Kent Academic Repository

Rando, Halie, Farré, Marta, Robson, Michael, Won, Naomi, Johnson, Jennifer, Buch, Ronak, Bastounes, Estelle, Xiang, Xueyan, Feng, Shaohong, Liu, Shiping and others (2018) *Construction of Red Fox Chromosomal Fragments from the Short-Read Genome Assembly.* Genes, 9 (6). ISSN 2073-4425.

Construction of Red Fox Chromosomal Fragments from the Short-Read Genome Assembly

Halie M. Rando [1,2] , **Marta Farré** [3] , **Michael P. Robson** [4] , **Naomi B. Won** [2] ,
Jennifer L. Johnson [2] , **Ronak Buch** [4] , **Estelle R. Bastounes** [2] , **Xueyan Xiang** [5] , **Shaohong Feng** [5] ,
Shiping Liu [5] , **Zijun Xiong** [5] , **Jaebum Kim** [6] , **Guojie Zhang** [5,7,8] , **Lyudmila N. Trut** [9] ,
Denis M. Larkin [3] **and Anna V. Kukekova** [2,*]

[1]   Illinois Informatics Institute, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA;
      rando2@illinois.edu

[2]   Department of Animal Science, College of Agricultural, Consumer and Environmental Sciences,
      University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA; krnommers@gmail.com (N.B.W.);
      jjohnso@illinois.edu (J.L.J.); bastoun2@illinois.edu (E.R.B.)

[3]   Department of Comparative Biomedical Science, Royal Veterinary College, London NW1 0TU, UK;
      mfbelmonte@rvc.ac.uk (M.F.); dlarkin@rvc.ac.uk (D.M.L.)

[4]   Department of Computer Science, College of Engineering, University of Illinois at Urbana-Champaign,
      Urbana, IL 61801, USA; mprobson@illinois.edu (M.P.R.); rabuch2@illinois.edu (R.B.)

[5]   China National Genebank, BGI -Shenzhen, Shenzhen 518083, Guangdong, China; x.xiang@uq.edu.au (X.X.);
      fengshaohong@genomics.cn (S.F.); liushiping@genomics.cn (S.L.); xiongzijun@genomics.cn (Z.X.);
      guojie.zhang@bio.ku.dk (G.Z.)

[6]   Department of Stem Cell and Regenerative Biology, Konkuk University, Seoul 05029, Korea;
      jbkim@konkuk.ac.kr

[7]   Section for Ecology and Evolution, Department of Biology, Universitetsparken 15, University of
      Copenhagen, DK-2100 Copenhagen, Denmark

[8]   State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology,
      Chinese Academy of Sciences, Kunming 650223, China

[9]   Institute of Cytology and Genetics of the Russian Academy of Sciences, Novosibirsk 630090, Russia;
      trut@bionet.nsc.ru

*   Correspondence: avk@illinois.edu; Tel.: +01-217-300-2425

Abstract: The genome of a red fox (*Vulpes vulpes*) was recently sequenced and assembled using next-generation sequencing (NGS). The assembly is of high quality, with 94X coverage and a scaffold N50 of 11.8 Mbp, but is split into 676,878 scaffolds, some of which are likely to contain assembly errors. Fragmentation and misassembly hinder accurate gene prediction and downstream analysis such as the identification of loci under selection. Therefore, assembly of the genome into chromosome-scale fragments was an important step towards developing this genomic model. Scaffolds from the assembly were aligned to the dog reference genome and compared to the alignment of an outgroup genome (cat) against the dog to identify syntenic sequences among species. The program Reference-Assisted Chromosome Assembly (RACA) then integrated the comparative alignment with the mapping of the raw sequencing reads generated during assembly against the fox scaffolds. The 128 sequence fragments RACA assembled were compared to the fox meiotic linkage map to guide the construction of 40 chromosomal fragments. This computational approach to assembly was facilitated by prior research in comparative mammalian genomics, and the continued improvement of the red fox genome can in turn offer insight into canid and carnivore chromosome evolution. This assembly is also necessary for advancing genetic research in foxes and other canids.

Keywords: *Vulpes vulpes*; comparative genomics; chromosome assembly; 10K Genomes Project; evolutionary breakpoints; Carnivora; Canidae; chromosome evolution; genome assembly

1. Introduction

At the turn of the millennium, the potential for mammalian comparative genomics to offer significant insights into both basic (e.g., adaptation and species formation) and applied (e.g., biomedical and agricultural) biology was already apparent [1]. However, early genome sequencing projects were costly investments, and mammalian genome sequencing projects were especially demanding given mammalian genomes' size and complexity. Thus, the original mammalian genome sequencing projects targeted popular model species such as human [2,3] and mouse [4]. Initial interest in comparative mammalian genomics was catalyzed by the recognized potential for comparative analysis to reveal regions of evolutionary constraint and to support the annotation of the human genome, but the high costs associated with genome assembly using Sanger sequencing technology limited the original comparative assembly project to the assembly of genomes from only 29 eutherian mammals [5]. More recently, the advent of low-cost, high-throughput next-generation sequencing (NGS) technology has revolutionized the scale at which comparative genomic projects can be approached. The effect of NGS technology on the scope of mammalian genomics is particularly manifest in the 10K Genomes Project [6], which seeks to sequence and assemble 10,000 vertebrate genomes and to ensure that every vertebrate genus is represented. As of 2015, 111 mammalian genome assembly projects, of which 105 were placental mammals, had been completed or were underway [7]. With this shift, many species outside the axis of traditionally-studied models have become the targets of mammalian genome sequencing projects.

One such species selected for genome sequencing through the Genome 10K Project is the red fox (*Vulpes vulpes*). In addition to being the widest-distributed wild terrestrial carnivore [8,9], the red fox has been bred to live on farms [10] and was also the subject of a unique experimental breeding program that began in 1959 at the Institute of Cytology and Genetics in Novosibirsk, Russia. For over 50 generations, one fox population has been bred to exhibit tame behavior towards humans similar to that of dogs while another has been bred to show heightened aggressive behavior towards humans [11]. Adaptation and response to selection in the red fox is therefore of interest not only to wildlife ecology, but also fields including behavioral and evolutionary genetics. The draft red fox genome assembly, which is currently in version 2.2 and known as vv2.2, has already proven valuable for studies of fox behavioral genetics [12], immunological adaptation [13], and population diversity [14].

While NGS can be credited with the proliferation of mammalian genome assemblies, the shift from Sanger to NGS technology is not without trade-off. Genome sequencing projects utilizing Sanger technology generated long sequencing reads (up to 800 bp), and also often included the development of physical or linkage maps. The length and low error rate [15] of Sanger reads rendered the de novo assembly of large fragments a computationally tractable problem with the overlap-layout-consensus (OLC) algorithm, which identifies a single best path through the sequencing reads [16]. This algorithm was used in the assembly of early eutherian mammal genome projects: for example, the dog genome [17] was assembled with ARACHNE [18,19], and the cat genome [20] with ARACHNE and Phusion [21]. By comparison, genomes today are often sequenced with short-read, error-prone technology whose limitations are offset by the reduced cost of sequencing at a very high coverage [15,22]. The OLC algorithm, however, becomes intractable if applied to the short-read data generated in modern genome sequencing projects because the short reads cause a proliferation of possible paths in repetitive regions [16,23]. Instead, short-read NGS genome assemblers rely on a heuristic approach to assembly that includes the deconstruction of the reads into *k*-mers followed by the traversal of a de Bruijn graph where the *k*-mers are the nodes [16,23,24]. This algorithm loses information by deconstructing the reads and is more susceptible to the introduction of errors by heterozygosity and repeats than the OLC approach is [25]. Utilizing libraries with a range of insert sizes, including long-range mate-pair libraries, is one critical step for reducing the uncertainty arising

from the use of short reads to improve assemblies and allow for assembly across repetitive regions, but it does not solve this problem [22,26]. Thus, the shift from Sanger sequencing to short-read NGS technologies has precipitated a new set of bioinformatic challenges related to assembly that affect the content of de novo assemblies in predictable ways.

The red fox genome was sequenced with short-read Illumina technology using 15 libraries with insert sizes ranging from 170 to 20,000 bp. Assembly was conducted using SOAPdenovo2, which uses de Bruijn graphs [27]. The genome assembly reports high coverage of 94X and a long scaffold N50 of 11.80 Mbp, but it is fragmented into 676,878 scaffolds [12]. The fragmentation of de novo assemblies sequenced with short reads typically occurs in repetitive regions [22,24] and is expected to introduce downstream challenges, such as the prediction of genes whose exons are split across multiple scaffolds [28]. Additionally, in the red fox genome, at least 70 of the 500 largest scaffolds (approximately 50 Kbp or larger) are chimeric [12], meaning they are likely to contain sequence from multiple fox chromosomes. Chimerism represents another problem known to affect de novo assemblies sequenced with short reads [22,27], as do within-chromosome structural errors such as the introduction of spurious rearrangements or segmental duplications [29–31]. Thus, although the high quality of the red fox genome assembly demonstrates once again the value of high-throughput, short-read technologies to mammalian genomics, the challenges of assembly from short reads have introduced some limitations that must be addressed in order to develop this genomic resource to its full potential.

The preferred outcome of improving the assembly would be to develop a full chromosomal assembly for the red fox. In the past, the large fragments assembled with Sanger sequencing facilitated the construction of chromosomes. The shift to short-read NGS technology has positioned the assembly of small fragments into chromosomes as a significant challenge in bioinformatics [32]. However, just as the assembly of additional genomes adds probative power to comparative genetic studies, comparison to assembled genomes can facilitate the assembly of chromosomes, or at least of chromosome-scale fragments. The program Reference-Assisted Chromosome Assembly (RACA), for example, utilizes synteny between the target genome of interest, a reference, and an outgroup to identify fragments that may be conserved. It then combines this syntenic information with the mapping of sequencing libraries made with a variety of insert sizes onto the genome of interest to identify the regions of the scaffolds that are robustly supported by both synteny and sequencing. In this way, the RACA pipeline can resolve assembly errors by breaking scaffolds into ungapped fragments colinear between the reference and target species (conserved blocks) from which it constructs larger, gapped colinear sequences (syntenic fragments) from the sequence content of multiple scaffolds. RACA then merges its syntenic fragments to produce sequences corresponding to predicted chromosome fragments (RACA fragments). RACA has previously been used in the development of chromosomal assemblies for the Tibetan Antelope [33] and the blind mole rat [34] as well as for the comparison of chromosome evolution among several avian species [35]. RACA can therefore be applied to the issue of fragmentation in the draft red fox genome, as the genomes of two close phylogenetic relatives of the fox have genome assemblies that include chromosomes. The dog (*Canis lupus familiaris*) and red fox shared an ancestor 9 to 10 million years ago [36], and the dog reference genome was sequenced and assembled into chromosomes using Sanger sequencing technology along with the development of tools including a Bacterial Artificial Chromosome (BAC) library and a radiation hybrid map of the dog genome [17]. Likewise, the cat (*Felis catus*) genome has been assembled into chromosomes [20] and offers a natural outgroup to the dog and fox, given that the canid and felid branches of Carnivora diverged about 50–60 million years ago [36].

