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*, 860–921.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*, 357–359.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*, 1754–1760.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., … Wang, J. (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, *20*, 265–272.

Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., Karlsson, E. K., Jaffe, D. B., Kamal, M., … Lander, E. S. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, *438*, 803–819.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., … Wang, J. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, *1*, 18.

Ma, J., Zhang, L., Suh, B. B., Raney, B. J., Burhans, R. C., Kent, W. J., … Miller, W. (2006). Reconstructing contiguous regions of an ancestral genome. *Genome Research*, *16*, 1557–1565.

Mardis, E., McPherson, J. D., Martienssen, R., Wilson, R. K., & McCombie, W. R. (2002). What is finished, and why does it matter? *Genome Research*, *12*, 669–671.

Menotti-Raymond, M., David, V. A., Schäffer, A. A., Tomlin, J. F., Eizirik, E., Phillip, C., … O'Brien, S. J. (2009). An autosomal genetic linkage map of the domestic cat, Felis silvestris catus. *Genomics*, *93*, 305–313.

Montague, M. J., Li, G., Gandolfi, B., Khan, R., Aken, B. L., Searle, S. M. J., … Warren, W. C. (2014). Comparative analysis of the domestic cat genome reveals genetic signatures underlying feline biology and domestication, *111*, 17230–17235.

Morey, D. F. (1994). Animal Evolution of the Domestic Dog can also be described as an evolutionary process. *American Scientist*, *82*, 336–347.

Murphy, W. J., Davis, B., David, V. A., Agarwala, R., Schäffer, A. A., Pearks Wilkerson, A. J., … Menotti-Raymond, M. (2007). A 1.5-Mb-resolution radiation hybrid map of the cat genome and comparative analysis with the canine and human genomes. *Genomics*, *89*, 189–196.

Murphy, W. J., Larkin, D. M., Everts-van der Wind, A., Bourque, G., Tesler, G., Auvil, L., … Lewin, H. A. (2005). Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science*, *309*, 613–617.

Murphy, W. J., Stanyon, R., & O'Brien, S. (2001). Evolution of mammalian genome organization inferred from comparative gene mapping. *Genome Biology*, *2*, 1–8.

Nash, W., Menninger, J., Wienberg, J., Padilla-Nash, H., & O'Brien, S. (2001). The pattern of phylogenomic evolution of the Canidae. *Cytogenetics and Cell Genetics*, *95*, 210–224.

Pevzner, P. A., Tang, H., & Waterman, M. S. (2001). An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America*, *98*, 9748–9753.

Pontius, J. U., Mullikin, J. C., Smith, D. R., Agencourt, S. T., Lindblad-Toh, K., Gnerre, S., … O'Brien, S. J. (2007). Initial sequence and comparative analysis of the cat genome. *Genome Research*, *17*, 1675–89.

Ribeiro, F. J., Przybylski, D., Yin, S., Sharpe, T., Gnerre, S., Abouelleil, A., … Jaffe, D. B. (2012). Finished bacterial genomes from shotgun sequence data. *Genome Research*, *22*, 2270–2277.

Rosenbloom, K. R., Armstrong, J., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., … Kent, W. J. (2014). The UCSC Genome Browser database: 2015 update. *Nucleic Acids Research*, *43*, D670–D681.

Salzberg, S. L., & Yorke, J. A. (2005). Beware of mis-assembled genomes. *Bioinformatics*, *21*, 4320–4321.

Sanger, F., Air, G., Barrell, B., Brown, N., Coulson, A., Fiddes, C., … Smith, M. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, *265*, 687–695.

Smit, A., Hubley, R., & Green, P. (2013). RepeatMasker Open-4.0. Retrieved from <http://www.repeatmasker.org>

Spady, T. C., & Ostrander, E. A. (2007). Canid genomics: Mapping genes for behavior in the silver fox. *Genome Research*, *17*, 259–263.

Statham, M. J., Trut, L. N., Sacks, B. N., Kharlamova, A. V., Oskina, I. N., Gulevich, R. G., … Kukekova, A. V. (2011). On the origin of a domesticated species: Identifying the parent population of Russian silver foxes (Vulpes vulpes). *Biological Journal of the Linnean Society*, *103*, 168–175.

Tajima, F. (1992). Statistical method for estimating the standard errors of branch lengths in a phylogenetic tree reconstructed without assuming equal rates of nucleotide substitution among different lineages. *Molecular Biology and Evolution*, *9*, 168–181.

Tavare, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, *17*, 57–86.

Trut, L. N. (1980). The Genetics and Phenogenetics of Domestic Behaviour. In D. K. Belyaev (Ed.), *Problems in General Genetics: Proceedings of the XIV International Congress of Genetics, Vol. II* (pp. 123–137). Moscow.

Trut, L. N. (1999). Early canid domestication: The farm-fox experiment. *American Scientist*, *87*, 160–169.

Trut, L., Oskina, I., & Kharlamova, A. V. (2009). Animal evolution during domestication: the domesticated fox as a model. *Bioessays*, *31*, 349–360.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., … Zhu, X. (2001). The sequence of the human genome. *Science*, *291*, 1304–1351.

Wagman, B. (2010, November 30). Genome 10K project announces first 101 species for genome sequencing. *CBSE in the News*. Santa Cruz, CA. Retrieved from https://cbse.soe.ucsc.edu/news/article/1820?ID=1820

Wang, W., & Kirkness, E. F. (2005). Short interspersed elements (SINEs) are a major source of canine genomic diversity. *Genome Research*, *15*, 1798–1808.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., … Lander, E. S. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, *420*, 520–562.

Wayne, R. K. (1993). Molecular evolution of the dog family. *Trends in Genetics*, *9*, 218–224.

Wetterstrand, K. (2014). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Retrieved June 10, 2015, from www.genome.gov/sequencingcosts

Wilkins, A. S., Wrangham, R. W., & Fitch, W. T. (2014). The "domestication syndrome" in mammals: A unified explanation based on neural crest cell behavior and genetics. *Genetics*, *197*, 795–808.

Wurster-Hill, D. H., Ward, O. G., Davis, B. H., Park, J. P., Moysiz, R. K., & Mayne, J. (1988). Fragile sites, telomeric DNA sequences, B chromosomes, and DNA content in raccoon dogs, Nyctereutes procyonoides, with comparative notes on foxes, coyote, wolf, and raccoon. *Cytogenetics and Cell Genetics*, *49*, 278–281.

Yang, F., Graphodatsky, A. S., O'Brien, P. C., Colabella, A., Solanky, N., Squire, M., … Ferguson-Smith, M. A. (2000). Reciprocal chromosome painting illuminates the history of genome evolution of the domestic cat, dog and human. *Chromosome Research*, *8*, 393–404.

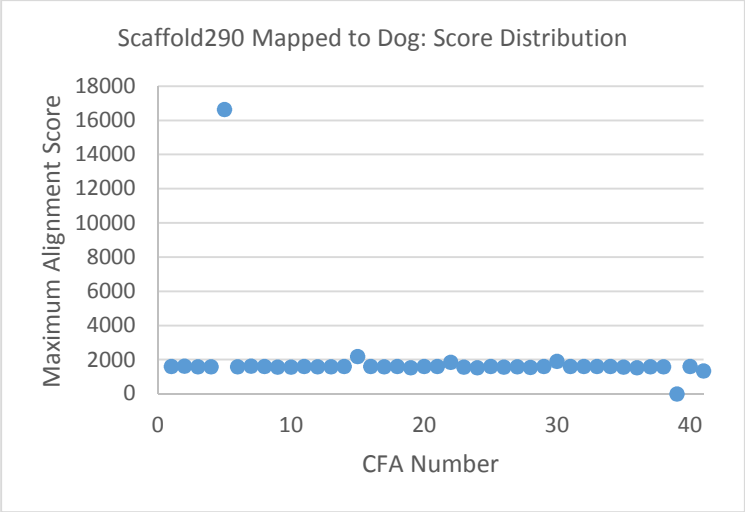
Yang, F., O'Brien, P. C., Milne, B., Graphodatsky, A. S., Solanky, N., Trifonov, V., … Ferguson-Smith, M. A. (1999). A complete comparative chromosome map for the dog, red fox, and human and its integration with canine genetic maps. *Genomics*, *62*, 189–202.

