Genome-Wide Polymorphism and Comparative Analyses in the White-Tailed Deer (Odocoileus virginianus): A Model for Conservation Genomics

Christopher M. Seabury¹*, Eric K. Bhattarai¹, Jeremy F. Taylor², Ganesh G. Viswanathan¹, Susan M. Cooper³, Donald S. Davis¹, Scot E. Dowd⁴, Mitch L. Lockwood⁵, Paul M. Seabury⁶

¹Department of Veterinary Pathobiology, College of Veterinary Medicine, Texas A&M University, College Station, Texas, United States of America, ²Division of Animal Sciences, University of Missouri, Columbia, Missouri, United States of America, ³Texas AgriLife Research, Texas A&M University System, Uvalde, Texas, United States of America, ⁴Research and Testing Laboratories, SpiroStat Technologies, Medical Biofilm Research Institute, Lubbock, Texas, United States of America, ⁵Texas Parks and Wildlife Department, Kerrville, Texas, United States of America, ⁶ElanTech, Inc., Greenbelt, Maryland, United States of America

Abstract

The white-tailed deer (Odocoileus virginianus) represents one of the most successful and widely distributed large mammal species within North America, yet very little nucleotide sequence information is available. We utilized massively parallel pyrosequencing of a reduced representation library (RRL) and a random shotgun library (RSL) to generate a complete mitochondrial genome sequence and identify a large number of putative single nucleotide polymorphisms (SNPs) distributed throughout the white-tailed deer nuclear and mitochondrial genomes. A SNP validation study designed to test specific classes of putative SNPs provides evidence for as many as 10,476 genome-wide SNPs in the current dataset. Based on cytogenetic evidence for homology between cow (Bos taurus) and white-tailed deer chromosomes, we demonstrate that a divergent genome may be used for estimating the relative distribution and density of de novo sequence contigs as well as putative SNPs for species without draft genome assemblies. Our approach demonstrates that bioinformatic tools developed for model or agriculturally important species may be leveraged to support next-generation research programs for species of biological, ecological and evolutionary importance. We also provide a functional annotation analysis for the de novo sequence contigs assembled from white-tailed deer pyrosequencing reads, a mitochondrial phylogeny involving 13,722 nucleotide positions for 10 unique species of Cervidae, and a median joining haplotype network as a putative representation of mitochondrial evolution in O. virginianus. The results of this study are expected to provide a detailed template enabling genome-wide sequence-based studies of threatened, endangered or conservationally important non-model organisms.

Introduction

Organisms traditionally considered of “minor” importance by national and international funding agencies generally suffer from a paucity of genome-wide sequence and polymorphism data which severely limits the implementation of genomic approaches for addressing biological questions in these species. One such underserved species is the white-tailed deer (Odocoileus virginianus), a highly successful and widely distributed ruminant mammal species of the order Artiodactyla and family Cervidae [1–3]. Within the genus Odocoileus, the primary species are the white-tailed deer (O. virginianus) and mule deer (O. hemionus; for review see [2,3]), with these species possessing equivalent karyotypes (2n = 70) [4].

Based primarily on geographic variation in body size, differences in antler growth, and other relatively minor morphological characteristics, as many as 38 subspecies of O. virginianus have been suggested, including the endangered Florida key deer (O. virginianus clavium), and Columbian white-tailed deer (O. virginianus leucurus) [1,3]. Currently, free ranging white-tailed deer are ubiquitously distributed throughout most of the United States, with representative populations extending from Canada to Mexico, Central America, and South America [2,3,5]. Moreover, white-tailed deer have adapted to a variety of landscapes and environmental conditions while also exhibiting exceptional potential for recruitment [6–8], and despite historic overexploitation, appear to possess relatively high levels of genetic diversity [7,9,10].

Citation: Seabury CM, Bhattarai EK, Taylor JF, Viswanathan GG, Cooper SM, et al. (2011) Genome-Wide Polymorphism and Comparative Analyses in the White-Tailed Deer (Odocoileus virginianus): A Model for Conservation Genomics. PLoS ONE 6(1): e15811. doi:10.1371/journal.pone.0015811

Editor: Rudy A. Harteskier, Royal Tropical Institute, The Netherlands

Received September 14, 2010; Accepted November 24, 2010; Published January 19, 2011

Copyright: © 2011 Seabury et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The authors acknowledge the financial support of Texas AgriLife Research, Department of Veterinary Pathobiology, College of Veterinary Medicine, and Texas A&M University. DNA used in the validation study was partially enabled by Texas Parks and Wildlife Department grant #183260 to C.M.S. ElanTech, Inc. provided no funding for research, but allows Paul M. Seabury to participate as a collaborator. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript, with the following exceptions: Mitch L. Lockwood (TPWD Biologist) provided suggested revisions and references regarding white-tailed deer biology; Paul M. Seabury (ElanTech) provided and deployed custom scripts to parse data, as directed by Christopher M. Seabury.

Competing Interests: The authors have declared that no competing interests exist. Paul M. Seabury of ElanTech, Inc. is an unpaid collaborator who engineers and deploys custom scripts for genomics research. Custom scripts are situationally relevant to research and input files presented herein, and do not represent universal (commercial) software packages. Nevertheless, the scripts are available upon request, and the authors’ affiliation with ElanTech, Inc. does not interfere with PLoS ONE policies on sharing data or materials.

