ALTERNATIVE APPLICATIONS OF WHOLE GENOME *DE NOVO* ASSEMBLY IN ANIMAL GENOMICS

A Dissertation Presented to the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment of the Requirements

for the Degree

Doctor of Philosophy

---

by

LYNSEY WHITACRE

Dr. Jared E. Decker, Dissertation Advisor

JULY 2017

APPROVAL PAGE

The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled:

ALTERNATIVE APPLICATIONS OF WHOLE GENOME *DE NOVO* ASSEMBLY IN ANIMAL GENOMICS

Presented by Lynsey Whitacre, a candidate for the degree of Doctor of Philosophy, and hereby certify that in their opinion it is worthy of acceptance.

_____

Dr. Jared E. Decker, Animal Sciences, UMC

_____

Dr. Robert D. Schnabel, Animal Sciences, UMC

_____

Dr. Jeremy F. Taylor, Animal Sciences, UMC

_____

Dr. J. Chris Pires, Biological Sciences, UMC

_____

Dr. Gavin C. Conant, Biological Sciences, NC State

ACKNOWLEDGEMENTS

I would like to acknowledge those individuals responsible for distinguishing my doctoral degree as a highly experiential learning experience. Thanks to Dr. Jared Decker, Dr. Jerry Taylor, Dr. Bob Schnabel, Dr. JaeWoo Kim, Dr. Gary Johnson, Dr. Kevin Wells, Dr. Susanta Behura, Dr. J. Chris Pires, Jesse Hoff, and the members of the Mizzou Animal Genomics group and Research Support Computing group for providing a collaborative learning experience and supporting my efforts. Each of these individuals contributed to my learning experiences during my doctorate in a unique manner that has shaped not only my degree, but my future. Thanks to Dr. Chi-Ren Shyu and Robert Sanders of the MU Informatics Institute for their dedication to the program and my success. Thanks to my fellow graduate students Harly Durbin, Troy Rowan, Sara Nilson, Maria Haag, and Cathy Bernhard for their friendship, advice, and support. Furthermore, special thanks to Dr. Gavin Conant (North Carolina State University) and Dr. Mark Wildhaber (United States Geological Survey) for their ongoing collaborations.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

REVIEW OF GENOME SEQUENCING AND *DE NOVO* ASSEMBLY

Introduction to genome sequencing

Genome sequencing is the process by which the sequence of deoxyribonucleic acid (DNA) residues that compromise the genome, or complete set of genetic materials of an organism or individual, is determined. Early genome sequencing efforts sequenced in an ordered fashion and required extensive laboratory work [1]. Next-generation sequencing (NGS) methods employ a different method that requires substantially less laboratory work, but demand great computational resources (Marguiles 2005). NGS is a high-throughput process that begins with fragmenting DNA so it can be sequenced in a massively parallel manner. The result is billions of short DNA sequences that must be put back together, or assembled. These methods can be applied to ribonucleic acid (RNA) after using the RNA to synthesize complementary DNA (cDNA), which is double-stranded, to determine the sequence of the transcriptome.

Down-stream analysis of NGS data requires that short reads be compiled into contiguous sequences either using *de novo* or reference-guided assembly, often referred to as read alignment [2]. For organisms like cattle with a mature reference genome, reads generated in the sequencing process are usually matched to the reference genome with a variety of alignment algorithms [3]. This is currently the most efficient way of transforming raw sequence reads from an individual animal into a whole genome consensus sequence or variant genotypes. However, one blatant limitation of this approach is that genetic differences are only detected with respect to the reference

genome. In addition, the analysis of the sequences is only as good as the reference to which the reads are mapped. This has led many projects to *de novo* genome assembly methods, which do not require any previous knowledge about the genome.

Elements of *de novo* genome assembly

A *de novo* genome assembly aims to create full length sequences of species or individuals that have not previously had their genome assembled. The overarching goal of *de novo* assembly is to overlap sequencing reads to construct contiguous sequences, also known as contigs or scaffolds, representing the genome [4]. Improvements in NGS technology over the last decade have enabled inexpensive deep sequencing of non-model organisms and facilitated many *de novo* genome assembly projects. However, despite significant technological advances in sequence and assembly technology, *de novo* assembly algorithms are often unable to construct long sequences with chromosome-level contiguity [5]. This is largely due to the impediments of graph theory and the inability to routinely determine a unique, correct path through the graph.