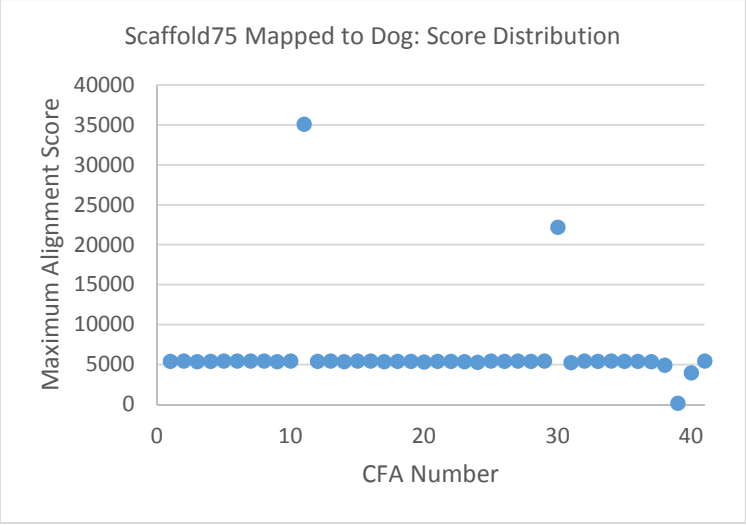
Yang, Z. (1994). Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution*, *39*, 105–111.

Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*, 821–829.

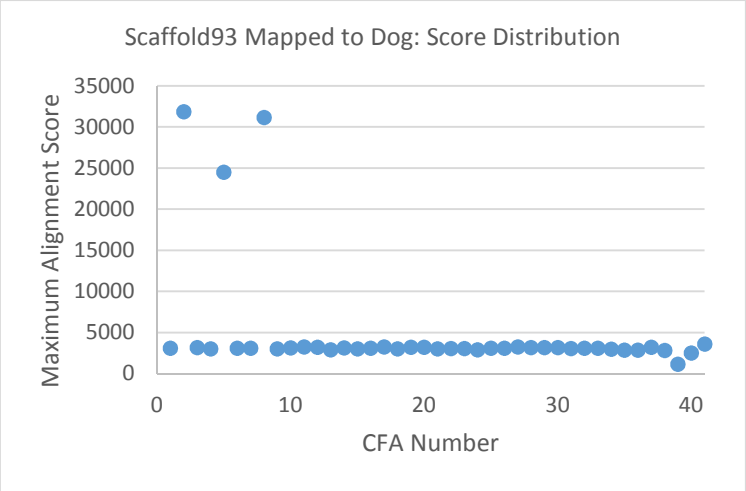
APPENDIX A

The 500 largest scaffolds in the fox genome were mapped against the dog genome to determine the orthologous region in the dog. Below, each graph represents a scaffold, and each dot represents the maximum score obtained for alignment of that scaffold against a particular dog chromosome. Chromosomes without ordinal identifiers are indicated as follows: M = CFA 39, X = CFA 40, Y = CFA 41. Scores of 0 correspond to no mapping between the scaffold and chromosome and were dropped from the distribution before z-score analysis (discussed in Chapter II). In Scaffold290, the score distribution indicates synteny with a single dog chromosome, whereas Scaffold75 maps to two chromosomes and Scaffold93 to three. Though these distributions are not normal, in all cases certain chromosomes map so much more significantly to the scaffold that the statistic is still able to differentiate them from the background noise.



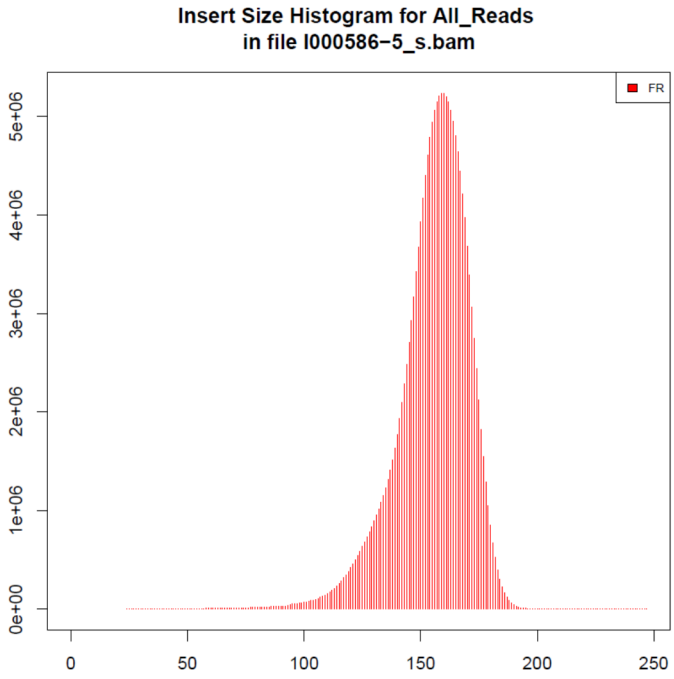


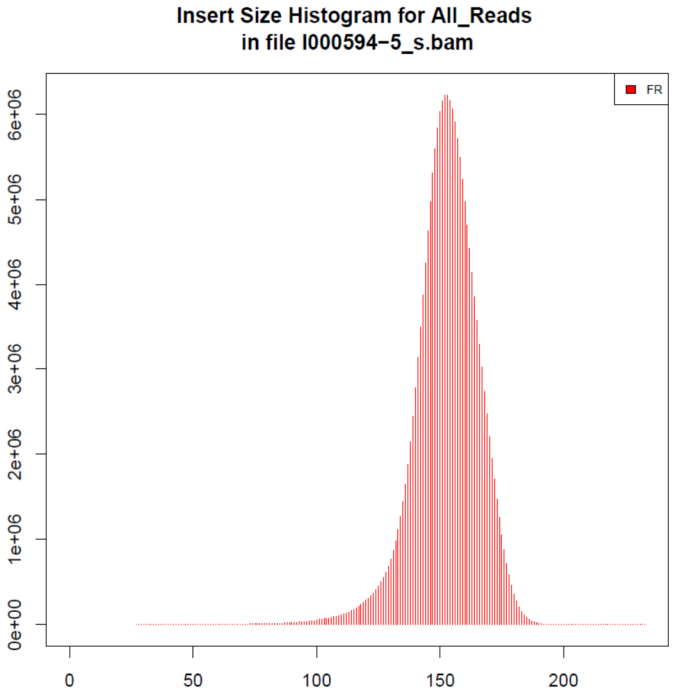
APPENDIX A (cont.)