An apparent limitation to using the dog and cat to assemble the fox chromosomes with RACA is the significant chromosomal rearrangement on the canid branch of the carnivore phylogenetic tree. The cat karyotype, which is comprised of 16 bi-armed and 2 single-armed autosomes [37], is thought to be similar to that of the ancestral carnivore [38,39]. While the fox's karyotype is composed of 16 metacentric autosomes, the sex chromosomes, and 0–8 supernumerary B chromosomes [37], the dog has 38 acrocentric autosomes in addition to its sex chromosomes [37]. Comparison of karyotypes within

the carnivore and canid clades suggests that the highly fragmented dog karyotype is very similar to the ancestral canid karyotype, and thus the dog-fox ancestor [40]. Syntenic blocks corresponding to regions of chromosomes conserved between the dog and fox have been characterized with comparative cytogenetic analysis [40,41] and with fluorescence in situ hybridization (FISH) using both flow-sorted dog chromosomes [42,43] and dog-derived BAC clones [44]. Synteny between the two species persists even at moderately high resolution: Construction of the red fox meiotic linkage map using canine microsatellite markers revealed that the order of the markers is almost completely conserved within dog and fox syntenic blocks [45–49]. These studies have demonstrated that each fox autosome is syntenic to at least two dog autosomes, and most dog autosomes map to the fox karyotype in a single syntenic block, although four dog chromosomes map in two blocks [40,42,44,46]. The relationship between the cat, fox, and dog chromosomes has also been established using FISH [50], and synteny between the dog and cat chromosomes has been characterized at a resolution of about 1 Mb using radiation-hybrid (RH) mapping [51]. Due to the lack of a meiotic or comparative RH-map, fox-cat synteny has never been examined at moderate- or high-resolution. The existing interspecies comparisons [50,51] suggest that dog and cat share much smaller syntenic blocks than dog and fox, consistent with the phylogeny.

The new red fox assembly sits at the intersection of traditional genome assembly pipelines and more modern ones: though it was assembled using the cost-effective short-read Illumina sequencing pipeline, two of its close phylogenetic relatives have Sanger-sequenced genomes that include full chromosome assemblies, and both physical and linkage mapping have established the relationships among the fox and dog chromosomes. The wealth of existing resources positions the new red fox genome assembly to both benefit from and contribute to the known relationships among carnivore and canid genomes. In the present study, the genomic sequence comprising the fox scaffolds was contextualized through sequence-level alignment to the dog reference chromosomes alongside a parallel alignment of the cat genome to the dog genome. RACA was used to combine this comparative genomic information with the raw sequence reads generated during genome assembly [12] to identify stretches of sequence colinear in dog and fox. These RACA fragments were then integrated with the dog-fox comparative chromosome map and the fox meiotic linkage map to assemble RACA's fragments into large fragments on the scale of fox chromosome arms.

Addressing and refining the errors introduced by assembly from short sequencing reads will facilitate evolutionary and population genetic analyses in the red fox and the mapping of the genetic architecture of quantitative traits of interest. Capitalizing on the products of early efforts in mammalian comparative genomics, namely the cat and dog reference genomes, to improve the fox genome assembly also advances the resolution at which karyotype evolution among the canids can be elucidated. The scaffolds of the red fox genome thus provide a valuable template for the construction of fox chromosomal fragments that can advance both organismal and comparative analysis of the red fox.

2. Materials and Methods

Syntenic chains and nets use interspecies alignments to identify runs of synteny between the species. In order to construct the chains and nets, the dog chromosomes comprising CanFam3.1 [17] were first partitioned into 40,010-bp pieces with 10 Kbp of overlap between pieces, and the 500 largest fox scaffolds (50 Kb and larger) (NCBI BioProject 378561) and the cat chromosomes from FelCat5 [5,20] were partitioned into 20-Kbp, nonoverlapping pieces. Fox and cat fragments were aligned against the dog fragments using LASTZ v1.02.00 [52,53]. For the fox-dog alignment, 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. Chains and nets had previously been developed for alignment of a previous version of the dog genome (CanFam2) with a previous version of the cat genome (FelCat3) [5], so the parameters from that alignment as listed by UCSC (University of California, Santa Cruz, CA, USA) Genome Browser were used in the current analysis. The LASTZ parameters were set to use the default scoring matrix and

O = 400, E = 30, K = 3000, L = 2200, H = 2000, and M = 50, with M specifying the threshold at which a sequence is considered to be repetitive and excluded from additional seeding. LASTZ alignments were run in parallel on the Carl R. Woese Institute for Genomic Biology's (University of Illinois at Urbana-Champaign, IL, USA) high-throughput computing cluster.

The alignments were merged to create chains and then nets, which are ungapped and gapped syntenic fragments (respectively), using the standard algorithms from the kentUtils v302 [54,55]: axtChain, chainSort, chainNet, and netSyntenic. The parameters in the chaining step for dog and fox included a minScore cut-off parameter of 5000 and a linearGap parameter set to medium. For cat and dog, the chaining parameters were set to a score threshold of 3000 and a linearGap of medium. The output of the chaining and netting step was a chain (.chain) and a net (.net) file corresponding to each dog chromosome compared to the fox, and the same for each chromosome compared to the cat.

The raw Illumina sequencing libraries (Table S1) generated in the original sequencing of the genome (NCBI BioProject 378561) were aligned against the largest 500 scaffolds in the draft genome using Burrows-Wheeler Alignment tool (BWA) 0.7.7 [56]. The alignments were converted into RACA's input format using the Perl script available on the RACA website [57]. Additionally, insert size statistics were measured for each of the 15 libraries using Picard v. 1.108 (Broad Institute, MA, USA) to calculate the mean and standard deviations of insert size mapping against the fox scaffolds (Table S1).

RACA (v. 0.9.1.1) [32] takes as input the dog-fox and dog-cat chains and nets, the insert size distributions, and the output of the Perl script from the RACA website. Additionally, RACA requires that the phylogenetic relationships among the three species be provided as a Newick tree, so this tree was estimated using the syntenic nets (Appendix A). The default RACA parameters were used for Window Size, Intracoverage Percentage, and Insert Size Threshold. RACA was run four times in order to test the block resolution parameter at 40, 80, 100, and 150 Kbp to identify its optimal value.

RACA evaluates the probability that fragments are adjacent based on support from overlapping sequencing reads; however, in the case of fox-dog synteny, strong a priori predictions of adjacency were available from previous analyses of interspecies synteny [40,42–44,46]. For this reason, the order of the markers and the known syntenic relationships between the dog and fox chromosomes were considered to be more reliable indicators of fragment order than RACA's fragment adjacency estimates. The number of dog chromosomes identified as syntenic to each scaffold by RACA was compared to their syntenic relationships with the dog genome as previously predicted [12]. The hypothesis that shorter scaffolds were more likely to be excluded by RACA, given that they would be less likely to contain runs of sequence longer than the Minimum Block Resolution, was tested using a one-tailed Welch Two Sample *t*-test in R [58] to compare the log10-transformed scaffold lengths.

The 414 microsatellite markers used previously in construction of the fox meiotic linkage map [45,46] were aligned to the fox genome. The primers for each marker were mapped with Bowtie 2 [59] to the fox scaffolds and then to the RACA fragments as though they were reads from paired-end sequencing. The insert size was set to a maximum of 700 and –D (the number of consecutive seed extension attempts that can fail before Bowtie 2 [59] skips ahead) was set to 50. Primer sets that mapped concordantly were considered to map robustly and were assigned a location.

The meiotic positions of these markers on the fox chromosomes, as well as their physical and genomic positions in the dog, are already known, so identifying them within the fox genome sequence served to bridge the gap between the linkage map and the contents of the genome. The RACA fragments at each of the tested block resolutions were then checked against the meiotic linkage map to identify extent of the concordance between the marker order in the fragments and their known order. The resolution at which the RACA fragments most faithfully recapitulated known marker order was 40 Kbp, so these fragments were used in all subsequent analyses.

Assembly of the fox chromosomes then proceeded in two steps: first, all RACA fragments mapping to a single dog chromosomes were concatenated based on the order of the microsatellite markers (first) and then according to known dog-fox synteny. When the dog sequence syntenic to two adjacent RACA fragments was not continuous, gaps were inserted based on the missing syntenic

sequence in the dog. In the next step, the new fragments were assigned positions along the fox chromosomes based on the known order and direction of the dog chromosomes relative to the fox chromosomes and the order of the markers on the fox chromosomes [42,46]. Unless they were mapped continuously in a single scaffold, segments of fox chromosomes that comprise distinct dog-fox syntenic blocks were assembled into separate fox chromosomal fragments because no assumptions could be made about the amount of sequence separating them. RACA also provided the position(s) in the cat genome corresponding to each RACA fragment, so the order of the fragments in cat was compared to cat-dog synteny as established by radiation hybrid mapping [51] and comparative cytogenetics [50] to confirm that the fragments recapitulated known cat-dog synteny.

Dog chromosomes CFA1, CFA13, CFA18, and CFA19 map in multiple syntenic blocks to the fox genome and correspond to evolutionary breakpoints. The exact breakpoints of the syntenic blocks along the fox chromosome have not previously been identified (Table S2). The RACA fragments syntenic to these four dog chromosomes were examined at high resolution and compared against the previous mapping of the breakpoints [44] to refine the syntenic positions of the breakpoints in the dog genome.