The quality of an assembly is often difficult to determine, but is generally measured by the length of the contigs that are assembled using a statistic called N50. The N50 is the length at which all of the contigs of that length or greater contains at least half of the total sum of the contig lengths [6]. By definition, this means that 50% of the assembled genome is found in contigs that are at least as large as the N50 value. One disadvantage of the N50 evaluation method is that this statistic only weights the contiguity of the assembly and does not consider the quality of the assembly or the probability of misassembly. The quality and contiguity of a *de novo* assembly can be affected by many factors including: efficiency of the assembly algorithm, library

construction, read quality and pre-processing, genome architecture, and depth of coverage [7–12].

Several bioinformatics algorithms have been developed for efficient and accurate *de novo* genome assembly. These algorithms are based on two primary methods: de Bruijn graph (DBG) and overlap layout consensus (OLC). The DBG method utilizes the compact representation of *k*-mers (short sequences of length *k*) to construct and traverse through a graph to infer the genome sequence [13]. The OLC method first finds overlaps between all the reads and then creates a graph of these overlaps where the nodes represent the reads and the edges represent overlaps [14]. From the traversal of this graph, the algorithm infers a consensus sequence.

Before assembly algorithms can be applied, each *de novo* genome project must start with the construction of high quality sequencing libraries. Sequencing libraries are constructed from DNA extracted from blood or tissue from the individual of interest [15]. Libraries are generally created by fragmenting the DNA to a specific average length and attaching oligonucleotide adapters to the ends of the fragments [16]. The size of the fragments is crucial for constructing a good library. The size of the library is determined by the average insert size, or the distance between the adapter sequences. Common insert sizes of paired end libraries range from 200 to 700 bp [17], while mate pair libraries generally range from 2 to 5 kilobases (kb) [18,19], and larger protocols, such as fosmid libraries, can hold DNA inserts up to 40 kb [20,21]. The insert size is important because it gives the assembler information about the physical distance between two sequences in the genome.

The most common types of libraries for short read sequencing are paired end and

mate pair. Paired end reads are sequenced from the same strand of DNA, but one read is sequenced in the forward direction and one in the reverse direction. Each read in the pair starts from an end of the fragment and is sequenced toward the center of the fragment. Mate pair reads are also sequenced from the same strand and in the opposite direction, but the sequencing starts in the middle of the fragment and goes outward due to a circularization step in the library preparation [18,19]. As sequencing technologies continue to be developed, there is a shift towards using longer reads to assemble genomes. These technologies can generate reads upwards of 10,000 bp in length, which improves the contiguity of assemblies, especially in repetitive regions of the genome [22,23].

One characteristic of the sequencing process is a probability of error, which is relayed to the user as a quality score. The quality score gives the probability that a base is incorrectly called. In sequencing-by-synthesis data, the quality tends to decrease at the 3' end of the read [24]. Therefore, the ends of the reads are often trimmed before being used. Sequencing errors can also be detected and reduced using *k*-mer frequency metrics where low depth *k*-mers are assumed to be erroneous [5,25]. Reads should also be pre-processed to trim off the adapters that were ligated to the DNA fragments in library construction. Failure to trim adapters will confuse the assembly algorithm leading to contaminated and poorly constructed contigs and will decrease the computational efficiency of the assembler [26].

Genome architecture also plays a large role in determining the quality of an assembly, primarily due to repeats. In organisms with large genomes like mammals, this presents a problem because upwards of 50% of the genome is generally comprised of

repetitive sequence [27,28]. This issue is often even more severe in plant genomes. Repetitive and low-complexity DNA in the genome complicate the assembly of short reads, especially when the size of the repeat is greater than the length of the read [29]. These repeats create ambiguity in the assembly graphs by creating branches or forks. If the assembler follows the incorrect branch, sequences will be falsely joined, creating a chimera. Many chimeras can be detected in post-processing stages of assembly by assessing characteristics such as mate-pair spacing and orientation, read coverage, and read breakpoints [30].

Genome assembly requires each segment of the genome to be sequenced multiple times. It has been demonstrated that an average of 24.2X coverage is required to cover 99.5% of a mammalian genome with at least one NGS read [31]. High depth of coverage is necessary to obtain coverage across the entire genome due to biases in the NGS chemistry. For example, GC rich regions are prone to low coverage with NGS techniques [32]. Increasing the depth of coverage also enhances the quality of an assembly around complex and heterozygous regions and is beneficial in determining whether a heterozygous site is due to actual genetic variation or a sequencing error. Therefore, 24.2X serves as a minimum for *de novo* assembly and coverage of approximately 50X is optimal [33,34].