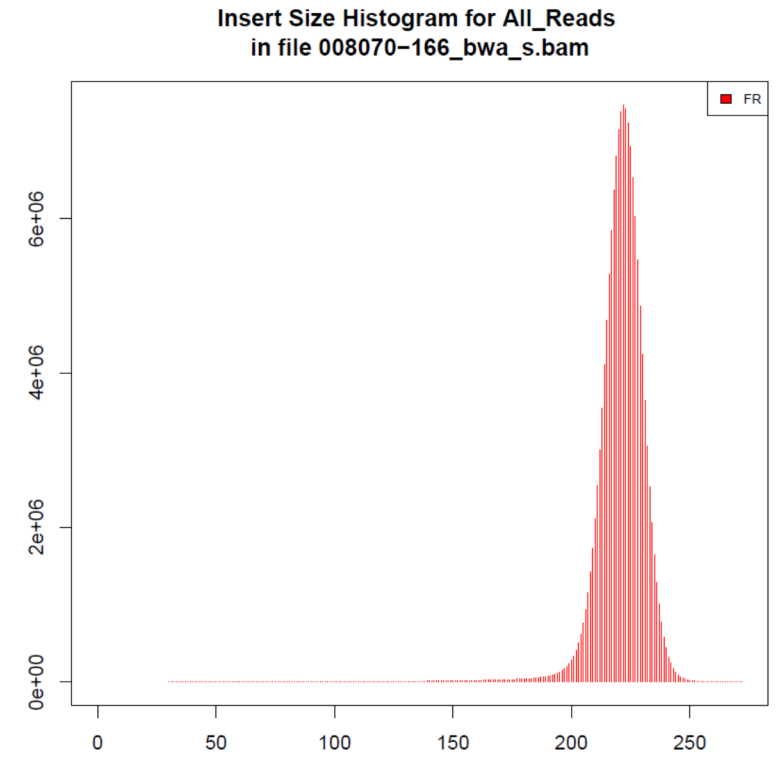
APPENDIX B

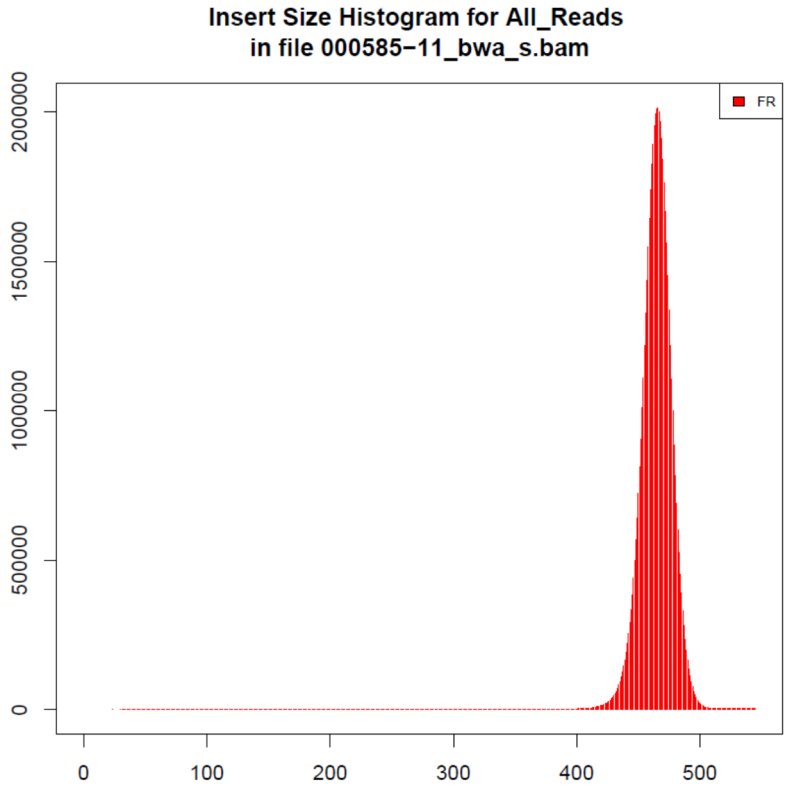
Library-by-library insert size distributions, as generated by Picard. Expected insert sizes can be found in Table 2 for comparison. FR libraries are in red, RF libraries are in blue.



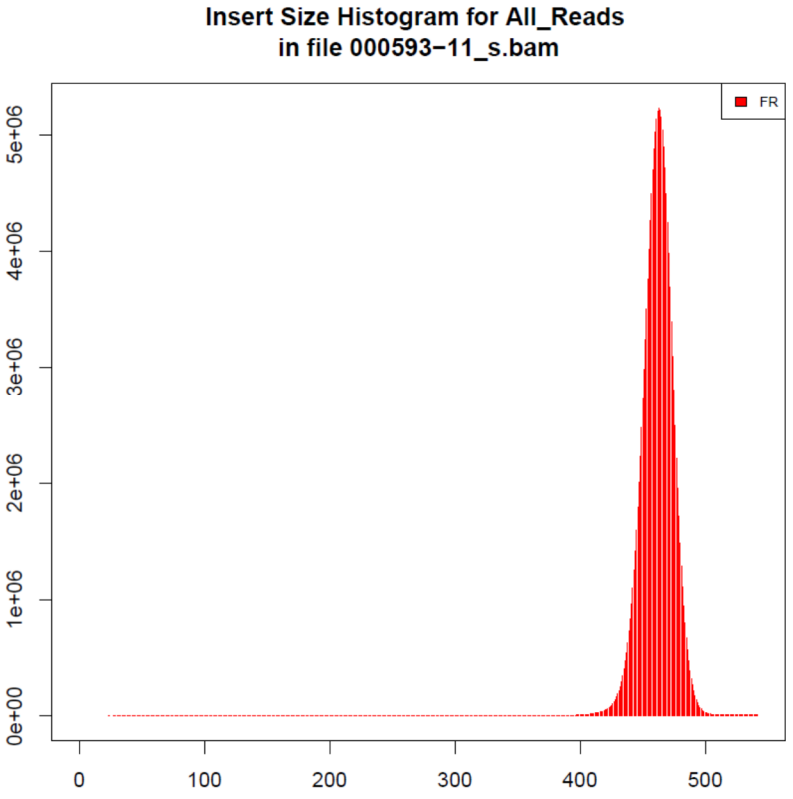


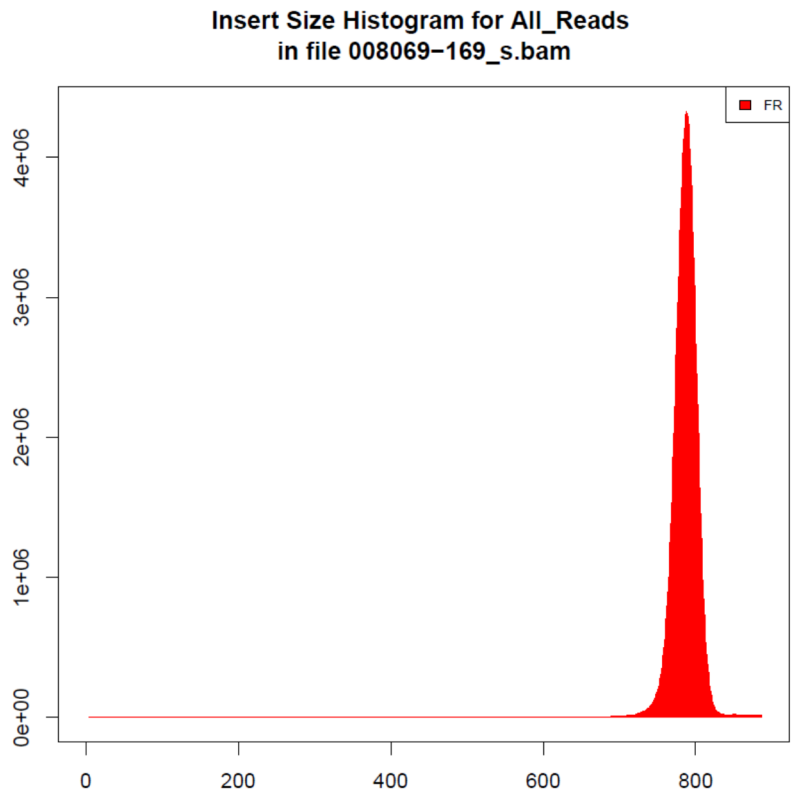
APPENDIX B (cont.)