Finally, the fox chromosomal fragments were assembled as FASTA files using a Python script (available at https://github.com/rando2/foxmap) to extract genomic sequence from the scaffolds according to the positional ranges identified by RACA. Directionality of the fragments was determined based on, first, the orientation of the scaffold relative to the rest of the chromosomal fragment, and, second, the likely direction of that fragment in the fox genome based on dog-fox synteny [46]. Sequences from the scaffolds were reversed and complemented using Python as necessary. Gaps of 100 bp were added between scaffolds within a RACA fragment according to RACA's recommendation. Gaps between RACA fragments within fox chromosomal fragments were determined based on the space between their respective syntenic positions in the dog.

Pairwise alignments using the new fox genome chromosomal assembly as the reference and the dog (CanFam3.1) and cat (FelCat8) genomes as targets were generated with LASTZ v.1.02.00 [52,53] using the following parameters C = 0, E = 30, H = 2000, K = 3000, L = 2200, and O = 400, with C=0 specifying chaining should not be used (–nochain) but gaps can be used (–gapped). The resulting pairwise alignments were converted into the UCSC chains and nets alignment formats with axtChain [54,55] (parameters: −minScore = 1000, –linearGap = medium, and –verbose = 0) followed by chainAntiRepeat, chainSort, chainPreNet, chainNet, and netSyntenic, all with default parameters. Pairwise homologous synteny blocks were defined using the maf2synteny tool [60] at 300-Kbp resolution and uploaded to Evolution Highway [61] (http://eh-demo.ncsa.illinois.edu/).

3. Results

### 3.1. Interspecies Synteny

Large-scale parallel alignment of the largest 500 fox scaffolds (94% of the sequence in the draft genome) against the dog genome with LASTZ [52,53] and analysis with the kentUtils [54,55] produced chain and net files that define collinear sequence fragments representing synteny between the fox scaffolds and the dog chromosomes. Comparable alignments were produced of the cat and dog genomes.

3.2. Reference-Assisted Chromosome Assembly

With 40-Kbp block resolution, RACA identified 537 conserved blocks ranging in size from 41.4 Kbp to 54.7 Mbp of fox sequence, with sizes of the blocks ranging from 42.3 Kbp to 53.8 Mbp in the dog. The conserved blocks represent single, continuous regions on both a dog chromosome and a fox scaffold. RACA also provided corresponding location(s) in the cat genome for fragments that were homologous in all three species. Once read alignment information from the original genome sequencing reads was integrated, RACA merged these conserved blocks and added additional blocks to fill in gaps, resulting

in a set of 128 RACA fragments. The RACA fragments ranged in size from 34.9 Kbp to 104.6 Mbp. RACA's assembly closely followed the structure of the individual dog chromosomes, except for one fragment spanning both dog chromosomes 12 and 33 (CFA12 and CFA33), which together comprise the larger arm of fox chromosome 1 (VVU1q) [40]. Mapping an independent set of fox paired-end reads revealed support for 12.2% of the scaffold adjacencies suggested by RACA (Appendix B).

RACA incorporated sequence from 398 scaffolds into the fragments it assembled, meaning that 102 scaffolds analyzed by RACA were not included in the assembled fragments. Although the scaffolds used by RACA spanned the full range of represented scaffold lengths (50 Kbp to 55 Mbp), the excluded scaffolds were concentrated at the short end of the range (Figure 1). Comparing the log10 of the nucleotide lengths of the 500 scaffolds revealed that the scaffolds excluded from the assembly (mean: 4.91; standard deviation: 0.22) were significantly shorter than those included (mean: 6.31; standard deviation: 0.76) based on a Welch two sample *t*-test ($t (491.01) = -31.845$, $p < 2.2 \times 10^{-16}$). Six of the excluded scaffolds have been reported [12] as mapping to the dog Y-chromosome, which was not included in the present analysis.
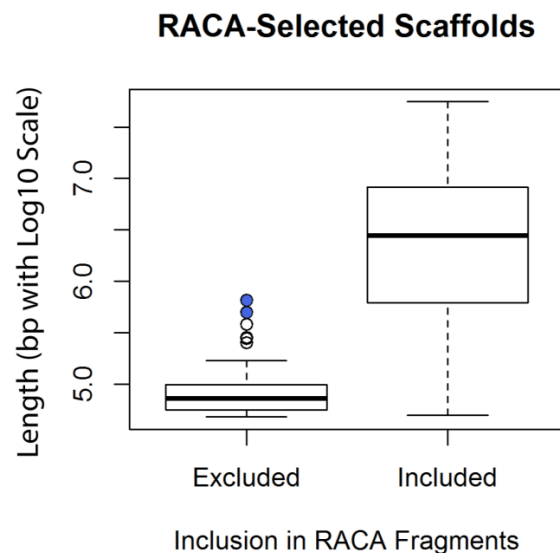
Figure 1. Box plot indicating lengths of scaffolds included and excluded from Reference-Assisted Chromosome Assembly's (RACA's) assembled fragments. RACA was provided with the largest 500 scaffolds, which ranged from 48 to 55,683 Kbp. The scaffolds included in the assembly spanned approximately the full range of sizes (50 to 55,683 Kbp) whereas the excluded scaffolds tended to be smaller (48 to 656 Kbp). The outlier points shaded in blue represent scaffolds 292 and 310, which were experimentally demonstrated to contain red fox Y-chromosome sequence [14], and therefore would not be expected to be included in RACA's assembled fragments.

The same prior analysis [12] had identified 70 scaffolds among the largest 500 as putative bioinformatic chimeras, meaning they were likely to contain misassembled sequence drawn from multiple fox chromosomes. Most of the scaffolds predicted to be syntenic to two or more dog chromosomes were split into the predicted number of fragments by RACA (Figure 2).

3.3. Assembly of Chromosome-Scale Fragments

The RACA fragments were integrated with the fox meiotic linkage map, which was constructed from dog-derived microsatellite markers whose order is known to be conserved along the fox chromosomal arms [46]. The microsatellite markers were mapped in silico onto the scaffolds and the RACA fragments to determine the order of the RACA fragments along the fox autosomes and X-chromosome. Of the 414 markers available, Bowtie 2 [59] mapped 373 concordantly to the largest 500 fox scaffolds, and all of those fell within RACA fragments (Table S3). Seventy-two of the 128 RACA

fragments contained at least one marker, and, of those, 57 contained two or more markers. The largest number of markers in a single RACA fragment was 25.

Presence of one or more markers allowed for the position of each fragment relative to the other fragments to be determined according to the fox meiotic linkage map, and when two or more markers mapped to a fragment, its direction relative to the other fragments could also be inferred. Fifty-six RACA fragments did not contain markers and were assigned positions along the fox chromosome arms according to their syntenic positions on the dog chromosomes and the previously identified patterns of fox-dog synteny (i.e., in the same order and orientation they would fall within the dog genome). Gaps were added between RACA fragments to be consistent with the corresponding unassembled dog sequence. Scaffolds 1 and 9, which are likely to overlap a historical fusion event, were recombined at the site where RACA had recommended splitting them because the order of the markers within each scaffold indicated that each scaffold contained sequence corresponding to two adjacent dog-fox syntenic blocks. Through this procedure, the number of fragments was reduced to 40. These chromosome-scale fragments ranged in size from 20.2 to 124.0 Mbp. All chromosomal fragments contain markers placed on prior linkage maps (Table S3).
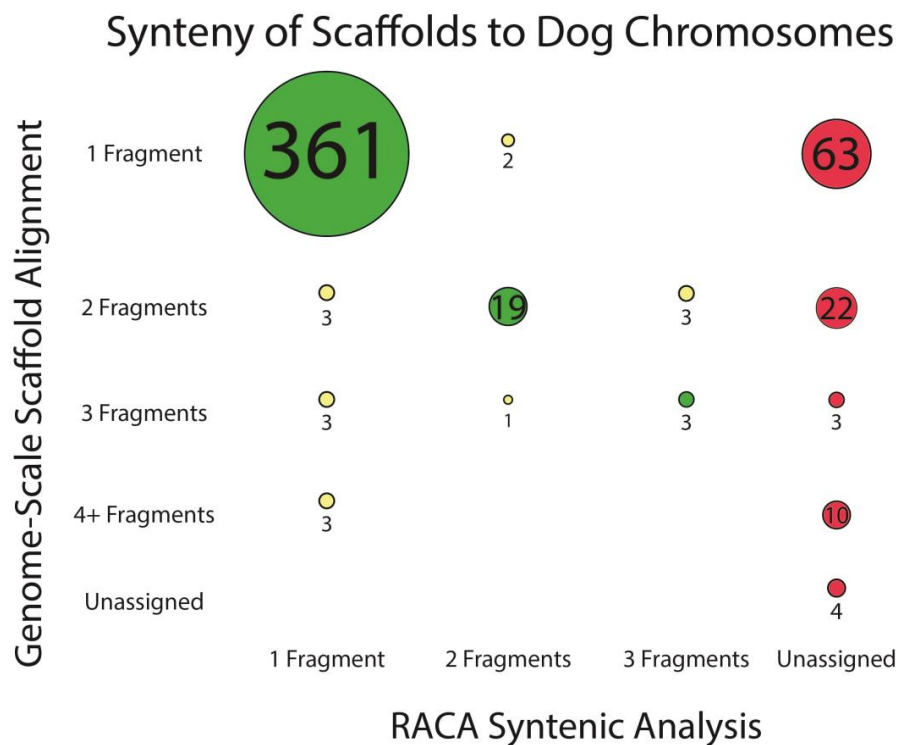
Figure 2. The number of distinct dog chromosomes syntenic to each scaffold, as predicted with two complementary methods: genome-scale alignment to identify syntenic dog chromosome(s) for each scaffold [12], and the number of dog chromosomes syntenic to each scaffold as identified by RACA. Circle diameter is proportional to the number of scaffolds, which is presented inside of or beside the circle. Green indicates that the same number of syntenic chromosomes was predicted by both methods; yellow that the methods predicted different numbers; and red that the scaffolds were excluded from the RACA assembly. Unlike the prior analysis, RACA did not compare the scaffolds to the dog Y-chromosome.

Comparison of the order of the 358 markers in RACA fragments that contained two or more markers (Table S3) to previous estimations of marker order [46–49] revealed a high level of congruence between the marker order in the fragments assembled by RACA, and as estimated with linkage. In all but 10 cases, the order of the markers in the assembly matched at least one previous map.