* E-mail: cseabury@cvm.tamu.edu
Unlike most traditional model organisms, and many free ranging wildlife species, white-tailed deer are distributed across a large range that includes both free ranging and captive populations, with a recent trend in North America toward establishing captively-propagated livestock lines from founders originating from a variety of wild populations. The farming of deer, and particularly white-tailed deer, has become a significant alternative livestock industry within the U.S. and Canada [11,12]. Consequently, U.S. captive and free-ranging populations of white-tailed deer have been subjected to intense surveillance for a variety of infectious diseases, with considerable efforts expended towards disease prevention and control [11–18]. Despite the growing numbers of captive and free ranging white-tailed deer under active management or surveillance in the U.S., little information exists concerning genomic diversity and divergence among populations. Previous genetic studies have primarily focused on genetic diversity, parentage, or the evaluation of population structure using relatively few microsatellite loci, with many of these markers derived from the domestic cow or sheep [7,9,10,19,20]. Recent research has targeted associations between naturally occurring polymorphisms within candidate genes and susceptibility to chronic wasting disease, a fatal transmissible spongiform encephalopathy (TSE) known to affect several species of Cervidae [21–24]. However, the inability to apply genome-wide approaches to address questions related to deer biology and evolution has severely limited progress. Thus, a substantial need exists to develop cost-effective de novo approaches which will rapidly enable sophisticated research programs for biologically important species for which research funds are limited.

We generated a reduced representation library (RRL) [25] to reduce the complexity of the white-tailed deer genome and a random shotgun library (RSL) to enable massively parallel pyrosequencing via the Roche 454 platform. The resulting random shotgun library (RSL) to enable massively parallel reduce the complexity of the white-tailed deer genome and a taurus

Based upon established homology between domestic cow (O. virginianus) and other species possessing lower coverage. Therefore, genomic sequence information derived from our RRL contigs will contribute to establishing an annotated white-tailed deer repeat library, and may also help elucidate potential copy number variants.

Alignment of the final white-tailed deer RRL contigs to the bovine genome sequence assembly (Btau4.0) via blastn resulted in 18,301 contigs producing 19,667 E-value informative hits (E-value≤1e-50) to either a single chromosome (BTA1-BTAX; MT; discrete unknown, chrUN; (≤3 chromosomal positions) or a single chromosome plus one unknown chromosome (≥3 chromosomal positions). These alignment criteria were chosen to maximize the likelihood of achieving unambiguous alignments while also allowing for potential gene family members, duplications, and limitations of the bovine genome assembly (i.e., assembly errors, chrUn unassigned sequence contigs). Overall, the average percent identity was approximately 92%, with an average alignment length of 306 bp, and 17,084 contigs (93%) produced a unique alignment to a bovine chromosome (Table S2).

Collectively, 6,877 putative SNPs (6,724 diallelic; 133 with ≥2 alleles) were detected within 18,301 blastn-aligned contigs using a 3X minimum depth of coverage for all potential variable sites (Table S3), with 5,710 (83%) putative SNPs derived from 17,084 contigs that produced one unique blastn hit. The average estimated minor allele frequency (MAF) for the 6,724 diallelic SNPs was 0.282. The distribution of blastn hits (n = 19,667) for all aligned contigs (n = 18,301) and putative SNPs (n = 6,877) against the bovine genome is shown in Figure S1, with similar results for

Results and Discussion

White-tailed Deer RRL Analysis

A white-tailed deer RRL, was constructed using 16 unrelated individuals representing northern and southern U.S. nuclear germplasm via A11 restriction enzyme digestion of a pooled DNA sample followed by a manual reduction in genome complexity via gel excision and purification of size-selected (≈350–400 bp) DNA fragments [26]. Collectively, 1,206,716 high quality reads comprising 285,269,784 bp of sequence were produced from this RRL on a Roche GS FLX instrument. A three stage assembly procedure employing strict requirements for read matching (sequence read length fraction = 0.90; similarity = 0.90) was used to create a reference assembly while also discouraging the creation of false contigs. The analytical workflow included an initial de novo assembly of RRL sequence reads, repeat masking of the resulting contigs, and utilization of the masked contig sequences to perform a reference assembly using the RRL sequencing reads (CLC Genomics Workbench 3.7.1). The resulting assembly contained 55,526 contigs comprising 19,207,189 bp of nucleotide sequence, with an average contig length of 346 bp. The minimum estimated repetitive DNA content for the 55,526 contigs was approximately 17%, as predicted by RepeatMasker (Human and/or Bovine Repeat Libraries). This relatively low estimate reflects our inability to mask all white-tailed deer repeats given the absence of a complete species-specific repeat library. Utilization of the masked contigs to perform a RRL reference assembly produced 44,385 final contigs averaging 338 bp, with approximately 6.2 sequence reads/contig, and a mean depth of 4.2X (Table S1). However, more than 95% of all contigs possessed <4X coverage (see Table S1 for coverage distribution), and when contigs possessing ≥20X coverage were excluded, the mean depth was approximately 2.1X (SD = 1.31). Unmasked repeats and/or potential copy number variants were apparent based on the observed depth of coverage achieved for the final contigs (see Table S1), with 392 contigs that possessed ≥20X coverage. However, it is also likely that some repeats and/or copy number variant positions are present in contigs possessing lower coverage. Therefore, genomic sequence information derived from our RRL contigs will contribute to establishing an annotated white-tailed deer repeat library, and may also help elucidate potential copy number variants.