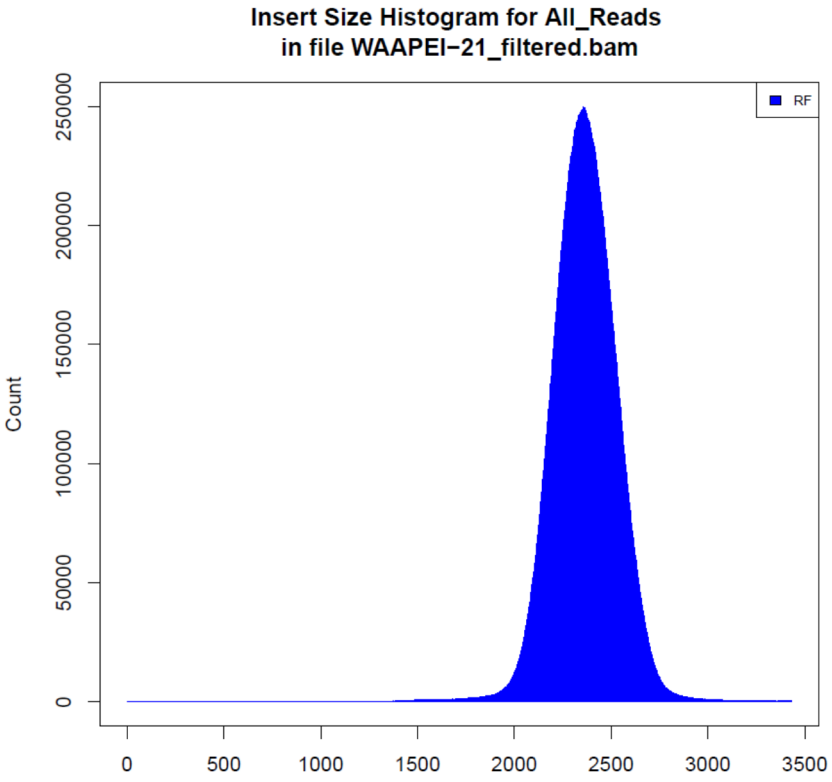


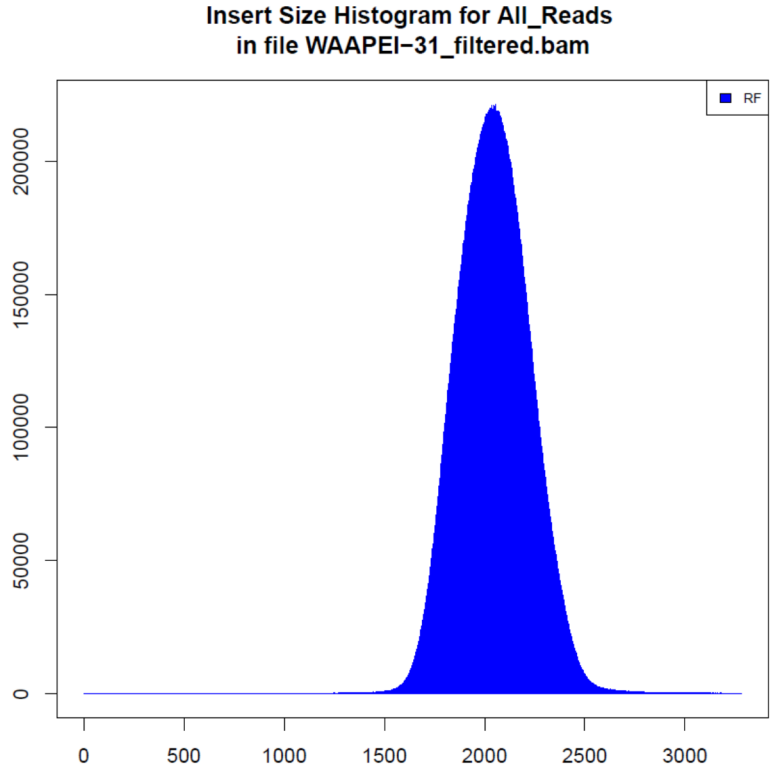
APPENDIX B (cont.)



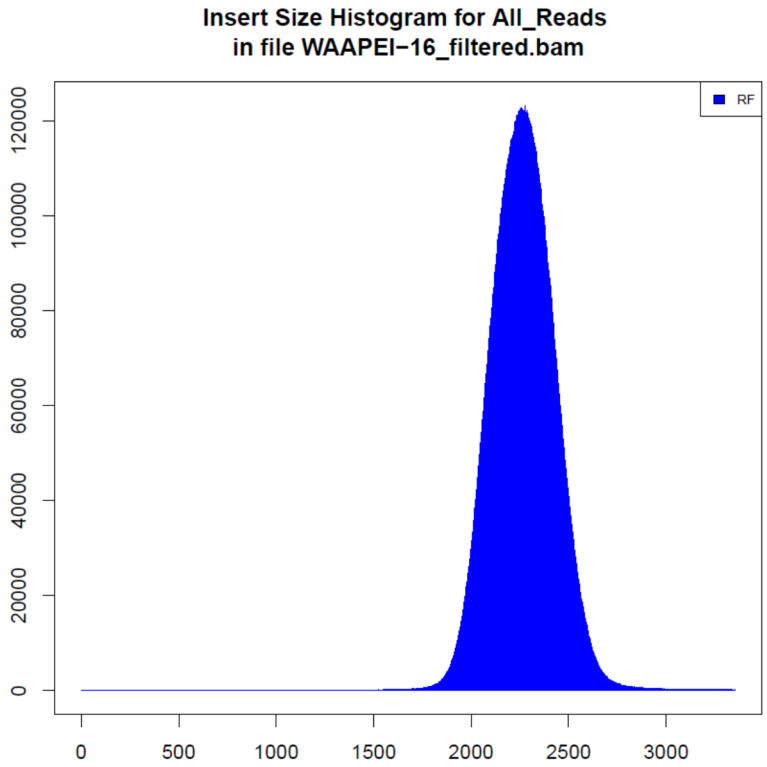


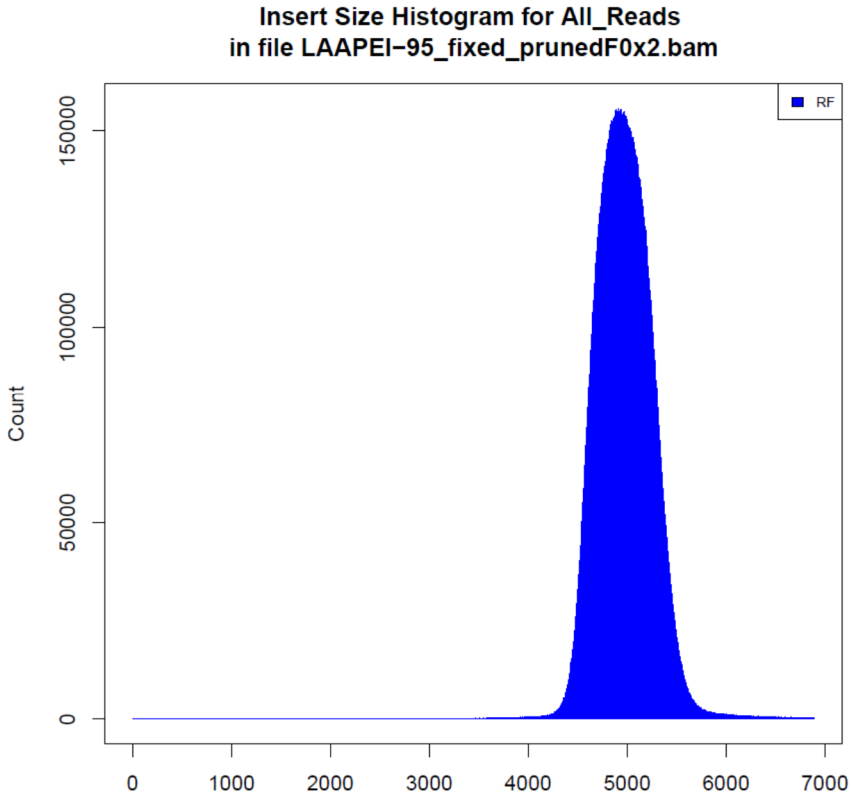
APPENDIX B (cont.)