Applications of *de novo* assembly

Once a high quality *de novo* assembly has been constructed, it is often used to determine variation in the genome or simply as a reference for future resequencing projects. However, the nature of *de novo* assembly algorithms lends the method to contribute information about other characteristics of an individual's genome. The

research presented in this dissertation aims to maximize the use of *de novo* assemblies from both RNA and DNA sequencing to answer questions about animal genomics in the absence of other established genomics or bioinformatics tools designated for that purpose. The following chapters present three publications to address these uses in detail.

In chapter 2, *de novo* assembly methods are used to assemble RNA and DNA reads sequenced from tissues from the bovine reference individual that did not map to the reference assembly. The assembly of these unmapped reads allowed insight into what regions of the genome were absent or misassembled. Particularly, by using both DNA and RNA, several thousand genes that were not correctly assembled in the reference were identified and constructed via *de novo* assembly. In addition, many sequences were found in the unmapped reads from DNA and RNA that were not of cattle origin. Once assembled, many of these sequences could be identified as arising from a particular species and gave insight into the pathogenic and commensal organisms using the cow as their host. Other sequences with a high similarity, but not identical, to known species, but likely representing a previously unsequenced or unidentified organism, were also detected. Based on the conclusions of this research, it is recommended that the unmapped reads of all animals sequenced be assembled and analyzed both to identify and construct sequence not represented in the reference genome and to classify pathogens that may contribute to subclinical illness in the animal [35].

In chapter 3, *de novo* assembly methods are used to help elucidate the genetic cause of a birth defect in water buffalo. Using the properties of assembly by traversal through a graph, contigs that are positive outliers for length were used to determine regions of the genome that were uniform and highly homozygous in affected animals.

Runs of homozygous can be determined by the largest contigs because they are the easiest to assemble. Along with other genomic data, these runs of homozygosity helped determine candidate regions for the causal disease loci [36].

In chapter 4, an extension of the *De Bruijn* graph method simultaneously assembles the genomes of multiple individuals to discover variants in a threatened species with no previous knowledge of genetic composition. The utility of *de novo* assembly is also explored in a more traditional sense in this chapter to assemble the genome of a threatened catfish species, but a novel statistical method is employed to analyze the conservation of genome sequence across closely related species. From these analyses, genes that contribute to important phenotypes can be coordinated with one another, either because they are significantly conserved or significantly diverged [37].

CHAPTER 2

WHAT'S IN YOUR NEXT-GENERATION SEQUENCE DATA? AN EXPLORATION OF UNMAPPED DNA AND RNA SEQUENCE READS FROM THE BOVINE REFERENCE INDIVIDUAL

Lynsey K. Whitacre[1,2], Polyana C. Tizioto[2,3], JaeWoo Kim[2], Tad S. Sonstegard[4,5], Steven G. Schroeder[4], Leeson J. Alexander[6], Juan F. Medrano[7], Robert D. Schnabel[1,2], Jeremy F. Taylor[2,*], and Jared E. Decker[1,2,*]

[1] Informatics Institute, University of Missouri, Columbia, Missouri 65211, USA

[2] Division of Animal Sciences, University of Missouri, Columbia, Missouri 65211, USA

[3] Embrapa Southeast Livestock, São Carlos, São Paulo 13560-970, Brazil

[4] Animal Genomics and Improvement Laboratory, USDA-ARS, Beltsville, Maryland 20705, USA

[5] Recombinetics Inc., 1246 University Ave W #301, St Paul, MN 55104, USA

[6] USDA-ARS (retired), LARRL, Fort Keogh Miles City, Montana 59301, USA

[7] Department of Animal Science, University of California-Davis, Davis, California 95616, USA

*Corresponding authors

Abstract

Background**:** Next-generation sequencing projects commonly commence by aligning reads to a reference genome assembly. While improvements in alignment algorithms and computational hardware have greatly enhanced the efficiency and accuracy of alignments, a significant percentage of reads often remain unmapped.