Alignment of the final white-tailed deer RRL contigs to the bovine genome sequence assembly (Btau4.0) via blastn resulted in 18,301 contigs producing 19,667 E-value informative hits (E-value≤1e-50) to either a single chromosome (BTA1-BTAX; MT; discrete unknown, chrUN; (≤3 chromosomal positions) or a single chromosome plus one unknown chromosome (≥3 chromosomal positions). These alignment criteria were chosen to maximize the likelihood of achieving unambiguous alignments while also allowing for potential gene family members, duplications, and limitations of the bovine genome assembly (i.e., assembly errors, chrUn unassigned sequence contigs). Overall, the average percent identity was approximately 92%, with an average alignment length of 306 bp, and 17,084 contigs (93%) produced a unique alignment to a bovine chromosome (Table S2).

Collectively, 6,877 putative SNPs (6,724 diallelic; 133 with ≥2 alleles) were detected within 18,301 blastn-aligned contigs using a 3X minimum depth of coverage for all potential variable sites (Table S3), with 5,710 (83%) putative SNPs derived from 17,084 contigs that produced one unique blastn hit. The average estimated minor allele frequency (MAF) for the 6,724 diallelic SNPs was 0.282. The distribution of blastn hits (n = 19,667) for all aligned contigs (n = 18,301) and putative SNPs (n = 6,877) against the bovine genome is shown in Figure S1, with similar results for
the 17,084 uniquely aligned contigs and 5,710 putative SNPs depicted in Figure S2. The average deer-to-bovine hit density was one deer contig every 142.7±27.7 kb. Absence of BTAY annotation precluded Y-specific comparative contig overlays between *B. taurus* and white-tailed deer.

Interestingly, we observed a disproportionately large number of SNPs for deer contigs that aligned with BTAY. Further investigation revealed two clusters comprising 14 total contigs (17,084 uniquely aligned contigs and 5,710 putative SNPs) between BTA28 and white-tailed deer chromosomes. Interestingly, the homologue of CHRM3 and ZVE33B (11.35–11.39 Mb; n = 12 contigs); and 2) Within a putative intronic region of LOC532077 (11.620918–11.620982 Mb; n = 2 contigs). Both bovine regions are near a small break in the cattle-human comparative map [27] that is also proximal to the HSA10 centromere. Furthermore, this region of BTA28 is near the break between previously described HSA1 and HSA10 homologous synteny blocks [27]. Collectively, these 14 deer contigs contained 424 putative SNPs (Table S2), and may represent a duplicated and/or expanded region of the white-tailed deer genome that also serves as a break point between cattle and white-tailed deer chromosomes. Interestingly, the homologue of BTA28 has not yet been identified in *Odocoileus* species, with both white-tailed and mule deer possessing five more autosomes than cattle [4].

To functionally characterize the sequence content of 18,301 blastn-aligned contigs, we performed functional annotation, pathway mapping, and putative ortholog matching by mapping the assembled sequences onto relevant classification schemes such as Gene Ontology (GO) terms [29], KEGG pathways [29], and Swiss Prot Protein keywords [30] using both Krakenblast [31] and the Database for Annotation, Visualization, and Integrated Discovery (DAVID) [32]. Despite the relatively low genomic sequence coverage, 1,801 contigs (9.8%) produced functional hits and detailed annotation data, with 322 functional ontology terms, categories, and keywords cumulatively identified (Table S4).

**Pooled White-tailed Deer RRL and RSL Analysis**

In addition to the RRL, we constructed a random shotgun library (RSL) from a single male deer included in the RRL and produced 778,792 sequence reads comprising 286,843,168 bp on a Roche GS FLX instrument. The pooled RRL and RSL reads (n = 1,985,508) were assembled via CLC Genomics Workbench 3.7.1 using the previously described three stage *de novo* assembly process, which resulted in 126,980 contigs (average contig length of 433 bp) representing 55,020,760 bp of sequence. The minimum estimated repetitive content predicted by RepeatMasker was approximately 2.1X (SD = 1.45). Examination of the distribution of blastn hits (n = 61,553) for all blastn-aligned contigs (n = 56,084) and putative SNPs (n = 17,813) with respect to the bovine genome are shown in Figure S3. A plot for the 51,087 uniquely aligned contigs and corresponding 13,962 putative SNPs is displayed in Figure 1. For the pooled analysis (RRL+RSL), the average deer-to-bovine hit density was one deer contig every 45.9±9.9 kb. Similar to the RRL only analysis, we again observed a disproportionately large number of putative SNPs predicted for 27 white-tailed deer contigs that aligned to the same two regions of BTA28. In total, those 27 contigs contained 900 putative SNPs (Table S7).