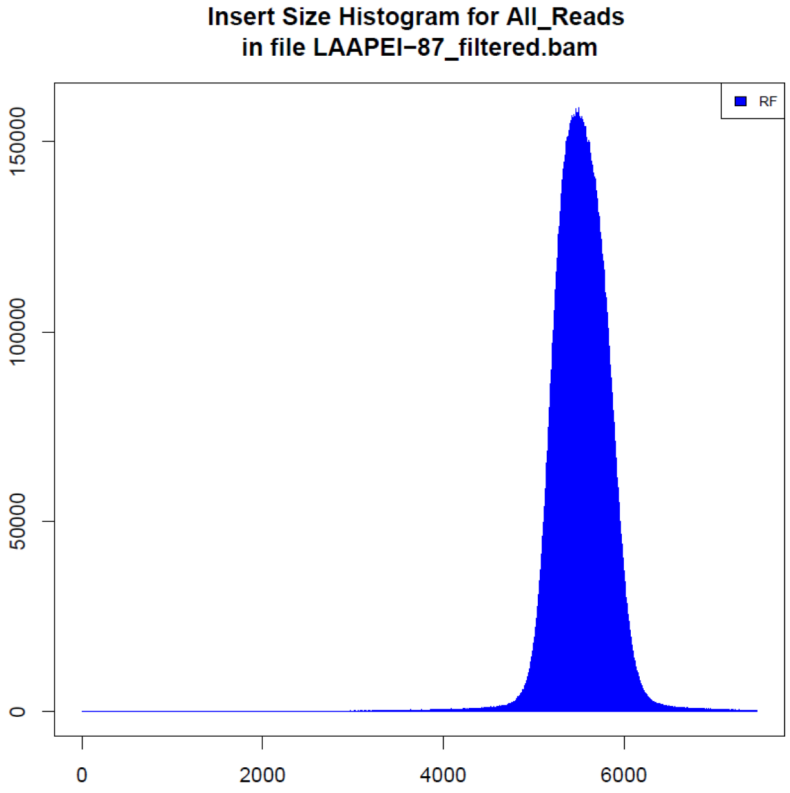


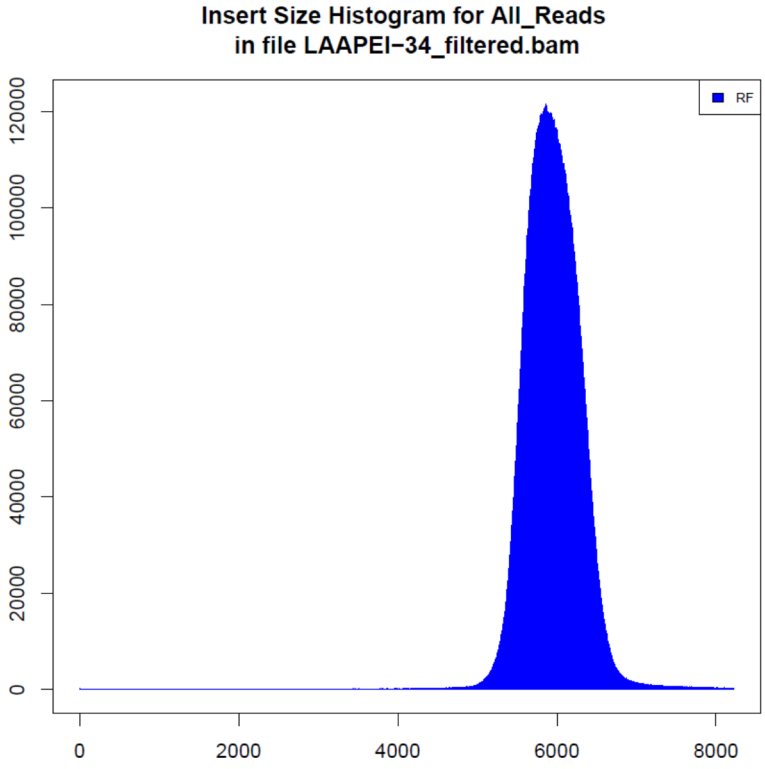
APPENDIX B (cont.)



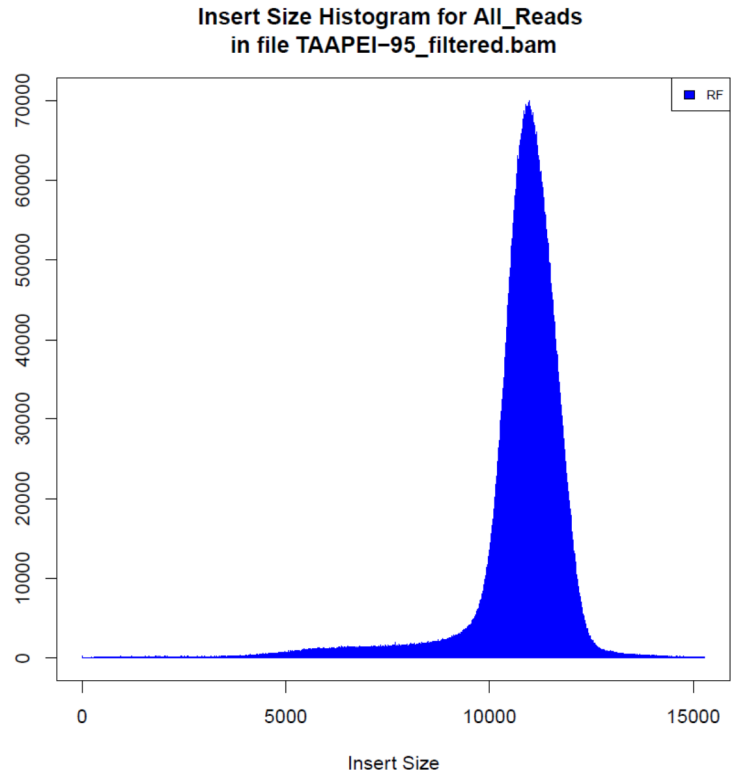


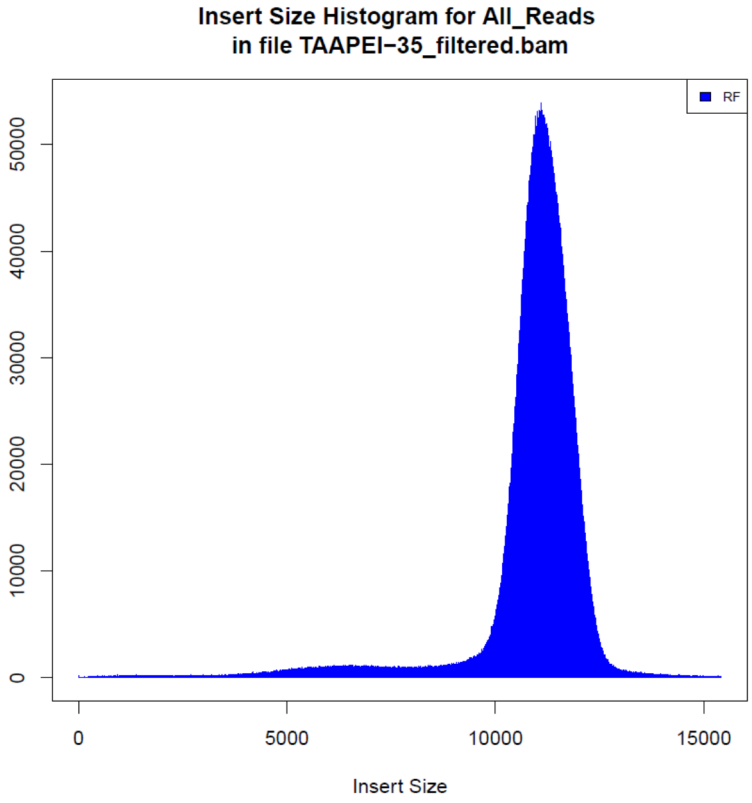
APPENDIX B (cont.)