Results**:** We generated *de novo* assemblies of unmapped reads from the DNA and RNA sequencing of the *Bos taurus* reference individual and identified the closest matching sequence to each contig by alignment to the NCBI non-redundant nucleotide database using BLAST. As expected, many of these contigs represent vertebrate sequence that is absent, incomplete, or misassembled in the UMD3.1 reference assembly. However, numerous additional contigs represent invertebrate species. Most prominent were several species of Spirurid nematodes and a blood-borne parasite, *Babesia bigemina*. These species are either not present in the US or are not known to infect taurine cattle and the reference animal appears to have been host to unsequenced sister species.

Conclusions**:** We demonstrate the importance of exploring unmapped reads to ascertain sequences that are either absent or misassembled in the reference assembly and for detecting sequences indicative of parasitic or commensal organisms.

Keywords**:** DNA sequencing, RNA sequencing, unmapped reads

Background

Next-generation sequencing technology has vastly increased the dimensionality of sequencing projects and routinely allows the generation of hundreds of millions or even billions of short reads. Analysis of these data requires that the short reads be assembled

into contiguous sequences either using *de novo* or reference-guided assembly. For organisms with a reference genome, reads generated in the sequencing process are usually matched to the reference sequence with a variety of alignment algorithms. This is currently the most efficient way of transforming the raw sequence reads into a consensus sequence. However, there are several limitations inherent to the alignment process, including alignment to repetitive regions, absent or misassembled sequence in the reference genome, and individual genetic divergence between the subject organism's genome and the reference genome [38]. Despite these challenges, the majority of reads produced from a sequencing experiment will adequately align to a reference assembly. Nevertheless, a small but significant fraction of reads frequently remain unmapped.

Unmapped reads have generally been disregarded and these data are often discarded. However, recent work has begun to focus on the development of bioinformatic tools for detecting pathogens in human sequence data by the computational subtraction of known human sequences [39–41]. Application of these pipelines in other recent studies has suggested that potentially biologically relevant information can be extracted from the unmapped reads [42,43]. Using an original alignment, assembly, and identification pipeline that can be applied to data from any species, we took advantage of a unique opportunity to explore the unmapped reads from the DNA and RNA sequencing of L1 Dominette 01449, the *Bos taurus* reference individual [44]. These data had not previously been used in the creation or annotation of the reference assembly.

Using sequence data produced from the reference individual, we minimized alignment challenges that are due to genetic variation among individuals. Thus, we expected to encounter meaningful biological information pertaining to sequences poorly

represented in the bovine reference assembly and sequences indicative of parasitic or commensal non-vertebrate organisms. We identified DNA and RNA contigs that were assembled *de novo* from unmapped reads that could generally be classified into one of three categories: 1) sequence from bovine; 2) sequence from other vertebrate species that was homologous to bovine; and 3) sequence from non-vertebrate species. This analysis unequivocally demonstrates that the unmapped reads contain important data pertaining to sequences from the organism that are missing from the reference assembly, represented by categories 1 and 2, and sequences that can be used to identify microbiota members, putatively represented by category 3.

Results

De novo assembly of unmapped reads

Approximately 111.7 million DNA sequence reads, 7.2% of the total, remained unmapped after alignment to the reference genome. A fraction of those reads could be used for assembly, due to a large number of sequences with low quality (https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-015-2313-7/MediaObjects/12864_2015_2313_MOESM1_ESM.xlsx; Supplementary Table 1). However, approximately 1.4 million reads were incorporated into 69,230 contigs with an N50 of 737 bp. Overall, the contigs comprised approximately 46.6 Mb. Additional assembly statistics are provided in Supplementary Table 1 (https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-015-2313-7/MediaObjects/12864_2015_2313_MOESM1_ESM.xlsx).

A median of approximately 6.7% of RNA-seq reads remained unmapped across each of the 17 tissue samples. *De novo* assembly of these reads yielded a total of 43,961 contigs, with a median of 1,792 contigs per tissue and an N50 of 324.5 bp. Overall, the contigs spanned 14.8 Mb with a median of 603 Kb per tissue. Assembly statistics for each tissue are in Supplementary Table 2 (https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-015-2313-7/MediaObjects/12864_2015_2313_MOESM1_ESM.xlsx).