In the RRL analysis, only six contigs provided evidence for a putative mitochondrial origin (Figure S1). However, our pooled RRL and RSL analysis effectively captured the complete white-tailed deer mitochondrial genome (GenBank Accession HQ332445) at an average coverage of 70.4X, with 2,965 reads included in the final assembly. Using the *B. taurus* and caribou (*Rangifer tarandus*) mitochondrial genome refseqs (GenBank accessions NC_006853.1; NC_007703) in conjunction with BLAST (blastn, tblastx, blastp; http://blast.ncbi.nlm.nih.gov/), we successfully annotated 13 white-tailed deer mitochondrial protein coding genes (*ND1*, *ND2*, *COX1*, *COX2*, *ATP8*, *ATP6*, *COX3*, *ND3*, *ND4L*, *ND4*, *ND5*, *ND6*, *CITB*) and two ribosomal RNA genes (12S, 16S; GenBank Accession HQ332445). Using tRNAscan-SE (http://lowelab.ucsc.edu/tRNAscan-SE/) [33], we also predicted 21 tRNA genes (GenBank Accession HQ332445). The consensus mitochondrial genome spanned 16,477 contiguous bp and possessed an average GC content of 36.9%. In total, 38 putative diallelic SNPs were detected with ≥3X coverage (average coverage 78X) and an overall average estimated MAF of 0.07 (Table S8). A blastn query of the white-tailed deer mitochondrial genome sequence against the nucleotide collection (nr/nt) using BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) produced a top hit (based on 100% coverage; 93% Max Identity) to the complete mitochondrial genome of *R. tarandus* (GenBank Accession NC_007703). To functionally characterize the sequence content of the 56,084 blastn-aligned contigs, the assembled sequences were again mapped onto the relevant classification schemes, with 3,057 contigs (9%) producing putative functional hits and detailed annotation data, including 606 identified functional ontology terms, categories, and keywords (Table S9). Pooling sequences obtained from the RRL and RSL resulted in a >2.8 fold increase in the total number of contigs predicted to possess at least partial gene sequences, thereby elucidating nucleotide sequence data for as many as 3,097 putative genes (see unique hits Table S9).

**SNP Validation and Mitochondrial Evolution**

Using a DNA panel of 96 white-tailed deer that included the 16 deer used to create the RRL and RSL, we estimated the proportion of total putative SNPs that were likely to be valid by fluorescent allele-specific PCR [34]. Putative nuclear SNPs were sampled from a range of discovery classes defined by differences in....
depth of coverage and observed minor allele count (Table 1), with efforts to avoid clustered SNPs and those predicted within or proximal to homopolymers. Every putative SNP detected within the white-tailed deer mitochondrial genome was also tested (n = 38). SNP validation provided evidence for as many as 10,448 nuclear SNPs (Table 1) and 28 mitochondrial SNPs (Table S7, S8). Lower validation rates for mitochondrial SNPs are likely a reflection of two primary issues: 1) Not all sequence compositions, including but not limited to tightly clustered SNPs, are amenable to genotyping by allele specific KASPar assays [34]; and 2) Some of the putative mitochondrial SNPs queried were predicted in regions that were rich in one or two specific nucleotides (including homopolymers), which likely increased local read error. Median joining haplotype networks constructed for both captive and free-ranging white-tailed deer revealed extensive mitochondrial haplotype variation and divergence (Figure 2). Using the white-tailed deer mitochondrial genome consensus sequence, we implemented a Bayesian approach [35] to estimate the evolutionary history for 10 species of Cervidae (Figure 3). Collectively, 13,722 nucleotide positions spanning the 12S rRNA, 16S rRNA, ND1, ND2, COX1, COX2, ATP8, ATP6, COX3, ND3, ND4L, ND4, ND5, ND6, and CYTB genes were used to produce the phylogeny. Distance, maximum composite likelihood, and maximum parsimony approaches (transitions+transversions) employing bootstrap resampling produced similar overall tree topologies, with the exception that *Elaphodus cephalophus* and *Hydropotes inermis* could not be unambiguously placed within the tree due to inadequate bootstrap support (bootstrap ≤ 70; trees not shown). Our Bayesian mitochondrial phylogeny (Figure 3) provides strong support for monophyly in Cervinae, Muntiacinae, and Odocoileinae for the surveyed taxa while also providing evidence that the single species representing the subfamily Hydropotinae (*H. inermis*) is more closely related to Odocoileinae (new world deer) than any other subfamily represented in our analysis. These findings are generally consistent with recent phylogenies produced from either partial mitochondrial genome sequences [36–38] and/or the hybrid combination of nuclear and mitochondrial sequence information [39]. Additionally, the position of *O. virginianus* within our tree is consistent with a nuclear phylogenomics study involving 39,695 parsimony informative characters derived from 40,843 nuclear SNPs [40]. Our ability to produce a robust mitochondrial phylogeny indicates that similar sequencing approaches in other non-model organisms are also likely to produce massive amounts of phylogenetically informative data.