3.4. Refinement of Fox-Dog Synteny at Chromosomal Breakpoints

All but four dog chromosomes correspond to single syntenic blocks in the fox genome. CFA 1, 13, 18, and 19 each split into two distinct syntenic blocks when compared to the fox chromosomes, and previous studies have sought to characterize the dog genomic positions of the breakpoints in dog-fox synteny (Table 1). The RACA fragments homologous to these four dog chromosomes were analyzed to determine the syntenic position of the each relative to the fox breakpoint. Several RACA fragments were anchored in regions with known syntenic placement on the fox map but extended into the regions of the dog chromosomes whose locations relative to the breakpoint was unknown. These RACA fragments allowed for the size of the region unmapped to either the syntenic block to be reduced and refined the dog genomic positions of the breakpoints in fox (Table 1). One RACA fragment, 1a, contains sequence that is syntenic to either side of the breakpoint on dog chromosome 1 (209,340–423,404 bp and 24,994,866–25,534,824 bp); a lack of markers in this fragment prevented confident assignment to either fox chromosome 1 or 5, but it has been assembled with RACA fragment 1b to form fox chromosomal fragment VVU5p-proximal, which is the segment of fox chromosome 5 located proximally (that is, adjacent to the centromere) on the p arm.

Table 1. Refinement of gaps in synteny between dog and fox. Four dog chromosomes map in more than one syntenic block to the fox genome, meaning a single dog chromosome is syntenic to more than one contiguous region of the fox as the result of either a fission event in the dog lineage or a fusion event in the fox lineage. The breakpoints and sizes of the unassigned regions as refined by previous analysis [44] are indicated in the third and fourth columns, respectively. High-resolution mapping in the present study allowed for refinement of the breakpoints, as indicated in the fifth and sixth columns. All genomic locations are based on CanFam3.1. RACA fragment 1a is excluded due to its lack of markers.

Fox Chromosomes	Dog Chromosome	Previously Unassigned Positions on Dog Chromosome (bp)	Previous Gap Size (Kbp)	Dog Chromosome Positions Still Unassigned (bp)	Updated Gap Size (Kbp)
1 & 5	1	24,600,000–25,700,000	1100	24,988,836–25,579,247	590.4
2 & 13	13	37,800,000–38,600,000	800	38,258,211–38,277,954	19.7
5 (2 fragments)	18	24,400,000–26,000,000	1600	25,259,332–25,332,083	72.8
4 & 5	19	18,800,000–22,100,000	3300	19,878,341–20,333,685	455.3

3.5. Assembled Chromosomes

The sequences of the 40 assembled chromosomes are available online as NCBI BioProject 421766 and represent version 2.4 of the red fox genome assembly (vv2.4). Comparing the order of the RACA fragments across the three assemblies allowed for an interspecies comparative chromosome map to be inferred (Figure 3). Synteny between the chromosomal fragments and the dog and cat chromosomes was consistent with previously characterized synteny among the three species [40,42–44,50] with one exception: The cat chromosomes found to be syntenic to the regions of VVU1 and VVU5 that are syntenic to CFA1 were swapped relative to the previous dog-fox-cat comparative karyotype [50], but the positions in the comparative karyotype proposed here were consistent with the dog genomic positions syntenic to each cat chromosome as characterized with RH-mapping [51] and the known fox breakpoints [44].

Additionally, the high-resolution analysis identified previously unknown, short (between 20 Kbp and 11.9 Mbp) cat-fox syntenic blocks, many of which are consistent with dog-cat synteny as characterized in the dog-cat comparative RH map. The cat-fox comparison in Figure 3 is inferred based on the two species' alignment to the dog. As previously reported [50], synteny is conserved between fox and cat but not between fox and dog on VVU4 and VVU13 (Table 2). The high-resolution analysis revealed three cases (VVU1, VVU6, and VVU7) where fox sequence syntenic to a continuous region in the cat genome was split across both arms of a single fox chromosome (Table S4). Synteny among

fox, dog, and cat, with positions from both the fox draft genome and the chromosome assembly, are visualized in Evolution Highway [61] (Figure 4).
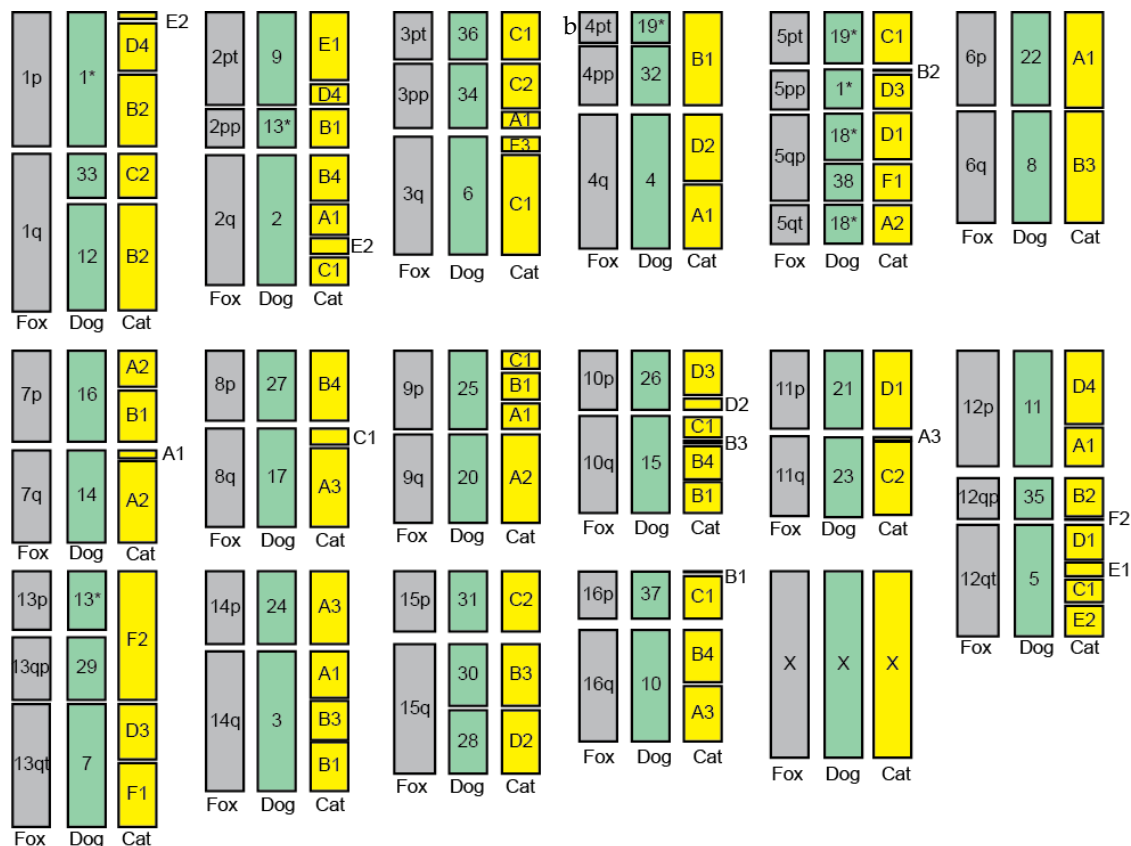
Figure 3. The fox chromosomal segments shown alongside the corresponding syntenic chromosomes in dog and cat. Fox chromosomal fragments are ordered to approximate the full assembled chromosome, and chromosome numbers appear inside of or alongside fragments. Asterisks indicate dog chromosomes that map in more than one syntenic block to fox. Dog and fox fragments are to scale; cat fragments are approximately to scale. Syntenic blocks smaller than 500 Kbp (found only for the cat) are not shown.

Table 2. Conserved synteny between cat and fox relative to dog. Alignment of the dog genome assembly (CanFam3.1) and cat genome assembly (FelCat5). Top: against VVU4; bottom: against VVU13. In these three cases, fox and cat map in a single syntenic block whereas fox and dog map in two blocks, as indicated by bold text. The cat syntenic blocks spanning two dog syntenic blocks are shaded. Genomic positions in CanFam3.1 and FelCat5 are indicated. * Indicates chromosome comprised by two dog-fox syntenic blocks.

Fox Fragment	Dog Chromosome	Dog Positions (Mbp)	Direction	Cat Chromosome	Cat Position (Mbp)
4pt	19 *	0–19.9	Forward	**B1**	**84.0–144.6**
4pp	32	0–38.7	Reverse		
4q	4	0–88.3	Forward	D2	8.6–48.2
				A1	172.6–227.4
13p	13 *	0–38.3	Reverse	**F2**	**0.6–82.8**
13qp	29	0–41.7	Reverse		
13qt	7	0–80.9	Reverse	D3	33.3–71.5
				F1	0.4–41.3; 65.7–68.7
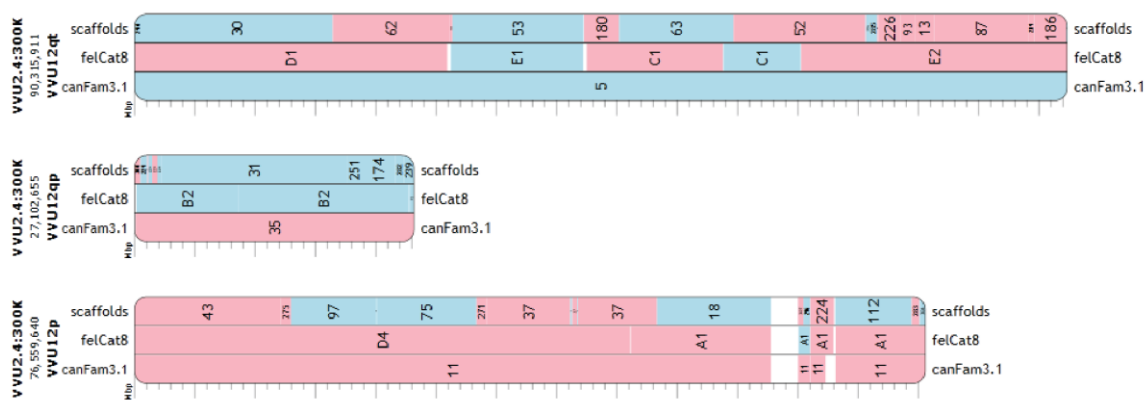
Figure 4. View of fox, dog, and cat synteny along fox chromosome 12. Interspecies chromosome-scale synteny has been visualized in Evolution Highway [61]. The fox scaffolds comprising vv2.2 and the final 40 fragments comprising vv2.4 are included in the visualization. Red represents fragments that run negative relative to the fox chromosome fragments (i.e., opposite strand), whereas blue fragments run in the same direction.