Pairwise alignment of contigs assembled from unmapped DNA reads to the non-redundant nucleotide database

Approximately 51% of the contigs generated from the unmapped DNA reads produced a significant alignment when queried against the non-redundant nucleotide (*nt*) database. The most common alignment was to other *Bos taurus* sequences (Figure 2.1). This result was expected given the draft quality of the bovine reference assembly and considering that we assembled paired reads if either one or both of the reads were unmapped to the reference assembly. However, the second most common alignment for these DNA contigs was to *Onchocerca ochengi*, a nematode known to infect indicine cattle that has been heavily researched due to its similarity to the parasite that causes African River Blindness in humans. We simulated paired-end sequence read data from the *O. ochengi* genome assembly by randomly shearing the genome and then aligned the produced paired-end reads to the bovine reference assembly and concluded that the *O. ochengi* assembly is contaminated with cattle sequences (Supplementary Note 1). Consequently, we excluded the *O. ochengi* assembly from any further analyses.

With subsequent analyses preventing alignment to *O. ochengi*, approximately 44% of the contigs produce a significant alignment against the *nt* database. A fraction of the contigs originally identified as *O. ochengi* were unambiguously matched to bovine sequences. However, the number of alignments to other filarial nematode sequences also increased. These included hundreds of contigs aligned to *Gonglyonema pulchrum* and *Wuchereria bancrofti,* and a few to *Parascaris equorum. G. pulchrum* and *W. bancrofti* belong to the order Spirurida, as does *O. ochengi*, but that are known to only infect humans. The alignments to each of these species had a percent identity of approximately 82% (Table 2.1), which is consistent with cattle not being a host for these nematodes and indicating that these alignments represent sequences from unsequenced sister species of *G. pulchrum* and *W. bancrofti*.

Also detected were sequences with high percent identities to *Babesia bigemina,* a blood-borne parasite known to cause bovine babesiosis, or Texas fever in cattle. While only 190 contigs aligned to *B. bigemina*, significantly less than the combined number of alignments to nematode species, ten were larger than 1,000 bp and the median identity was 91.10% (Table 2.1). A complete summary of significant alignments, both vertebrate and non-vertebrate, is presented in Supplementary Table 3 (https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-015-2313-7/MediaObjects/12864_2015_2313_MOESM1_ESM.xlsx).

Alignments to other vertebrates represent cattle sequences that are not currently well represented in the *Bos taurus* database. Thus, the number of alignments to these organisms is a function of the completeness of the available data for each species and the phylogenetic relationship between the species and cattle. For example, human (*Homo*

sapiens), being the most complete, has the largest number of alignments of the other vertebrate species, followed by pig (*Sus scrofa*), and while bison (*Bison bison bison*) and water buffalo (*Bubalus bubalis*) are more closely related to cattle than human or pig, these bovids have less sequence data available and thus do not produce as many alignments.

Pairwise alignment of contigs assembled from unmapped RNA-seq reads to the non-redundant nucleotide database

The pairwise alignment of the *de novo* assembled contigs generated from the unmapped RNA-seq reads to the *nt* database produced similar results to the alignment of the DNA contigs. Overall, 81% of the RNA-seq contigs had significant alignments to sequences in the *nt* database. Across all tissues, *Bos taurus* produced the most largest number of alignments. Also prominent were alignments to *Bison bison bison, Bubalus bubalis,* and *Bos mutus*, all species that are closely related to cattle (Figure 2.2). Significant BLAST alignments of the RNA-seq unmapped read contigs to cattle or these other closely related species indicates the existence of coding regions that are missing or misassembled in the reference assembly. By mapping the GI number of the most significant BLAST alignment to a gene symbol, we detected alignments to 4,412 *B. taurus* and 4,029 *B. bison bison, B. bubalis*, or *B. mutus* genes. As the total number of *Bos taurus* genes reported by Ensembl is 19,994 [45], this suggests that as many as 42.2% of the bovine protein coding genes are misassembled (although these misassemblies likely represent a small fraction of total transcriptome base pairs). Additionally, approximately 5% of RNA alignments failed to map to a gene with an

assigned symbol, likely corresponding to unannotated structural or regulatory RNAs. Further results and discussion of these analyses are included in Supplementary Note 2. As was the case for the pairwise alignment of unmapped DNA contigs, there were also numerous alignments to other vertebrate and non-vertebrate species. The most common alignments to non-vertebrate species included uncultured bacterium, bovine herpesvirus 6, *Onchocerca flexuosa* and *B. bigemina* (Table 2.2). Bovine herpesvirus 6 was previously discovered as a contaminant in the UMD3.1 build by Merchant *et al*. [46], who concluded that Dominette must have been host to the virus. Alignments to *O. flexuosa* and *B. bigemina* support the hypothesis generated from the analysis of the unmapped DNA read contigs that Dominette was also host to a nematode of the Spirurida order and an unsequenced relative of *B. bigemina*. Several additional fungal and bacterial species were also detected in the unmapped read RNA-seq contigs at low levels. Nearly all of the detected non-vertebrate organisms had alignments from multiple tissues, which would be expected for blood-borne parasites. The number of alignments for each tissue was a function of the total number of sequencing reads from that tissue. A complete summary of alignments from all 17 tissues is presented in Supplementary Tables 4 and 5 (https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-015-2313-7/MediaObjects/12864_2015_2313_MOESM1_ESM.xlsx).