Figure 1. Comparative Contig Overlay with Putative SNPs. Histogram displaying the bovine chromosome locations (Btau4.0) of 51,087 uniquely aligned (E ≤ 1e-50) white-tailed deer (WTD) sequence contigs and 13,962 putative SNPs (≥ 3X coverage) derived from sequencing a WTD reduced representation library (RRL) and random shotgun library (RSL; Also See Table S6, S7, and S8). Note, one WTD contig that aligned to BTA13 (Consensusfrom53142contig) was removed by NCBI filtering. doi:10.1371/journal.pone.0015811.g001
would cost approximately $70,000 per GS FLX instrument run, with each fully optimized run producing approximately 400 Mb of sequence data, a theoretical 1X coverage of the white-tailed deer genome (10481Y; (Codon 211; Ile→Ile); Table S8) was also variable (C/T) in at least four lineages of Bos, including B. taurus, B. indicus, B. frontalis, and B. javanicus (Genbank Accessions HM045018.1, EU177870.1, EF685907.1, AY079130.2, AB542189.1, FJ356556, AY689188.1, FJ997262.1). Importantly, amino acid 211 of the CYTB gene falls within a domain of unknown function according to the Simple Modular Architecture Research Tool online (http://smart.embl-heidelberg.de/). The unusual nature of this shared synonymous polymorphism provides some potential evidence for either selective and/or functional constraint at amino acid 211 across several highly divergent taxa and should be further investigated.

**Prospects for a complete white-tailed deer genome sequence**

Generation of a white-tailed deer genome sequence assembly would provide a valuable resource for population geneticists and physiologists studying the basis of adaptation and fecundity [5–8]. In addition to interesting biological characteristics, burgeoning physiologists studying the basis of adaptation and fecundity [5–8].

At a cost of $10,000 per GS FLX instrument run, with each fully optimized run producing approximately 400 Mb of sequence data, a theoretical 1X coverage of the white-tailed deer genome would cost approximately $70,000, assuming a genome size similar to that of the domestic cow (B. taurus, 2.87 Gb genome; [48]). Moreover, a hybrid combination of currently available long (Roche GS FLX) and short read sequencing technologies (Illumina or Applied Biosystems) would generate sufficient sequence data from mate-pair and paired-end libraries to allow a de novo white-tailed deer assembly at a modest cost, with single molecule sequencing [49] soon to further reduce the cost of de novo genome assemblies. However, every genome sequence needs a good map [50], and the SNPs identified in this study will enable linkage and/or linkage disequilibrium approaches to producing a robust physical and/or genetic map capable of guiding the white-tailed deer sequence assembly.

**Conclusions and Future Studies**

This study demonstrates that small, highly targeted investments in sequencing for “minor” species can rapidly enable genomic and phylogenomic research programs via adaptation of bioinformatic procedures developed for human or well-funded food animal species [25,26,51,52]. We have generated a substantial number of high quality white-tailed deer nuclear sequence contigs which were unambiguously aligned to the bovine genome assembly despite 32.5 million years of species divergence, as predicted from a mitochondrial CYTB and 16S rRNA analysis [53]. Our relatively small, targeted sequencing initiative lead to the generation of a complete white-tailed deer mitochondrial genome sequence and phylogeny while also providing a new resource for population-based studies and genetic discrimination analyses for Odocoileus species. The mitochondrial genome sequence, nuclear genome sequence contigs, putative and validated SNPs, and functional annotation will directly facilitate conservation, population, management, and genetic epidemiology research in O. virginianus and O. hemionus while also enabling conservation genomics in the endangered Florida Key deer.

**Materials and Methods**

**Construction and Sequencing of the RRL and RSL**

In the absence of a genome sequence against which to perform an in silico whole genome digestion [51], a molecular prediction based upon a domestic animal species was employed. Using knowledge from the construction of a pig (Sus scrofa) RRL [26], two restriction enzymes (Alul; HauIII, New England Biolabs) which recognize four base sequences [AG]CT; [GG]CC) and produce blunt-ended fragments were tested via overnight digestion of 2.8 μg of white-tailed deer DNA (O. virginianus; 37°C; 8 U enzyme/μg DNA). Alul was selected for RRL preparation based on modest repetitive content observed within size-selected swine digestion fragments [26], its apparent ability to fully digest white-
tailed deer DNA, presence of the restriction site in deer mitochondrial sequences (Genbank accessions U12869.1; M35874.1), and the paucity and distribution of visually apparent bands following digestion (Figure S4).

Equal amounts of DNA were pooled from 16 (n = 7 females; n = 9 males) unrelated white-tailed deer representing a sampling of northern and southern U.S. nuclear germplasm derived from both captive and free ranging populations. Ethical clearance is not applicable to samples obtained from lawfully harvested white-tailed deer. A total of 14 mg of DNA was digested overnight (37°C; 8 U enzyme/mg DNA) with AluI, as suggested by the manufacturer (New England Biolabs). The resulting fragments were separated via 2.5% ULTRA-SIEVE high resolution agarose (IBI Scientific) gel electrophoresis according to instructions for enhanced separation (IBI Scientific), with ethidium bromide staining for visualization. Three regions of the gel were excised for purification: A) Fragments <350–400 bp, with efforts to avoid a faint band at <340 bp; B) Fragments <220–280 bp; and C) Fragments <160–180 bp. Restriction fragments were purified using the Qiaquick Gel Extraction Kit (Qiagen), with an extended elution incubation period (10–12 min) and reduced elution volume (20 µl). Following gel purification, the smaller fragments were archived and the largest fragment population (<350–400 bp) was utilized for pyrosequencing (Figure S4). A sequencing library was constructed via random ligation of sequencing adaptors provided with the GS FLX titanium library kit (Roche Applied Science). All library preparation, emulsion PCR, quantitation, and sequencing steps followed manufacturer protocols (Roche Applied Science).