In almost all cases, the assembled fox chromosomal fragments split at locations where synteny breaks with the dog, such that the fragment is syntenic to a single dog chromosome. Three exceptions were found. Fox chromosomal fragment 1q was assembled by RACA based on read support along scaffold 7 to allow the fusion of regions syntenic to dog chromosomes 33 and 12. Additionally, fragments VVU5qp (the proximal segment of the larger (q) arm of fox chromosome 5) and 15q are cases where a scaffold bridges two adjacent dog-fox syntenic blocks. Although RACA did not suggest leaving the scaffold intact, the marker order indicated that the scaffold was likely to overlap an historical fox fusion. Thus, there are 40 fox chromosomal fragments in the current assembly, corresponding to 43 syntenic blocks between fox and dog. A full comparison of synteny among the three species across all fragments is provided in Table S4.

4. Discussion

The Genome 10K Project seeks to propel the coming-of-age of comparative genomics by sequencing a phylogenetically comprehensive set of vertebrate genomes. This objective was advanced, at least in the case of the red fox genome, by earlier research in comparative genomics that provides the resources needed to overcome the computational limitations of genome assembly with short-read NGS technology. Here, the scaffolds comprising the red fox genome were examined through the lens of comparative genomics to construct 40 fragments corresponding to partial or full fox chromosome arms. This pipeline allowed for the resolution of some assembly errors (e.g., chimerism) and for the red fox genome sequence to be reconciled with the red fox meiotic linkage map. The assembly of the red fox draft genome into chromosome-scale fragments will facilitate future genomic analysis of the fox and is thus a necessary step in the development of this genome.

The red fox genome project is positioned to build on resources developed over more than two decades that include the Sanger sequencing and OLC assembly of the chromosomes of two phylogenetically close species [5,17,20], characterization of the red fox chromosomes relative to the dog chromosomes through comparative cytogenetic analysis and FISH [40–44], and the development and refinement of a red fox meiotic linkage map using dog-derived markers [46–49]. RACA identified synteny between the red fox scaffolds and the dog chromosomes, and the placement of the markers along RACA's fragments allowed for the assembly of 40 fragments ranging in size from 20.2 to 124.0 Mbp. The assembled fragments comprise 2.38 Gbp of sequence including 2.34 Gbp of scaffold-derived sequence (i.e., excluding the gaps added by RACA or to combine RACA fragments). The chromosomal fragments contain 93.7% of the 2.5 Gbp of sequence originally reported [12].

Additionally, a minimum of 0.05% of the sequence contained in the original genome is derived from the Y-chromosome [14]. Thus, although the refined fox assembly incorporates genomic information from only 398 large scaffolds of the 676,878 scaffolds assembled, this corresponds to a loss of less than 3.5% of sequence information from the draft genome.

The fox chromosomal fragments are highly consistent with previous work using meiotic linkage mapping, with the order of the markers in the RACA fragments being almost identical to the established marker order on the fox map (Table S3). This concordance allowed for marker-guided assembly of the RACA fragments into the full and partial chromosomal arms. The ten discrepancies in marker order along the RACA fragments were placed in or near regions known to be affected by high levels of pericentromeric suppression in the fox [46] (Table S3). The placement of the markers in the RACA fragments therefore not only facilitated the assembly of the chromosomal fragments, but also offered new insight into regions that have been difficult to resolve using meiotic linkage.

Furthermore, explicitly developing the fox as a resource can provide additional support for research in chromosome evolution. The Canidae and Felidae branches of Carnivora diverged approximately 60 million years ago, and within the canids, four major clades are recognized: red-fox-like canids, South American canids, wolf-like canids, and the basal *Urocyon* clade [17]. The dog is a wolf-like canid, and the red fox belongs to its eponymous clade. While comparison of chromosomal rearrangements between the fox, dog, and cat clearly support the closer phylogenetic relationship between the dog and fox, there are at least two regions where synteny between the cat and fox chromosomes is more highly conserved than between dog and fox. The regions of VVU4 syntenic to cat chromosome B1 and of VVU13 syntenic to cat chromosome F2 have been found to form continuous syntenic blocks in cat [50] and canid lineages other than the wolf-like canid clade [44]. The present high-resolution analysis revealed regions where the evolutionary history may be more complex than previously known (Table S4). For example, on VVU6, RACA identified the regions corresponding to the telomeres of CFA22 and CFA8 (e.g., the VVU6 centromere) as being syntenic to cat chromosome A1. However, it does not appear that this region forms a continuous syntenic block with cat because there is 6.5 Mbp of missing cat sequence that RACA identified as being syntenic to VVU7 and these regions are not adjacent in other canids [39,44]. Similarly, sequences syntenic to cat chromosome B2: 31.9–154.2 Mbp split into multiple syntenic blocks across VVU1p and VVU1q, whereas these regions are syntenic to distinct chromosomes in the dog (CFA1 and CFA12, respectively). A similar pattern is found for the regions of VVU7p and VVU7q that are syntenic to cat chromosome A2. Whether or not these breakpoints have been reused [62–64] is a question for future analysis. Regardless, the strong support for at least two derived fissions in the wolf-like canids as identified by both physical [44,50] and now sequence-level synteny suggests that two modifications to the estimated ancestral canid karyotype [39] are needed. These modifications would reduce the number of unknown fragments in the ancestral canid karyotype (Appendix C).

High-resolution mapping of dog-fox syntenic blocks can elucidate regions associated with chromosome evolution. Twenty-six fusions and 4 fissions separate the red fox karyotype from the dog karyotype [42], with most fox chromosome arms corresponding to one or more acrocentric dog chromosomes. Identifying the sequence of the fox chromosomes presents an opportunity to characterize the regions associated with these fissions and fusions, especially in the cases where a single scaffold spans a historical fusion in the fox lineage (VVU1q, VVU5qp, and VVU15q). Likewise, analyzing the genomic content of the regions surrounding the four fissions can provide insights into chromosomal evolution in Canidae and beyond. The four fissions correspond to the breaks in dog-fox synteny seen on dog chromosomes 1, 13, 18, and 19. These loci are considered evolutionary breakpoint regions (EBRs) [44], which are positions along the chromosomes known to be the sites of a large number of fission events [39,40,44,50,65,66]. Based on the syntenic patterns observed among canid species [43,44,50], the breakpoints in the fox associated with CFA1, CFA13, and CFA19 all emerged from the same pattern of chromosome evolution: They likely belonged to distinct chromosomal fragments in the ancestral canid that fused in the ancestor of modern wolf-like and South American

canids, so they remain separate fragments in the rest of the canids, including the red fox. However, the evolutionary history of the breakpoint on CFA18 is more complex and suggests reuse of this EBR within recent canid chromosome evolution. Previous analysis [44] identified two syntenic blocks corresponding to CFA18 in both the red fox and the grey fox (*Urocyon cinereoargenteus*), but these regions form a single syntenic block in not only the wolf-like and South American canids but also some of the red-fox-like canids (i.e., raccoon dog and beat-eared fox). Becker and colleagues [44] mapped the breakpoints in the red fox and grey fox to the same regions along CFA18 but concluded that two distinct fusion events would best explain the pattern observed. They also noted that this syntenic block is adjacent to the CFA38-syntenic block not only in the red fox, but also in the grey fox, which belongs to the mode basal canid clade; this pattern is consistent with two separate events where these syntenic blocks were fused in the past 10 million years of canid evolution [44]. On VVU5qp in the proposed assembly of the fox chromosomal fragments, we present the sequence composition of an interstitial boundary of syntenic blocks corresponding to CFA38 and CFA18, although CFA38 and CFA18 share two boundaries on VVU5q, and the boundary assembled in scaffold9 is not the one recapitulated in the *Urocyon* karyotype. While the sequence of this EBR would be of particular interest to future analysis, the red fox chromosomal assembly, as a whole, will allow for higher resolution characterization of EBRs and the genomic content underlying chromosome evolution.

The chromosome fragments assembled here are essential for the advancement of genomic studies in the fox. Previous studies seeking to map behavioral traits in the fox have been forced to choose between relying on low-resolution microsatellite markers (e.g., [49]), interpreting against the dog genome assembly (e.g., [67]), or analyzing short genomic fox fragments (e.g., [12]). The large fragments will facilitate the use of high-resolution markers and the use of the fox as a reference genome for sequence alignment, while still allowing continuous analysis along fragments on the scale of chromosomal arms. The scale of the new fragments will facilitate the mapping of phenotypic traits, including the extreme behavioral phenotypes of the Novosibirsk farmed populations that influenced the inclusion of the red fox in the Genome 10K Project, and will advance future studies using in evolutionary and population genomics. Additionally, comparative analysis of wild fox populations and farmed foxes selected for behavior or other traits will be facilitated by the longer, continuous runs of sequence. For example, selective sweeps that reduce heterozygosity in regions surrounding a locus of interest will be more easily detected and the correct loci targeted when the sweep is recognized as contiguous, rather than being split across multiple scaffolds. Thus, the fragments themselves provide a new resource for analysis in the red fox.

Though reorganizing the scaffolds as chromosomes is important to the usability of the red fox genome, this assembly remains unfinished. As sequencing technologies and pipelines continue to evolve, new approaches to resolution of repetitive regions are becoming increasingly feasible. With the new technologies sometimes referred to as third-generation sequencing (3GS), sequencing reads of up to 20 Kbp are now available. The major limitation to many 3GS technologies is a high error rate compared to Sanger or short-read sequencing technologies [68]. Integrating reads from long-read platforms such as PacBio or Oxford Nanopore with short-read Illumina sequencing that can be used to correct errors has proven a valuable strategy for resolving complex regions of genomes [69,70]. In fact, new assembly algorithms have been developed to leverage both short- and long-read NGS technologies. One example is DBG2OLC, where short reads are assembled with de Bruijn graphs into contigs that are then used to correct errors in long reads, allowing for use of an OLC approach [68]. Approaches to sequencing itself have also emerged that obtain spatial information and sequence information simultaneously. These include Hi-C, which elucidates chromatin structure [71], and optical mapping, where fingerprints indicating the positions of restriction enzymes are visualized along long stretches of the genome [72]. Selection of individuals for sequencing can also play a role in improving de novo assemblies through strategies such as trio-binning, where sequencing parents alongside offspring facilitates the identification of specific haplotypes in a departure from the conventional choice to sequence inbred individuals [73]. Short-read-assembled genomes like that of the red fox will benefit

significantly from the improvements available with these approaches as 3GS becomes more widely available and economical.