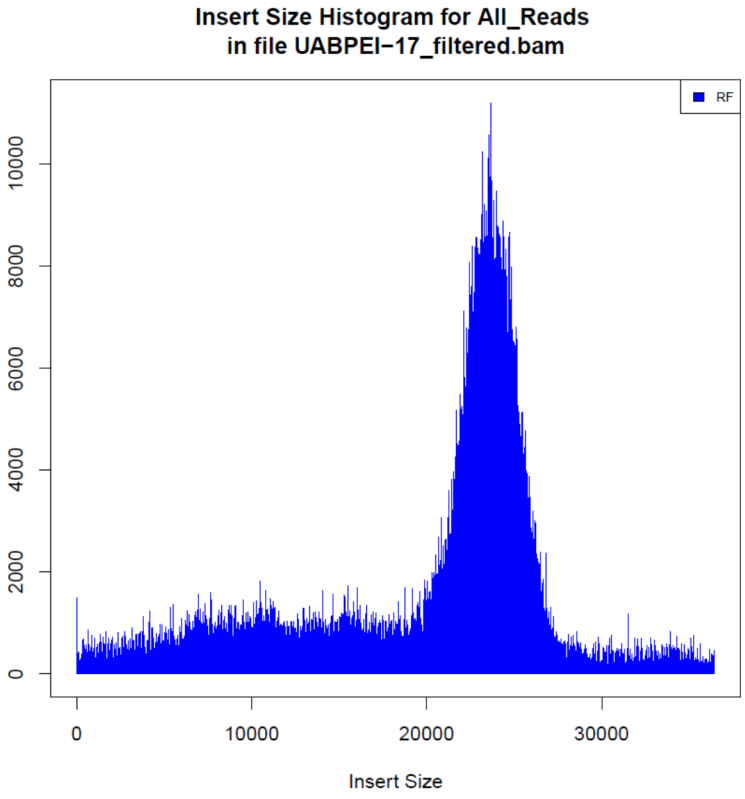


APPENDIX B (cont.)





APPENDIX B (cont.)



APPENDIX C

Z-scores estimated in the analysis from Chapter II by which chimeras were identified. Each scaffold was analyzed individually, and the z-score was estimated based on the maximum alignment score for each chromosome compared to the maximum alignment score for all other chromosomes.

Scaffold	Number of SFs	Dog Chromosome	Z-Score
1	2	chr28	3.843059
		chr30	4.866379
2	1	chr3	6.239366
3	1	chr4	6.202728
4	1	chr29	6.299715
5	1	chr2	6.139409
6	1	chr20	6.279371
7	2	chr33	5.381958
		chr12	3.083004
8	1	chr12	6.255059
9	3	chr18	3.440666
		chr38	3.374353
		chr20	3.810001
10	1	chr15	6.25924
11	1	chr10	6.274766
12	2	chr32	1.942638
		chr3	5.919194
13	3	chr14	5.073155
		chr21	2.404547
		chr5	2.56943
14	1	chr13	6.28015
15	1	chr16	6.274541
16	1	chr36	6.222104
17	1	chr17	6.237469
18	2	chr13	2.713183
		chr11	5.549493
19	1	chr27	6.235581
20	1	chr31	6.242674
21	2	chrX	1.71426
		chr9	5.953597
22	2	chr6	4.669528
		chr8	3.961067
23	1	chr25	6.222003
24	1	chr7	6.300224

APPENDIX C (cont.)

25	1	chrX	6.27793
26	1	chr1	6.239893
27	1	chr1	6.237079
28	2	chr32	4.800156
		chr3	3.882589
29	2	chr24	3.464685
		chr30	5.196447
30	1	chr5	6.242228
31	1	chr35	6.201027
32	1	chr9	6.243766
33	1	chr27	6.278959
34	1	chr8	6.192043
35	2	chr23	4.566402
		chr2	4.159232
36	1	chr7	6.290854
37	1	chr11	6.256604
38	1	chr1	6.237562
39	1	chr10	6.223381
40	1	chr37	6.227959
41	2	chr24	3.662532
		chr1	5.031662
42	1	chr37	6.266531
43	1	chr11	6.269544
44	2	chr26	4.274866
		chr12	4.423673
45	1	chr6	6.320156
46	1	chr7	6.290728
47	1	chr13	6.205645
48	1	chr22	6.229497
49	1	chr14	6.235257
50	1	chr18	6.24255
51	1	chr6	6.291251
52	1	chr5	6.243653
53	1	chr5	6.278143
54	1	chr4	6.212639
55	1	chr12	6.287769
56	1	chr7	6.229787
57	3	chr18	4.616767
		chr23	3.630935
		chr9	1.723589

APPENDIX C (cont.)

58	1	chr25	6.239471
59	1	chr23	6.111406
60	2	chrX	2.762326
		chr7	5.527326
61	1	chr18	6.243225
62	1	chr5	6.170434
63	1	chr5	6.241687
64	1	chr34	6.223283
65	1	chr12	6.236631
66	1	chr8	6.243436
67	1	chr34	6.244093
68	1	chr25	6.21394
69	1	chr19	6.197273
70	1	chr15	6.13373
71	1	chr21	6.244538
72	1	chr8	6.229415
73	1	chr4	6.296738
74	1	chr22	6.242684
75	2	chr11	5.411449
		chr30	2.982014
76	1	chrX	6.252709
77	1	chr17	6.242795
78	1	chr9	6.238238
79	1	chr22	6.241498
80	1	chr24	6.30313
81	1	chr1	6.242289
82	1	chr21	6.168771
83	1	chr28	6.221597
84	1	chr14	6.243197
85	1	chr10	6.237812
86	1	chr6	6.238301
87	1	chr5	6.256936
88	1	chr2	6.242491
89	1	chr2	6.241015
90	1	chr14	6.287714
91	1	chrX	6.242873
92	1	chr24	6.243827
93	3	chr5	2.845548
		chr2	3.915152
		chr8	3.812378

APPENDIX C (cont.)

94	1	chr8	6.236069
95	1	chr18	6.235492
96	1	chr15	6.227364
97	1	chr11	6.20881
98	1	chr1	6.237815
99	1	chr32	6.247655
100	2	chr36	2.314716
		chr4	5.724101
101	2	chr23	4.436519
		chr2	4.28937
102	1	chr1	6.180256
103	1	chr19	6.242897
104	1	chr2	6.242881
105	1	chr24	6.306676
106	1	chr26	6.177945
107	1	chrX	6.225442
108	1	chr17	6.142627
109	1	chr24	6.207376
110	1	chr21	6.284546
111	1	chrX	6.157319
112	1	chr11	6.231215
113	1	chr23	6.233873
114	1	chr19	6.216196
115	1	chr32	6.183612
116	1	chr19	6.229766
117	1	chr23	6.24369
118	1	chr16	6.257471
119	2	chr26	4.059638
		chr24	4.639109
120	1	chr1	6.234821
121	1	chr17	6.17866
122	1	chr9	6.099407
123	1	chr23	6.241004
124	1	chr34	6.203069
125	1	chr13	6.119622
126	1	chr15	6.172099
127	1	chr22	6.220927
128	1	chr4	6.23891
129	1	chrX	5.593785
130	1	chr19	6.199989

APPENDIX C (cont.)