Sequence Assembly, SNP Detection, BLAST annotation, and tRNA Prediction

Two assemblies, each with SNP detection and BLAST annotation analyses were performed using identical workflows. For the first assembly, only sequence reads generated from the RRL were utilized (RRL only), whereas the second assembly included sequence reads from both the RRL and the RSL (Pooled RRL+RSL). In both cases, a three step assembly procedure was utilized in conjunction with CLC Genomics Workbench 3.7.1 (CLC Bio) and RepeatMaster (http://www.repeatmasker.org/). Unmasked sequence reads were imported into CLC Genomics Workbench 3.7.1 and a de novo assembly was performed with user defined parameters: add conflict annotations = yes, conflict resolution = vote (A,T,C,G), create report = yes, create sequence list = yes, non-specific matches = ignore, minimum contig length = 200 bp, mismatch cost = 2, insertion cost = 3, deletion cost = 3, minimum read length fraction = 0.90, minimum fraction of identity (similarity) = 0.90. Contigs produced were processed with RepeatMasker (http://www.repeatmasker.org/; RepBase15.0.2), and the masked contigs became the reference sequences used for SNP discovery; an approach similar to that used for the rainbow trout [52]. Masked contig reference sequences were exported for subsequent filtering by local blastn searches against Btau4.0 (E≤1e-50) and scenario specific (RRL only; RRL+RSL) SNP detection.
analyses. Custom scripts were engineered to parse contig assignments to bovine chromosomes as follows: E-value informative hit to BTA1-BTX; MT; discrete chrUN (≤3 chromosomal positions); to a single chromosome plus one discrete chrUN (≤3 chromosomal positions), and to a unique chromosomal location. The distribution and density of all contig alignments were examined relative to their predicted locations within the bovine genome (Btau4.0).

SNP detection analyses employed the Neighborhood Quality Standard algorithm [25,54] within CLC Genomics Workbench 3.7.1, with the following user defined and default parameters: annotate consensus = yes, annotate reference = yes, create table = yes, maximum coverage = 1000X, maximum gap and mismatch count = 2, minimum average quality = 15, minimum central quality = 20, minimum coverage = 3X, minimum variant frequency = 10% or count 3, SNP analysis window = 11 bp. Custom scripts were developed to parse putative SNP locations from contigs aligned to Btau4.0, and their genomic distribution was assessed against Btau4.0.

**Functional Annotation Analyses**

White-tailed deer consensus contig sequences were mapped to functional classification schemes such as Gene Ontology (GO) terms [28], KEGG pathways [29], and Swiss Prot Protein

---

**Figure 3. Mitochondrial Phylogeny.** Majority rule consensus phylogenetic tree for 10 species of Cervidae inferred from concatenated 12S rRNA, 16S rRNA, ND1, ND2, COX1, COX2, ATP8, ATP6, COX3, ND3, ND4L, ND4, ND5, ND6, and CYTB gene sequences. Entries in parenthesis represent Genbank Accession Numbers. The *Odocoileus virginianus* consensus sequence was derived from this study. Start codons, stop codons, and overlapping regions of genes were excluded. Numbers beside branches indicate posterior probabilities estimated from a Bayesian analysis employing $2 \times 10^6$ generations [35].

doi:10.1371/journal.pone.0015811.g003
Keywords [30], via Kraken (www.krakenblast.com) which is a high throughput distributed BLAST engine based upon WND-BLAST [31] using the BLASTx algorithm and a database containing all functionally annotated protein sequences contained within Uniprot (March 2010 version), with cross reference mapping performed using DAVID [32].

SNP Validation, Mitochondrial Phylogenetics, and Network Analysis

Nuclear SNPs were selected from white-tailed deer contigs that produced a single blastn hit (E-value ≤ 1e-50) to one discrete bovine chromosomal coordinate (Table S7). All SNPs were selected based on the coverage and minor allele count criteria contained in Table 1, with putative SNPs predicted within or near homopolymers excluded. We also endeavored to select putative SNPs from contigs that uniquely aligned to different bovine autosomes and BTAX. Putative nuclear SNPs closely flanked by another predicted polymorphism were avoided; however, all putative mitochondrial SNPs were evaluated. Collectively, 116 putative SNPs were genotyped using the KASPar allele-specific fluorescent genotyping system [Kbiosciences; Table S7, S8] [34]. Assays exhibiting poor genotype clustering after two rounds of optimization were considered technical failures due to quality control violations [34].