Beyond the red fox itself, the chromosome fragments here and future improvements to them will offer a resource for analyses of chromosome evolution within Canidae. The scaffolds that contain sequence from two adjacent syntenic blocks offer an opportunity to analyze the genomic sequence comprising breakpoints during recent canid karyotype evolution, and, in one case an EBR, because the red fox is the first canid other than dog for which a genome has been assembled. Therefore, just as the red fox assembly has benefited greatly from the work of comparative mammalian genomics, it can also offer new insights into karyotype evolution in Canidae, the family with one of the most highly rearranged karyotypes [63]. The assembly of the fox draft genome into large fragments is thus a critical next step for genomic research in the red fox itself and in the red fox genome's potential to contribute to comparative mammalian genomics.

Appendix A

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 were converted to .maf format using netToAxt and axtToMaf from the KentUtils [54,55], then concatenated, and finally analyzed using phyloFit [74]. Within phyloFit, nucleotide substitution rates were estimated under the reversible nucleotide substitution model [75], as suggested in a previous work [76]. The following Newick Tree was produced: ((canFam3:0.0156945,vv2:0.00985954):0.0890152,felCat5:0.0890152)

The Newick tree itself is used by RACA, but Polydendron [77] offers a way to visualize the Newick tree for easier interpretation (Figure A1).
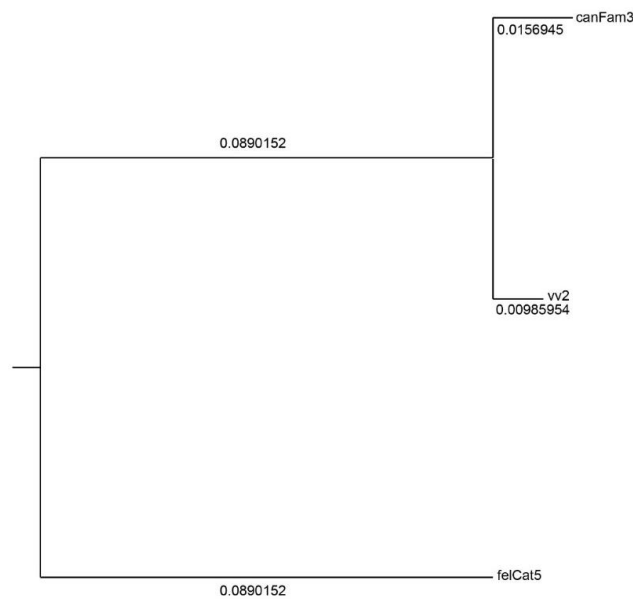
Figure A1. Newick tree visualization including branch lengths.

Appendix B

The goal of the assembly of the fox chromosome fragments was to reorganize the sequence contained in the scaffolds of the draft genome to more accurately recapitulate the structure of the red fox genome. To assess whether the chromosomal fragments represent an improved organization of the fox genome's sequence content, we analyzed whether some of the scaffold junctions suggested by RACA represented adjacent blocks of sequence. Breaks in the assembly corresponding to different scaffolds would be expected to occur in regions that are difficult to assemble, such as those rich with repetitive elements. Sequencing reads overlapping these boundaries may therefore be relatively rare, but would offer support for RACA's ordering of the fox scaffold sequences. To identify reads overlapping the RACA scaffold junctions, we analyzed pooled sequencing data from 10 red foxes (NCBI BioProject PRJNA376561; [12]) mapped to both the chromosome fragment assembly (vv2.4) and to the set of 398 scaffolds from vv2.2 that RACA used to assemble the RACA fragments.

The reads were first filtered to remove adaptor contamination and then aligned to both assemblies using Bowtie 2 [59]. For the 360 adjacencies where RACA added a 100-bp gap between the scaffolds, reads were counted if they spanned the gap (i.e., one end on either side of the gap), were oriented with both ends reading towards the gap, and aligned no more than 2000 bp apart based on the insert size reported in the SAM file. These reads were identified using SAMtools version 1.7 [78] to select the concordantly mapped reads that were placed within 5000 bp upstream or 5000 bp downstream from the gaps. The resulting SAM files were analyzed with Python 2.7 (Python Software Foundation, Beaverton, OR, USA) to count cases where one read in a pair was upstream of a gap and the other downstream with no more than 2000 bp between paired ends.

The mapping of the sequencing reads to the chromosome fragment assembly and to the 398 scaffolds used in the RACA assembly produced almost identical results. In both cases, the overall alignment rate was 94.06% and the percent of the reads that mapped concordantly exactly one time was 84.70%. The exact number of reads mapping concordantly exactly one time was slightly different: 345,168,075 pairs mapped to the chromosome fragment assembly and 345,170,193 pairs mapped to the scaffolds. This is a very small difference of 2,118 pairs of reads, or 0.0005% of all reads. This difference is likely due to RACA removing the ends of some of the scaffolds that were used in the chromosome fragment assembly that remained present in the set of scaffolds. In total, 44 of 360 (12.2%) of scaffold adjacencies were covered by sequencing reads.

Appendix C

The evolutionary origin of modern canid species has been a popular question even since the time of Darwin, and many approaches have been used to address it. One avenue of research has focused on evolutionary analysis of the modern canine/lupine karyotype, which has a large number of chromosomes (2n = 78). Using, at first, interspecies comparisons of chromosome banding patterns [79,80] and, later, comparative cytogenetic analysis with FISH [38,39], studies have sought to reconstruct karyotypes for extinct ancestral species within Carnivora (Ancestral Carnivore Karyotype, or ACK) and Canidae (Ancestral Karyotype of Extant Canids, or AKEC). The proposed ACK is very similar to that of the modern cat, with 2n = 42 [38,39], while the AKEC was estimated to have approximately 2n = 82, with substantial similarity to the modern dog karyotype [38,39]. At the time that the AKEC was published, however, the syntenic positions of several dog chromosomes were unknown. Advances in genomics have allowed for these regions to be mapped in high resolution, and therefore the present study proposes two updates to the AKEC from 2001.

Specifically, two regions were identified where the arrangement of the syntenic blocks in the ancestral canid was likely to have been more similar to the modern cat karyotype than to that of the modern dog. Therefore, the present results, in addition to evidence of these same patterns in other canid lineages including the basal *Urocyon* clade [44], suggest two modifications should be made to the AKEC (Figure A2). The extended homology of AKEC29 and VVU4 would encompass a larger region of Nash's ACK2q [39], which is homologous to cat B1q. The p arm of VVU4 is not the only region in which complexity would be reduced: based on the results of a comparison across multiple canid lineages [44], AKEC29 would also span Arctic Fox chromosome 6p and Raccoon Dog 5p in their entireties. In the second case, AKEC31 is homologous to cat chromosome F2 and would also be homologous to the full Arctic Fox chromosome 9 and a large interstitial region of Raccoon Dog 1, as depicted previously [39]. Though small changes, these updates reduce the complexity of the relationship of the AKEC and extant canid species and reduce the similarity between the AKEC and the modern dog in favor of species whose genomes have historically been less well characterized.
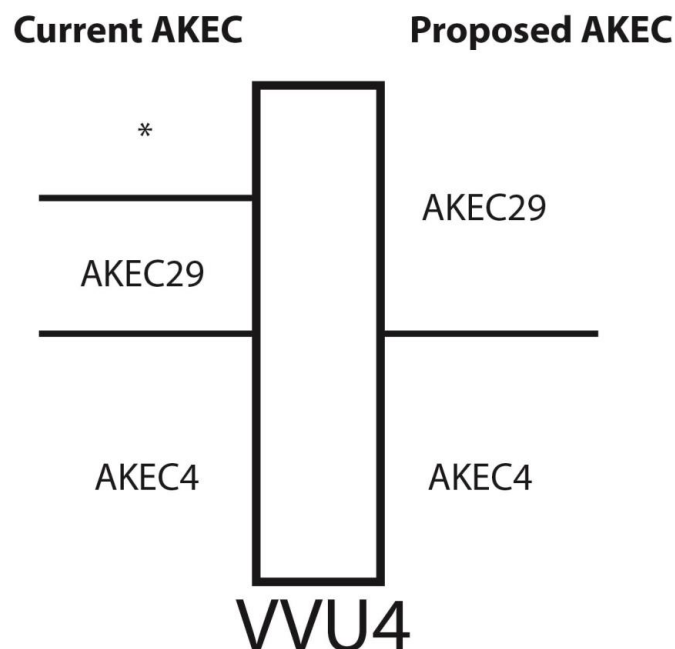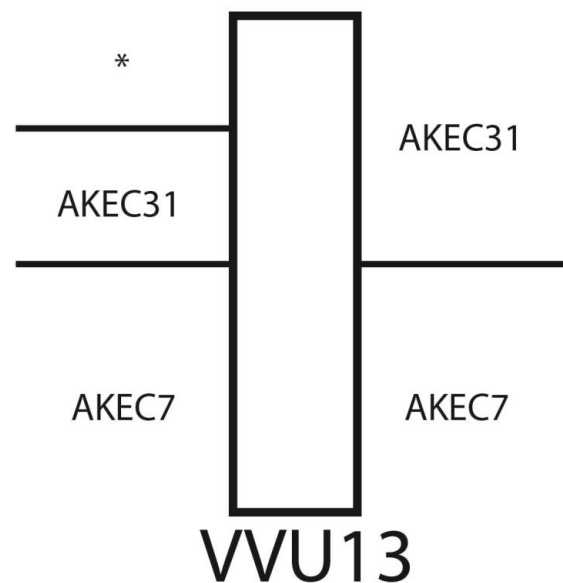

Figure A2. *Cont.*

Figure A2. Proposed updates to the Ancestral Karyotype of Extant Canids (AKEC) [39]. Asterisks denote regions of the fox chromosomes where the ancestral chromosome number was unknown in [39]. The original AKEC underestimates the coverage of VVU4 and VVU13 by the previously identified homologous ancestral chromosomes. The results of comparative genome mapping support a simpler homology between the ancestral chromosomes and the modern vulpine chromosomes, as indicated to the right.