131	1	chr20	6.236613
132	1	chr4	6.24273
133	1	chr26	6.19592
134	1	chr22	6.240746
135	1	chr10	6.234624
136	1	chr6	6.269378
137	1	chr25	6.243159
138	1	chr32	6.142453
139	1	chr34	6.244093
140	1	chr22	6.243971
141	1	chr19	6.216044
142	1	chr10	6.316647
143	1	chrX	6.236773
144	1	chr16	6.186411
145	1	chr7	6.242313
146	1	chr6	6.240564
147	1	chrX	6.229206
148	2	chr15	5.795614
		chr1	2.177689
149	1	chr21	6.244543
150	1	chr16	6.212964
151	1	chr16	6.298772
152	1	chr36	6.242104
153	1	chr8	6.180801
154	1	chr19	6.219847
155	1	chr23	6.244098
156	1	chr21	6.244882
157	1	chr14	6.235029
158	1	chr1	6.244872
159	1	chr26	6.217144
160	1	chr17	6.23514
161	1	chr8	6.22598
162	1	chr14	6.233143
163	1	chr34	6.320269
164	1	chr20	6.151134
165	1	chr6	6.243742
166	1	chr10	6.060747
167	2	chr26	2.985925
		chr9	5.406319
168	1	chr16	6.272617

APPENDIX C (cont.)

169	1	chr3	6.16463
170	1	chr7	6.243105
171	1	chr22	6.241694
172	1	chrX	6.191241
173	1	chr8	6.24395
174	2	chr35	5.906844
		chr19	1.872934
175	1	chr26	6.243441
176	1	chrX	6.221773
177	1	chr17	6.184109
178	1	chr8	6.233782
179	1	chrX	6.187619
180	1	chr5	6.214943
181	1	chr17	6.203133
182	1	chr31	6.223858
183	1	chr20	6.233249
184	1	chr15	6.227885
185	1	chr6	6.234554
186	1	chr5	6.24386
187	1	chr32	6.218608
188	1	chr12	6.240073
189	1	chrX	6.2386
190	1	chr17	6.210456
191	1	chr18	6.235427
192	1	chr4	6.243597
193	1	chr1	6.241398
194	1	chr1	6.239615
195	2	chr1	5.082789
		chr10	3.480455
196	1	chr25	6.237649
197	1	chr1	6.248001
198	1	chr16	6.244142
199	1	chr16	6.242848
200	1	chrX	6.317362
201	1	chr6	5.869959
202	1	chrX	6.202646
203	1	chr13	6.243548
204	1	chr13	6.266989
205	1	chr31	6.244635
206	1	chr34	6.224341

APPENDIX C (cont.)

207	1	chr1	6.228254
208	1	chr23	6.244708
209	1	chr31	6.232586
210	1	chr31	6.22802
211	1	chr19	6.184039
212	1	chr8	6.163537
213	1	chr20	6.244684
214	1	chr14	6.242024
215	1	chrX	6.224022
216	1	chr26	6.233104
217	1	chr13	6.088055
218	1	chr24	6.232955
219	1	chr6	6.242611
220	1	chr38	6.244415
221	1	chr3	6.232944
222	1	chr18	6.211941
223	1	chr31	6.063622
224	1	chr11	6.213921
225	1	chrX	6.226083
226	1	chr5	6.244929
227	1	chr22	6.218679
228	1	chr31	6.234257
229	1	chr35	6.225505
230	1	chr34	6.306824
231	1	chr26	6.24447
232	1	chr6	6.238061
233	1	chrX	6.228033
234	1	chr4	6.224015
235	1	chrX	6.238006
236	1	chr14	6.241644
237	1	chrX	6.220201
238	1	chr9	6.237753
239	2	chr35	4.241293
		chr5	4.307898
240	1	chrX	6.225878
241	1	chr17	6.203374
242	1	chr9	6.202422
243	1	chr22	6.123873
244	1	chr21	6.287691
245	1	chr22	6.19457

APPENDIX C (cont.)

246	1	chr31	6.23652
247	1	chr22	5.797196
248	1	chr7	6.241911
249	1	chr14	6.031946
250	1	chr18	6.166118
251	1	chr35	6.244843
252	1	chr13	6.191658
253	1	chrX	6.17262
254	1	chr25	6.217429
255	2	chr18	3.963774
		chr27	4.382976
256	1	chr1	6.225395
257	1	chr23	6.312672
258	1	chrX	6.237912
259	1	chrX	6.233034
260	1	chr31	6.240255
261	1	chr9	6.23264
262	1	chr7	6.244294
263	1	chr16	6.234706
264	1	chrX	6.24326
265	1	chr1	6.234297
266	1	chrX	6.206468
267	1	chr21	6.244474
268	1	chr16	6.231623
269	1	chr26	6.140089
270	1	chr1	6.242561
271	1	chr11	6.239312
272	1	chr19	6.244237
273	1	chr14	6.049229
274	1	chrX	6.038874
275	1	chr11	6.22138
276	1	chr16	6.226242
277	1	chrX	6.196331
278	1	chr21	6.229978
279	1	chrX	6.226127
280	1	chr20	6.244499
281	1	chr1	6.193964
282	1	chr35	6.215329
283	1	chr11	6.178477
284	1	chr31	6.191415

APPENDIX C (cont.)

285	1	chr5	6.242307
286	1	chr36	6.201462
287	1	chr1	6.243441
288	1	chr8	6.242306
289	1	chrX	6.228298
290	1	chr5	6.236642
291	1	chrX	6.241496
292	2	chrY	5.592
		chr8	1.655975
293	1	chr14	6.227101
294	1	chr18	6.230614
295	1	chr17	6.222931
296	1	chr11	6.195792
297	1	chr17	6.202917
298	1	chr38	6.098711
299	2	chr16	1.657821
		chr21	5.970576
300	1	chr35	5.274263
301	1	chrX	6.163806
302	1	chr31	6.112549
303	1	chr18	6.197971
304	1	chr11	6.180266
305	1	chr19	6.241836
306	1	chrX	6.170733
307	1	chr11	6.042552
308	1	chr1	6.218727
309	1	chr26	5.965679
310	1	chrY	3.556809
311	1	chrX	5.960777
312	1	chrX	6.209778
313	1	chrX	6.192623
314	1	chrX	6.157012
315	1	chr19	5.979252
316	1	chr5	6.214394
317	2	chr16	1.929878
		chr28	2.006761
318	1	chr16	5.811778
319	1	chr26	6.244367
320	1	chrX	6.236596
321	1	chrX	6.137837

APPENDIX C (cont.)

322	2	chr14	2.476812
		chr5	5.653956
323	1	chr22	6.244111
324	1	chrX	5.819683
325	1	chrX	6.175069
326	2	chr3	2.616712
		chr10	5.428106
327	2	chr3	1.962858
		chr1	1.859552
328	1	chrX	6.160084
329	1	chr1	6.240705
330	1	chr21	6.238274
331	1	chrX	6.223985
332	1	chr26	6.243383
333	1	chrX	6.207551
334	1	chrX	6.191969
335	1	chr8	6.209516
336	4	chr31	3.017558
		chr32	2.820281
		chr24	1.687448
		chr12	2.790023
338	1	chrX	6.162449
339	1	chr4	6.208267
340	1	chr32	5.659594
341	1	chrX	6.135633
342	1	chr13	5.887574
343	1	chr2	1.717884
344	1	chr34	6.244668
345	3	chr14	2.65753
		chr18	3.236975
		chr2	3.959582
346	1	chr19	6.204373
347	1	chrX	6.095731
348	5	chr32	1.817767
		chr18	1.96664
		chr22	1.795082
		chr4	1.732697
		chr8	2.82301
349	1	chr8	6.212203
350	1	chr2	4.859706

APPENDIX C (cont.)