All mitochondrial sequences were aligned using the ClustalW2 webserver (http://www.ebi.ac.uk/Tools/clustalw2/index.html), with manual adjustment (alignment available upon request). A Bayesian approach implemented within the program MrBayes 3.1.2 [35] was used to estimate the evolutionary history of 10 species of Cervidae by evaluating random trees using 2 × 10^6 generations, GTR model with gamma-distributed rate variation, and four simultaneous Markov chains [38]. Start codons, stop codons, and overlapping regions of genes were excluded. A majority rule consensus tree rooted using the B. taurus outgroup was constructed and visualized within TreeView 1.6.6. For comparison, we also constructed phylogenetic trees using distance (Kimura 2 parameter), maximum composite likelihood, and maximum parsimony (transitions-transversions) with bootstrap resampling (1000 replicates; trees not shown), as referenced and implemented within the program MEGA4 [53]. All median joining haplotype networks were constructed using Network 4.5.1.0 (Fluxus Technology LTD) [34].

Supporting Information

Figure S1 Comparative Contig Overlay with Putative SNPs. Histogram displaying the bovine chromosome locations (Btaa4.0) of 19,667 blastn hits (E-value ≤ 1e-50) for 18,301 white-tailed deer (WTD) sequence contigs and 6,877 putative SNPs (≥3X coverage) derived from sequencing a WTD reduced representation library (RRL).

Figure S2 Comparative Contig Overlay with Putative SNPs. Histogram displaying the bovine chromosome locations (Btaa4.0) of 17,084 uniquely aligned (E-value ≤ 1e-50) white-tailed deer (WTD) sequence contigs and 5,710 putative SNPs (≥3X coverage) derived from sequencing a WTD reduced representation library (RRL).

Figure S3 Comparative Contig Overlay with Putative SNPs. Histogram displaying the bovine chromosome locations (Btaa4.0) of 61,553 blastn hits (E-value ≤ 1e-50) for 56,084 white-tailed deer (WTD) sequence contigs and 17,813 putative SNPs (≥3X coverage) derived from sequencing a WTD reduced representation library (RRL) and random shotgun library (RSL).

Acknowledgments

We thank Shawn Prince and Jannick Bendtsen for CLC support. C.M.S. thanks Jeff Skelton and Enrique Terrazas for workstation support, Dr. Cain and John Sewell for access to archived DNA, and Steven Hoffman for Roche support. We also thank Texas A&M College of Veterinary Medicine and Texas Agrilife Research for support. C.M.S. thanks Dr. Mike Heaton for insightful RRL discussions. Data deposition: Contig fasta files have been deposited in Genbank (WGS Accessions AEGY00000000; AEGZ00000000). The contig accessions described in this paper are the first version. Note, 43 contigs were removed by NCBI filtering. A complete mitochondrial genome was deposited in Genbank (Accession HQ332445). Validated mitochondrial and nuclear SNP data were deposited in dbSNP (Accessions ss263191552–ss263191579; ss263191363–ss263191425).
Author Contributions
Conceived and designed the experiments: CMS. Performed the experiments: CMS EKB SED. Analyzed the data: CMS. Contributed reagents/materials/analyses: CMS GGV. Provided the manuscript: CMS. Engineered custom scripts: PMS GGV. Provided