No evidence of horizontal gene transfer

With deep sequencing, it is possible to expose rare horizontal gene transfer events. To address this possibility, we searched for mate-pair reads from the large insert DNA libraries where one mate was uniquely mapped to the cattle reference genome and the other mate mapped uniquely to a non-vertebrate sequence. No such mate-pairs were

identified that met these criteria. Additionally, in our BLAST results we also searched for chimeric contigs (contigs that partially mapped uniquely to cattle and partially mapped uniquely to a non-vertebrate species). Again, no such contigs were identified that met these criteria.

Discussion

To our knowledge, this is the first formal investigation into the nature and identity of unmapped reads from the resequencing of an individual used for the generation of a reference genome assembly. These data allowed us to directly compare reads to the reference assembly without alignment challenges due to genetic variation between the reference and the resequenced genome. Second, the opportunity to compare independently generated datasets from the same individual provided unequivocal support for our discovery of concordant non-vertebrate sequences within the whole genome and transcriptome sequences of the bovine host. In addition to our sequencing of cDNA generated from RNA isolated from 17 tissues, we also sequenced genomic DNA that had been isolated from both liver and white blood cells at three separate facilities. Endogenous contaminants were detected in the reads that were generated from all three sequencing runs. Nearly all of the contigs assembled *de novo* from unmapped reads that were identified as representing a non-vertebrate species were comprised of reads that originated from multiple libraries sequenced at separate facilities. These attributes facilitated both the discovery and validation of the parasitic and commensal species sequences found in this study.

Despite the continuing exponential increases in sequences submitted to NCBI's databases, the number of represented species still comprises only a small proportion of

existing species. While we detected several sequence alignments to spirurid nematodes in both the DNA and RNA sequence data, none of these species are known to be present in the US or to infect taurine cattle. Therefore, we postulate that the actual species present within the tissues of Dominette either represent undiscovered species or a previously recognized, but unsequenced, organism such as *Onchocerca gutturosa* or *Onchocerca lienalis* from the Spirurida order. Both *O. gutturosa* and *O. lienalis* are known to infect taurine cattle in various parts of the United States [47]. However, these species have not been sequenced other than for a few selected genes used to generate data for phylogenetic analyses [48–56]. In this study, we assembled nearly 1,000 contigs that we believe represent novel sequence from a Spirurid nematode that infects taurine cattle in North America.

The precise identity of the species generating the sequence matching *B. bigemina* in both the RNA-seq and genomic DNA data is also ambiguous. As no fever like symptoms were reported in this cow who spent her life at a USDA research facility near Miles City, Montana and babesiosis has been reported to have been eradicated in the United States with vaccination no longer being required [57], we suspect that Dominette was asymptomatically infected with a non-pathogenic strain of *Babesia* spp., as has previously been reported in Turkey [58], Syria [59] and Thailand [60]. Although it is currently not possible to determine the exact species of parasite, we can estimate the animal's parasite burden via deep sequence data by evaluating the number of species to which the contigs of unmapped reads align and the number of contigs that align to each species. Parasite burden negatively impacts animal health and profitability [61,62] and can serve as a reservoir for later infections [59]. Although symptoms were not visible and

the animal appeared healthy, the detection of subclinical parasite burden, even from non-pathogenic parasites, is important because a physiological response to the infection from the host must still occur. This response reduces fitness, causes a decrease in production traits such as feed intake and feed efficiency [62,63] and can also influence the interpretation of RNA-seq experiments.

An alternate explanation for the identification of non-vertebrate sequences in a vertebrate animal is the actual integration of these DNA sequences into the animal's genome. Recently, horizontal gene transfer has been reported to occur at a low level in many animal species [64–66]. It has also been reported that there can be integration of foreign DNA released by dead cells into healthy host cells [67]. However, we were unable to find evidence for the integration of non-vertebrate DNA into this animal's genome and must exclude horizontal gene transfer based on our data.