References

1. O'Brien, S.J.; Menotti-Raymond, M.; Murphy, W.; Nash, W.G.; Wienberg, J.; Stanyon, R.; Copeland, N.G.; Jenkins, N.A.; Womack, J.E.; Marshall Graves, J.A. The promise of comparative genomics in mammals. *Science* **1999**, *286*, 458–462, 479–481. [CrossRef] [PubMed]

2. Lander, E.S.; Linton, L.M.; Birren, B.W.; Nusbaum, C.; Zody, M.C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; et al. Initial sequencing and analysis of the human genome. *Nature* **2001**, *409*, 860–921. [CrossRef] [PubMed]

3. Venter, J.C.; Adams, M.D.; Myers, E.W.; Li, P.W.; Mural, R.J.; Sutton, G.G.; Smith, H.O.; Yandell, M.; Evans, C.A.; Holt, R.A.; et al. The sequence of the human genome. *Science* **2001**, *291*, 1304–1351. [CrossRef] [PubMed]

4. Waterston, R.H.; Lindblad-Toh, K.; Birney, E.; Rogers, J.; Abril, J.F.; Agarwal, P.; Agarwala, R.; Ainscough, R.; Alexandersson, M.; An, P.; et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* **2002**, *420*, 520–562. [CrossRef] [PubMed]

5. Lindblad-Toh, K.; Garber, M.; Zuk, O.; Lin, M.F.; Parker, B.J.; Washietl, S.; Kheradpour, P.; Ernst, J.; Jordan, G.; Mauceli, E.; et al. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **2011**, *478*, 476–482. [CrossRef] [PubMed]

6. Genome 10K Consortium of Scientists. Genome 10K: A proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J. Hered.* **2009**, *100*, 659–674. [CrossRef]

7. Koepfli, K.-P.; Paten, B.; O'Brien, S.J. The Genome 10K Project: A way forward. *Annu. Rev. Anim. Biosci.* **2015**, *3*, 57–111. [CrossRef] [PubMed]

8. Lariviere, S.; Pasitschniak-Arts, M. *Vulpes vulpes*. *Mamm. Species* **1996**, *537*, 1–11. [CrossRef]

9. Schipper, J.; Chanson, J.S.; Chiozza, F.; Cox, N.A.; Hoffmann, M.; Katariya, V.; Lamoreux, J.; Rodrigues, A.S.L.; Stuart, S.N.; Temple, H.J.; et al. The status of the world's land and marine mammals: Diversity, threat and knowledge. *Science* **2008**, *322*, 225–230. [CrossRef] [PubMed]

10. Dearborn, N. *Silver Fox Farming in Eastern North America*; US Department of Agriculture: Washington, DC, USA, 1915; pp. 1–35.

11. Trut, L.N.; Oskina, I.; Kharlamova, A.V. Animal evolution during domestication: The domesticated fox as a model. *Bioessays* **2009**, *31*, 349–360. [CrossRef] [PubMed]

12. Kukekova, A.V.; Johnson, J.L.; Xiang, X.; Feng, S.; Liu, S.; Rando, H.M.; Kharlamova, A.V.; Herbeck, Y.; Serdyukova, N.A.; Xiong, Z.; et al. The red fox genome assembly facilitates identification of the genomic regions associated with selection for tame and aggressive behavior. *Nat. Ecol. Evol.* **2018**.

13. Donaldson, M.E.; Rico, Y.; Hueffer, K.; Rando, H.M.; Kukekova, A.V.; Kyle, C.J. Development of a genotype-by-sequencing immunogenetic assay as exemplified by screening for variation in red fox with and without endemic rabies exposure. *Ecol. Evol.* **2018**, *8*, 572–583. [CrossRef] [PubMed]

14. Rando, H.M.; Stutchman, J.T.; Bastounes, E.R.; Johnson, J.L.; Driscoll, C.A.; Barr, C.S.; Trut, L.N.; Sacks, B.N.; Kukekova, A.V. Y-chromosome markers for the red fox. *J. Hered.* **2017**, *108*, 678–685. [CrossRef] [PubMed]

15. Metzker, M.L. Sequencing technologies—The next generation. *Nat. Rev. Genet.* **2010**, *11*, 31–46. [CrossRef] [PubMed]

16. Li, Z.; Chen, Y.; Mu, D.; Yuan, J.; Shi, Y.; Zhang, H.; Gan, J.; Li, N.; Hu, X.; Liu, B.; et al. Comparison of the two major classes of assembly algorithms: Overlap-layout-consensus and de-bruijn-graph. *Brief. Funct. Genomics* **2012**, *11*, 25–37. [CrossRef] [PubMed]

17. Lindblad-Toh, K.; Wade, C.M.; Mikkelsen, T.S.; Karlsson, E.K.; Jaffe, D.B.; Kamal, M.; Clamp, M.; Chang, J.L.; Kulbokas, E.J.; Zody, M.C.; et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **2005**, *438*, 803–819. [CrossRef] [PubMed]

18. Batzoglou, S.; Jaffe, D.B.; Stanley, K.; Butler, J.; Gnerre, S.; Mauceli, E.; Berger, B.; Mesirov, J.P.; Lander, E.S. ARACHNE: A whole-genome shotgun assembler. *Genome Res.* **2002**, *12*, 177–189. [CrossRef] [PubMed]

19. Jaffe, D.B.; Butler, J.; Gnerre, S.; Mauceli, E.; Lindblad-Toh, K.; Mesirov, J.P.; Zody, M.C.; Lander, E.S. Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **2003**, *13*, 91–96. [CrossRef] [PubMed]

20. Pontius, J.U.; Mullikin, J.C.; Smith, D.R.; Agencourt, S.T.; Lindblad-Toh, K.; Gnerre, S.; Clamp, M.; Chang, J.L.; Stephens, R.; Neelam, B.; et al. Initial sequence and comparative analysis of the cat genome. *Genome Res.* **2007**, *17*, 1675–1689. [CrossRef] [PubMed]

21. Mullikin, J.C.; Ning, Z. The phusion assembler. *Genome Res.* **2003**, *13*, 81–90. [CrossRef] [PubMed]

22. Henson, J.; Tischler, G.; Ning, Z. Next-generation sequencing and large genome assemblies. *Pharmacogenomics* **2012**, *13*, 901–915. [CrossRef] [PubMed]

23. Pevzner, P.A.; Tang, H.; Waterman, M.S. An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 9748–9753. [CrossRef] [PubMed]

24. Compeau, P.E.C.; Pevzner, P.A.; Tesler, G. How to apply de Bruijn graphs to genome assembly. *Nat. Biotechnol.* **2011**, *29*, 987–991. [CrossRef] [PubMed]

25. Pop, M. Genome assembly reborn: Recent computational challenges. *Brief. Bioinform.* **2009**, *10*, 354–366. [CrossRef] [PubMed]

26. Li, R.; Zhu, H.; Ruan, J.; Qian, W.; Fang, X.; Shi, Z.; Li, Y.; Li, S.; Shan, G.; Kristiansen, K.; et al. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **2010**, *20*, 265–272. [CrossRef] [PubMed]

27. Luo, R.; Liu, B.; Xie, Y.; Li, Z.; Huang, W.; Yuan, J.; He, G.; Chen, Y.; Pan, Q.; Liu, Y.; et al. SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *Gigascience* **2012**, *1*, 18. [CrossRef] [PubMed]

28. Yao, G.; Ye, L.; Gao, H.; Minx, P.; Warren, W.C.; Weinstock, G.M. Graph accordance of next-generation sequence assemblies. *Bioinformatics* **2012**, *28*, 13–16. [CrossRef] [PubMed]

29. Eichler, E.E. Segmental duplications: What's missing, misassigned, and misassembled—And should we care? *Genome Res.* **2001**, *11*, 653–656. [CrossRef] [PubMed]

30. Kelley, D.R.; Salzberg, S.L. Detection and correction of false segmental duplications caused by genome mis-assembly. *Genome Biol.* **2010**, *11*, R28. [CrossRef] [PubMed]

31. Alkan, C.; Sajjadian, S.; Eichler, E.E. Limitations of next-generation genome sequence assembly. *Nat. Methods* **2011**, *8*, 61–65. [CrossRef] [PubMed]

32. Kim, J.; Larkin, D.M.; Cai, Q.; Asan; Zhang, Y.; Ge, R.-L.; Auvil, L.; Capitanu, B.; Zhang, G.; Lewin, H.A.; Ma, J. Reference-assisted chromosome assembly. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 1785–1790. [CrossRef] [PubMed]

33. Ge, R.-L.; Cai, Q.; Shen, Y.-Y.; San, A.; Ma, L.; Zhang, Y.; Yi, X.; Chen, Y.; Yang, L.; Huang, Y.; He, R.; et al. Draft genome sequence of the Tibetan antelope. *Nat. Commun.* **2013**, *4*, 1858. [CrossRef] [PubMed]

34. Fang, X.; Nevo, E.; Han, L.; Levanon, E.Y.; Zhao, J.; Avivi, A.; Larkin, D.; Jiang, X.; Feranchuk, S.; Zhu, Y.; et al. Genome-wide adaptive complexes to underground stresses in blind mole rats *Spalax. Nat. Commun.* **2014**, *5*, 3966. [CrossRef] [PubMed]

35. Damas, J.; O'Connor, R.; Farré, M.; Lenis, V.P.E.; Martell, H.J.; Mandawala, A.; Fowler, K.; Joseph, S.; Swain, M.T.; Griffin, D.K.; Larkin, D.M. Upgrading short read animal genome assemblies to chromosome level using comparative genomics and a universal probe set. *Genome Res.* **2016**, *27*, 1–10. [CrossRef] [PubMed]

36. Wayne, R.K. Molecular evolution of the dog family. *Trends Genet.* **1993**, *9*, 218–224. [CrossRef]

37. O'Brien, S.J.; Menninger, J.C.; Nash, W.G. *Atlas of Mammalian Chromosomes*; O'Brien, S.J., Menninger, J.C., Nash, W.G., Eds.; John Wiley and Sons: Hoboken, NJ, USA, 2006.