References
1. Baker RH (1984) White-tailed Deer: Ecology and Management Hills LK, ed. Harrisburg Stackpole Books, pp 1–18.
2. O’Rourke KI, Walker’s Mammals of the World Volume II, sixth ed. Baltimore: The Johns Hopkins University Press. pp 1114–1118.
3. Wilson DE, Reeder DM (2005) Mammal Species of the World: A Taxonomic and Geographic Reference, third ed. Baltimore: The Johns Hopkins University Press. pp 657–685.
4. Gallagher DS, Jr., Derr JN, Womack JE (1994) Chromosome conservation among the advanced pachytes and determination of the primitive bovid karyotype. J Hered 85: 204–210.
5. Warren RJ (1997) The challenge of deer overabundance in the 21st century. Wildlife Society Bulletin 25: 213–214.
6. Bubenik GA, Brown RD, Schams D, Bartos L (1999) The effect of ACTH on the Odocoileus virginianus restoration in Mississippi. Mol Ecol 12: 3237–3252.
7. Locke SL, Coakley LS, Havens DA, Davis DS, Lopez RL, et al. (2007) Effectiveness of Spayvac® for reducing white-tailed deer feral. J Wildl Dis 43: 726–730.
8. Anderson JD, Honeycutt RL, Gonzales RA, Gee KL, Skow LC, et al. (2002) Development of microsatellite DNA markers for the automated genetic characterization of white-tailed deer populations. J Wildl Manage 66: 67–74.
9. DeYoung RW, Demarais S, Honeycutt RL, Gonzales RA, et al. (2009) Evaluation of a deer microsatellite panel useful for genetic exclusion studies in white-tailed deer. Wildlife Society Bulletin 31: 220–232.
10. Waters WR, Palmer MV, Thacker TC, Orloski K, Nol P, et al. (2008) Blood surveillance to detect chronic wasting disease in white-tailed deer in Wisconsin. J Wildl Dis 44: 1025–1030.
11. Decker JE, Pires JC, Conant GC, McKay SD, Heaton MP, et al. (2009) The diversity and evolution of 11 immune genes in Bos taurus taurus and Bos taurus indicus cattle. Proc Natl Acad Sci USA 106: 151–156.
12. Pang H, Lui W, Chen Y, Fang L, Zhang X, et al. (2000) Identification of complete mitochondrial DNA of the tufted deer. Mitochondrial DNA 19: 1572–1574.
13. Douzery E, Randi E (1997) The mitochondrial control region of Cervidae: Evolutionary patterns and phylogenetic content. Mol Biol Evol 14: 1154–1166.
14. Wada K, Nishibori M, Yokohama M (2007) The complete nucleotide sequence of mitochondrial genome in the Japanese Sika deer (Cervus nippon), and a phylogenetic analysis between Cervidae and Bovidae. Small Ruminant Res 69: 46–54.
15. DeYoung RW, Demarais S, Honeycutt RL, Gonzales RA, et al. (2009) Evaluation of white-tailed deer fertility. J Wildl Dis 45: 153–164.
16. DeYoung RW, Demarais S, Honeycutt RL, Gonzales RA, et al. (2009) Evolution of white-tailed deer. J Wildl Dis 45: 989–997.
17. Keane D, Barr D, Osborn R, Langenberg J, O’Rourke K, et al. (2009) Surveillance to detect chronic wasting disease in white-tailed deer in Wisconsin. J Wildl Dis 45: 989–997.
18. Waters WR, Palmer MV, Thacker TC, Orloski K, Nol P, et al. (2008) Blood surveillance to detect chronic wasting disease in white-tailed deer in Wisconsin. J Wildl Dis 44: 1025–1030.
19. Wobeser G (2009) Bovine tuberculosis in Canadian wildlife: An Updated history. J Wildl Dis 45: 989–997.
20. Ernest HB, Hoar BR, Well JA, O’Rourke KI (2010) Molecular genalogy tools for white-tailed deer with chronic wasting disease. Can J Vet Res 74: 153–156.
21. Williams ES, Miller MW (2002) Chronic wasting disease in deer and elk in North America. Rev Sci Tech 21: 305–316.
22. O’Rourke KI, Spraker TR, Hames LS, Besser TE, Brayton KE, et al. (2004) Detection of a novel reassortant epizootic hemorrhagic disease virus (EHDV-2) in the USA containing RNA segments derived from both exotic (EHDV-6) and endemic (EHDV-2) serotypes. J Gen Virol 91: 430–439.
23. DeYoung RW, Demarais S, Gonzales RA, Honeycutt RL, Gee KL (2002) Multiple paternity in white-tailed deer (Odocoileus virginianus) revealed by DNA microsatellites. J Mammal 83: 884–892.
24. Ernest HB, Hoar BR, Well JA, O’Rourke KI (2010) Molecular genalogy tools for white-tailed deer with chronic wasting disease. Can J Vet Res 74: 153–156.
25. Williams ES, Miller MW (2002) Chronic wasting disease in deer and elk in North America. Rev Sci Tech 21: 305–316.
26. Cutler SJ, Fooks AR, van der Poel WHM (2010) Public health threat of new emerging infectious diseases. Nature 451: 990–993.
27. DeYoung RW, Demarais S, Honeycutt RL, Gonzales RA, et al. (2009) Evaluation of white-tailed deer. J Wildl Dis 45: 989–997.
28. Allison AB, Geokjian VH, Postger AC, Wilson WC, Johnson DJ, et al. (2010) Detection of a novel reassortant epizootic hemorrhagic disease virus (EHDV-2) in the USA containing RNA segments derived from both exotic (EHDV-6) and endemic (EHDV-2) serotypes. J Gen Virol 91: 430–439.
29. DeYoung RW, Demarais S, Honeycutt RL, Gee KL (2002) Multiple paternity in white-tailed deer (Odocoileus virginianus) revealed by DNA microsatellites. J Mammal 83: 884–892.
30. Ernst HB, Hoar BR, Well JA, O’Rourke KI (2010) Molecular genalogy tools for white-tailed deer with chronic wasting disease. Can J Vet Res 74: 153–156.
31. Williams ES, Miller MW (2002) Chronic wasting disease in deer and elk in North America. Rev Sci Tech 21: 305–316.
32. O’Rourke KI, Spraker TR, Hames LS, Besser TE, Brayton KE, et al. (2004) Polymorphisms in the prion protein functional gene but not the pseudogene are associated with susceptibility to chronic wasting disease in white-tailed deer. J Gen Virol 85: 1339–1346.
33. Blanchong JA, Haysell DM, Struebour KT, Libants SV, Johnson C, et al. (2009) Genomic susceptibility to chronic wasting disease in free-ranging white-tailed deer: complement component C1q and PRP polymorphisms. Infect Genet Evol 9: 1329–1335.
34. Wilson GA, Nakada SM, Bollinger TK, Pybus MJ, Merrill EH, et al. (2009) Polymorphisms in the PRNP gene influence susceptibility to chronic wasting disease in two species of deer (Odocoileus Spp) in western Canada. J Toxicol Environ Health A 72: 1025–1029.
35. Ashburner D, Pollaro JV, Cowles CR, Eiten WJF, Baldwin J, et al. (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 407: 513–516.
36. Wiedmann RT, Smith TPL, Nonneman DJ (2008) SNP Discovery in ovine by reduced representation and high-throughput pyrosequencing. BMC Genet 9: 81.