351	1	chrX	6.232807
352	3	chr13	4.267999
		chr16	2.301837
		chr5	2.094578
353	1	chr16	6.234014
354	3	chr14	3.212599
		chr38	1.944473
		chr9	4.793937
355	1	chr12	1.755633
356	1	chr15	6.244696
357	1	chr17	6.234295
358	1	chr15	6.242415
359	1	chr13	6.244345
360	1	chrY	6.234744
361	1	chr18	5.849309
362	1	chr20	6.146074
363	2	chr15	2.271105
		chr20	2.563075
365	2	chr29	4.747066
		chr10	1.879342
366	1	chr1	6.146728
367	1	chrY	6.241795
368	4	chr31	3.257417
		chr32	1.995913
		chr6	1.847618
		chr12	2.692295
369	1	chrX	4.839704
370	1	chr22	6.222296
371	1	chr1	6.194118
372	1	chr13	2.352621
373	4	chr13	3.198452
		chr22	1.723698
		chr3	2.103073
		chr9	1.814534
374	2	chr31	2.025098
		chr16	2.2745
375	1	chr5	6.238649
376	1	chrX	5.760944
377	1	chr31	6.021246

APPENDIX C (cont.)

378	6	chr18	1.879524
		chr26	2.395687
		chr22	1.724551
		chr29	1.922017
		chr4	1.7308
		chr3	1.96576
379	1	chr31	6.136048
380	1	chr26	5.864659
381	1	chrX	6.130121
382	1	chr8	3.981162
383	1	chrX	5.535396
384	4	chr14	2.230055
		chrY	1.976023
		chr19	1.971091
		chr26	4.257378
385	1	chr31	4.555447
386	1	chrX	6.192445
388	1	chr1	6.199129
389	5	chr15	2.171436
		chr16	1.806358
		chr25	1.843611
		chr5	2.16771
		chr1	1.790215
390	1	chr5	6.236128
391	2	chr26	4.308978
		chr11	4.222707
392	1	chr7	1.891506
393	1	chr19	6.234125
394	1	chr19	6.239001
395	1	chr8	5.936556
396	1	chr19	6.243982
397	1	chr7	6.23998
398	1	chr1	5.613863
399	3	chr34	3.005563
		chr16	1.711776
		chr8	1.939332
400	1	chr16	6.007933
401	1	chr1	6.181853
402	1	chr22	6.153821
403	1	chr13	6.228726

APPENDIX C (cont.)

404	1	chr32	6.199222
405	1	chr1	6.195066
406	1	chr18	6.154244
407	1	chr8	5.92234
408	4	chr28	1.759806
		chr6	2.032779
		chr3	2.290586
		chr1	2.106077
409	1	chr19	5.352516
410	1	chr26	6.080336
411	1	chr17	6.241793
412	2	chr25	4.745281
		chr2	3.906641
413	1	chr16	5.596203
414	3	chr19	3.153658
		chr18	3.095834
		chr8	3.368431
415	1	chr16	5.484782
416	2	chr16	5.221452
		chr1	3.241398
417	1	chr26	6.18485
418	2	chr15	1.655477
		chr23	4.790009
419	2	chr3	3.787931
		chr10	4.841881
420	2	chr19	4.593976
		chr28	3.049925
421	1	chr19	6.238506
422	1	chr13	6.230945
423	3	chrX	2.607772
		chr11	2.664216
		chr8	2.669861
424	1	chr14	6.242327
425	1	chr14	6.228849
426	2	chr15	2.289047
		chr19	5.117716
427	1	chr8	6.227525
429	1	chrX	6.233725
430	1	chr8	6.23715

APPENDIX C (cont.)

431	3	chr21	2.930981
		chr28	2.39837
		chr11	3.316045
432	4	chr35	1.670657
		chr37	2.885846
		chr18	1.796555
		chr3	1.918804
433	1	chr11	6.24098
434	1	chr7	6.109032
435	1	chr1	6.195397
436	1	chr9	6.097478
437	1	chr1	6.158699
438	1	chr16	6.187989
439	1	chr18	6.121831
440	1	chr14	6.139723
441	1	chr19	6.029452
442	1	chrX	5.940926
444	1	chrY	6.239254
445	4	chr34	2.040331
		chr28	2.040331
		chr3	1.670855
		chr1	1.89993
446	1	chr26	6.243452
447	2	chr22	4.530107
		chr5	4.144082
448	1	chr22	6.224017
449	1	chr12	6.23906
450	1	chr19	6.222835
451	1	chr31	6.243137
452	1	chr31	5.903824
453	6	chr15	1.852496
		chrX	2.314096
		chr32	2.099032
		chr6	2.109523
		chr1	2.238037
		chr11	1.86561
454	1	chr22	6.139487
455	2	chr25	2.896812
		chr2	5.372216

APPENDIX C (cont.)

456	2	chr37	4.153199
		chr3	3.948103
457	1	chr11	5.623739
458	2	chr15	3.864481
		chr17	4.629638
459	6	chrX	1.683027
		chr34	1.97687
		chr18	2.523316
		chr21	2.069663
		chr4	1.951095
		chr11	2.098016
460	1	chr7	6.224008
461	1	chrX	6.24448
462	1	chr14	6.160044
463	1	chr14	1.854912
464	1	chr9	6.238184
465	1	chr18	2.556901
466	1	chr31	6.237649
467	1	chr2	5.640785
468	1	chr5	2.104379
469	1	chr11	1.728143
470	1	chr26	5.638699
471	1	chr22	6.240187
472	1	chr14	6.188812
473	1	chr15	6.242888
474	1	chr6	6.236977
475	2	chr19	2.603611
		chr22	5.606853
476	1	chr19	6.216503
477	1	chr32	6.238165
478	2	chr16	4.251126
		chr28	1.891649
479	1	chr11	6.179559
480	2	chr32	4.614874
		chr16	2.173337
481	2	chr34	3.912901
		chr16	3.93893
482	2	chr25	3.957681
		chr2	4.532137
483	1	chr34	6.179373

APPENDIX C (cont.)

484	2	chr14	5.035841
		chr9	3.548914
485	1	chr5	6.189977
486	1	chr13	6.243697
487	1	chr22	6.145503
488	1	chr19	6.207895
489	1	chr17	6.229816
490	1	chr1	6.201809
491	1	chr18	5.980626
492	1	chr19	6.213017
493	1	chr34	6.244675
494	1	chr14	6.225417
495	1	chr16	6.226951
496	1	chrX	6.135112
497	6	chrX	1.881921
		chr18	2.484051
		chr22	2.160646
		chr4	2.067028
		chr3	2.14788
		chr2	1.962773
498	2	chr22	2.290112
		chr5	5.728842
499	2	chr25	3.846796
		chr2	4.758033
500	2	chr34	2.664546
		chr12	5.544293

APPENDIX D

As an example, seven syntenic fragments identified by RACA. This data is taken from RACA's output file "Orthology.Blocks". For each block, the first line (e.g. ">1") represents the SF number, and the subsequent lines indicate the position of the fragment in each genome. Locations in dog (canFam3) and fox (vv2) must be unique and continuous, but more flexibility is permitted for the alignment to cat (felCat5).

>1
canFam3.chr11:5945-590625 +
vv2.scaffold304:40686-568218 -
felCat5.chrA1:89962061-90407488 +
felCat5.chrA1:99042298-99042300 -

>2
canFam3.chr11:802529-1605945 +
vv2.scaffold283:18015-767885 +
felCat5.chrA1:90467572-91115585 +

>3
canFam3.chr11:1605945-8923437 +
vv2.scaffold112:46444-7386489 -
felCat5.chrA1:91128525-98879548 +

>4
canFam3.chr11:9055290-9511633 +
vv2.scaffold224:0-436188 -
felCat5.chrA1:100068836-100514394 -

>5
canFam3.chr11:9594529-10212568 +
vv2.scaffold296:271-637648 -
felCat5.chrA1:99447990-100041503 -

>6
canFam3.chr11:10212568-10755694 +
vv2.scaffold307:3702-546607 +
felCat5.chrA1:99138499-99447990 -
felCat5.chrA1:99091502-99137856 +
felCat5.chrA1:98875896-99091311 -

>7
canFam3.chr11:11475533-11865091 +
vv2.scaffold224:372093-795279 -
felCat5.chrA1:100445107-100849313 -