38. Murphy, W.J.; Stanyon, R.; O'Brien, S.J. Evolution of mammalian genome organization inferred from comparative gene mapping. *Genome Biol.* **2001**, *2*, 0005.1–0005.8. [CrossRef]

39. Nash, W.; Menninger, J.; Wienberg, J.; Padilla-Nash, H.; O'Brien, S. The pattern of phylogenomic evolution of the Canidae. *Cytogenet. Cell Genet.* **2001**, *95*, 210–224. [CrossRef] [PubMed]

40. Graphodatsky, A.S.; Perelman, P.L.; Sokolovskaya, N.V.; Beklemisheva, V.R.; Serdukova, N.A.; Dobigny, G.; O'Brien, S.J.; Ferguson-Smith, M.A.; Yang, F. Phylogenomics of the dog and fox family (Canidae, Carnivora) revealed by chromosome painting. *Chromosom. Res.* **2008**, *16*, 129–143. [CrossRef] [PubMed]

41. Graphodatsky, A.S.; Beklemisheva, V.R.; Dolf, G. High-resolution GTG-banding patterns of dog and silver fox chromosomes: Description and comparative analysis. *Cytogenet. Cell Genet.* **1995**, *69*, 226–231. [CrossRef] [PubMed]

42. Yang, F.; O'Brien, P.C.; Milne, B.; Graphodatsky, A.S.; Solanky, N.; Trifonov, V.; Rens, W.; Sargan, D.R.; Ferguson-Smith, M.A. A complete comparative chromosome map for the dog, red fox, and human and its integration with canine genetic maps. *Genomics* **1999**, *62*, 189–202. [CrossRef] [PubMed]

43. Graphodatsky, A.S.; Yang, F.; O'Brien, P.C.; Serdukova, N.A.; Milne, B.; Trifonov, V.; Ferguson-Smith, M.A. A comparative chromosome map of the Arctic fox, red fox and dog defined by chromosome painting and high resolution G-banding. *Chromosom. Res.* **2000**, *8*, 253–263. [CrossRef]

44. Becker, S.E.D.; Thomas, R.; Trifonov, V.A.; Wayne, R.K.; Graphodatsky, A.S.; Breen, M. 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. *Chromosom. Res.* **2011**, *19*, 685–708. [CrossRef] [PubMed]

45. Kukekova, A.V.; Trut, L.N.; Oskina, I.N.; Kharlamova, A.V.; Shikhevich, S.G.; Kirkness, E.F.; Aguirre, G.D.; Acland, G.M. A marker set for construction of a genetic map of the silver fox (*Vulpes vulpes*). *J. Hered.* **2004**, *95*, 185–194. [CrossRef] [PubMed]

46. Kukekova, A.V.; Trut, L.N.; Oskina, I.N.; Johnson, J.L.; Temnykh, S.V.; Kharlamova, A.V.; Shepeleva, D.V.; Gulievich, R.G.; Shikhevich, S.G.; Graphodatsky, A.S.; et al. A meiotic linkage map of the silver fox, aligned and compared to the canine genome. *Genome Res.* **2007**, *17*, 387–399. [CrossRef] [PubMed]

47. Kukekova, A.V.; Trut, L.N.; Chase, K.; Kharlamova, A.V.; Johnson, J.L.; Temnykh, S.V.; Oskina, I.N.; Gulevich, R.G.; Vladimirova, A.V.; Klebanov, S.; et al. Mapping loci for fox domestication: Deconstrution/reconstruction of a behavioral phenotype. *Behav. Genet.* **2011**, *41*, 593–606. [CrossRef] [PubMed]

48. Kukekova, A.V.; Temnykh, S.V.; Johnson, J.L.; Trut, L.N.; Acland, G.M. Genetics of behavior in the silver fox. *Mamm. Genome* **2012**, *23*, 164–177. [CrossRef] [PubMed]

49. Nelson, R.M.; Temnykh, S.V.; Johnson, J.L.; Kharlamova, A.V.; Vladimirova, A.V.; Gulevich, R.G.; Shepeleva, D.V.; Oskina, I.N.; Acland, G.M.; Rönnegård, L.; Trut, L.N.; Carlborg, Ö.; et al. Genetics of interactive behavior in silver foxes (*Vulpes vulpes*). *Behav. Genet.* **2017**, *47*, 88–101. [CrossRef] [PubMed]

50. Yang, F.; Graphodatsky, A.S.; O'Brien, P.C.; Colabella, A.; Solanky, N.; Squire, M.; Sargan, D.R.; Ferguson-Smith, M.A. Reciprocal chromosome painting illuminates the history of genome evolution of the domestic cat, dog and human. *Chromosom. Res.* **2000**, *8*, 393–404. [CrossRef]

51. Davis, B.W.; Raudsepp, T.; Pearks Wilkerson, A.J.; Agarwala, R.; Schäffer, A.A.; Houck, M.; Chowdhary, B.P.; Murphy, W.J. A high-resolution cat radiation hybrid and integrated FISH mapping resource for phylogenomic studies across Felidae. *Genomics* **2009**, *93*, 299–304. [CrossRef] [PubMed]

52. Harris, R.S. *Improved Pairwise Alignment of Genomic DNA*; The Pennsylvania State University: State College, PA, USA, 2007.

53. Harris, R.S. Miller Lab, LASTZ. Available online: http://www.bx.psu.edu/~rsharris/lastz/ (accessed on 17 March 2015).

54. Kent, J. kentUtils. Available online: https://github.com/ENCODE-DCC/kentUtils (accessed on 19 February 2015).

55. Kent, W.J.; Baertsch, R.; Hinrichs, A.; Miller, W.; Haussler, D. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 11484–11489. [CrossRef] [PubMed]

56. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [CrossRef] [PubMed]

57. Kim, J. RACA: Reference-Assisted Chromosome Assembly. Available online: http://bioen-compbio.bioen.illinois.edu/RACA/ (accessed on 17 April 2014).

58. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2014. Available online: http://www.R-project.org/ (accessed on 5 June 2018).

59. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [CrossRef] [PubMed]

60. Kolmogorov, M.; Raney, B.; Paten, B.; Pham, S. Ragout—A reference-assisted assembly tool for bacterial genomes. *Bioinformatics* **2014**, *30*, 302–309. [CrossRef] [PubMed]

61. Evolution Highway (fox). Available online: http://eh-demo.ncsa.illinois.edu/fox (accessed on 5 February 2018).

62. Murphy, W.J.; Larkin, D.M.; Everts-van der Wind, A.; Bourque, G.; Tesler, G.; Auvil, L.; Beever, J.E.; Chowdhary, B.P.; Galibert, F.; Gatzke, L.; et al. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* **2005**, *309*, 613–617. [CrossRef] [PubMed]

63. Larkin, D.M.; Pape, G.; Donthu, R.; Auvil, L.; Welge, M.; Lewin, H.A. Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. *Genome Res.* **2009**, *19*, 770–777. [CrossRef] [PubMed]

64. Sankoff, D. The where and wherefore of evolutionary breakpoints. *J. Biol.* **2009**, *8*, 66. [CrossRef] [PubMed]

65. Wayne, R.K.; Nash, W.; O'Brien, S.J. Chromosomal evolution of the Canidae. I. Species with high diploid numbers. *Cytogenet. Cell Genet.* **1987**, *44*, 123–133. [CrossRef] [PubMed]

66. Wayne, R.K.; Nash, W.; O'Brien, S.J. Chromosome evolution of the Canidae. II. Divergence from the primitive carnivore karyotype. *Cytogenet. Cell Genet.* **1987**, *44*, 134–141. [CrossRef] [PubMed]

67. Johnson, J.L.; Wittgenstein, H.; Mitchell, S.E.; Hyma, K.E.; Temnykh, S.V.; Kharlamova, A.V.; Gulevich, R.G.; Vladimirova, A.V.; Fong, H.W.F.; Acland, G.M.; et al. Genotyping-by-sequencing (GBS) detects genetic structure and confirms behavioral QTL in tame and aggressive foxes (*Vulpes vulpes*). *PLoS ONE* **2015**, *10*, e0127013. [CrossRef] [PubMed]

68. Ye, C.; Hill, C.M.; Wu, S.; Ruan, J.; Ma, Z. DBG2OLC: Efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Sci. Rep.* **2016**, *6*, 1–9. [CrossRef] [PubMed]

69. Jain, M.; Koren, S.; Miga, K.H.; Quick, J.; Rand, A.C.; Sasani, T.A.; Tyson, J.R.; Beggs, A.D.; Dilthey, A.T.; Fiddes, I.T.; et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **2018**, *36*, 338–345. [CrossRef] [PubMed]

70. Jiao, W.B.; Schneeberger, K. The impact of third generation genomic technologies on plant genome assembly. *Curr. Opin. Plant Biol.* **2017**, *36*, 64–70. [CrossRef] [PubMed]

71. Lieberman-Aiden, E.; van Berkum, N.; Williams, L.; Imakaev, M.; Ragoczy, T.; Telling, A.; Amit, I.; Lajoie, B.R.; Sabo, P.J.; Dorschner, M.O.; et al. Comprehensive mapping of long range interactions reveals folding principles of the human genome. *Science* **2009**, *326*, 289–293. [CrossRef] [PubMed]

72. Schwartz, D.; Li, X.; Hernandez, L.; Ramnarain, S.; Huff, E.; Wang, Y. Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. *Science* **1993**, *262*, 110–114. [CrossRef] [PubMed]

73. Koren, S.; Rhie, A.; Walenz, B.P.; Dilthey, A.T.; Bickhart, D.M.; Kingan, S.B.; Hiendleder, S.; Williams, J.L.; Smith, T.P.L.; Phillippy, A. Complete assembly of parental haplotypes with trio binning. *bioRxiv* **2018**, 271486. [CrossRef]

74. Hubisz, M.J.; Pollard, K.S.; Siepel, A. PHAST and RPHAST: Phylogenetic analysis with space/time models. *Brief. Bioinform.* **2011**, *12*, 41–51. [CrossRef] [PubMed]

75. Tavare, S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math. Life Sci.* **1986**, *17*, 57–86.

76. Yang, Z. Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* **1994**, *39*, 105–111. [CrossRef] [PubMed]

77. Gilbert, D.G. Polydendron version 0.8d. 1999. Available online: http://iubio.bio.indiana.edu/treeapp/treeprint-form.html (accessed on 6 July 2016).

78. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef] [PubMed]

79. Wurster-Hill, D.H.; Gray, C.W. The interrelationships of chromosome banding patterns in procyonids, viverrids, and felids. *Cytogenet. Cell Genet.* **1975**, *15*, 306–331. [CrossRef] [PubMed]

80. Dutrillaux, B.; Couturier, J. The ancestral karyotype of Carnivora: Comparison with that of platyrrhine monkeys. *Cytogenet. Cell Genet.* **1983**, *35*, 200–208. [CrossRef] [PubMed]