Conclusions

In conclusion, we alert researchers that many sequences of interest may be found in the reads that fail to align to a reference assembly. We demonstrate that the unmapped reads contain biologically significant information relative to genes that are either partially or completely missing from the reference assembly, as well as information regarding the identity and magnitude of commensal or parasitic organisms. The large number of missing or misassembled bovine protein coding genes must significantly impact the interpretation of RNA-seq studies, warrants further research, and is likely more severe in the less complete reference genomes of other livestock species. Continuation of unmapped read mining will also expand our knowledge of the extent of internal parasitic infections and may lead to the discovery of previously unknown symbiotic relationships.

These metagenomic inferences are an additional source of information from whole-genome sequencing data that can be used as phenotypes or covariates in downstream analyses. As the quality of reference assemblies improves and the scope of sequenced microorganisms broadens, the detection of parasitic infections and other symbiotic relationships will become more explicit.

Methods

Ethics statement

Tissues from L1 Dominette 01449 were sampled according to IACUC No. 081711-1, which was approved by the USDA-ARS Fort Keogh Livestock and Range Animal Care and Use Committee.

DNA and RNA sequencing

DNA was extracted from liver and whole blood samples from L1 Dominette 01449 (referred to here as "Dominette"), a Hereford cow used to generate the *Bos taurus* Sanger reference assembly [44], and was sent to three separate facilities for sequencing. Paired-end and mate-pair libraries were constructed and DNA was 2 x 100 bp sequenced to an average coverage of approximately 55X. Further details regarding the sequencing of each library is provided in Supplementary Table 6 (https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-015-2313-7/MediaObjects/12864_2015_2313_MOESM1_ESM.xlsx).

RNA was extracted using Trizol Reagent (Invitrogen, Carlsbad, CA) as described elsewhere [68] from 17 tissue samples including ampula, blood, cerebral cortex, endometrium sampled from caruncular regions contralateral (car con) and ipsilateral (car

ips) to the corpeus luteum, gallbladder, heart, ileum, infundibulum, jejunum, kidney, liver, mesenteric lymph nodes, pons, ribeye muscle, semitendinosus muscle, and spleen. Preparation of the mRNA samples for sequencing was performed by Global Biologics (Columbia, MO) using the TruSeq Stranded mRNA Library Prep Kit (Illumina®, San Diego, CA) and sequenced 2 x 100 bp using Illumina technology, with the exception of blood which used the TruSeq RNA Sample Preparation Kit and was sequenced 1 x 100 bp.

Pre-processing and alignment of reads

Error correction was performed on DNA sequence reads using the QuorUM error correction algorithm [25]. After filtering duplicate and low quality reads, 1,622,097,087 unique reads remained. Paired reads were aligned to the UMD3.1 cattle reference assembly using NextGENe 2.4.1 (SoftGenetics, LLC, State College, PA) requiring at least 35 contiguous bases with ≥95.0% overall match, up to 2 allowable mismatched bases, and up to 100 allowable alignments of equal probability genome-wide.

RNA sequence reads were filtered for quality and adapter sequences and were then trimmed using a custom Perl script already described [68]. Computations were performed on the HPC resources at the University of Missouri Bioinformatics Consortium (UMBC). TopHat v2.0.6 [69] was used to map the reads to the *Bos taurus* UMD3.1 reference genome. A total of 2 mismatches and up to 3 bp indels were allowed in alignment.

De novo assembly of unmapped reads

Reads from DNA sequencing that remained unmapped following alignment to the reference genome were assembled using MaSuRCA 2.3.2 [70]. Reads from RNA

sequencing that remained unmapped following alignment were assembled using Trinity version r20140717 [71]. To maintain a paired read file structure, reads where both the forward and reverse read were unmapped or where one of the reads was unmapped but the other was mapped were collectively used for assembly.

Pairwise alignment of unmapped contigs to the nt database

Prior to pairwise alignment, contigs assembled from the unmapped DNA reads were sorted by size and only contigs greater than 500 bases were aligned (n = 42,086). Due to the smaller size of the RNA contigs, they were not filtered by size prior to pairwise alignment. Using the blastn algorithm of BLAST+ 2.2.30 [72,73], each DNA and RNA contig was aligned to the NCBI non-redundant nucleotide database and the most significant alignment was returned. The BLAST output was then parsed to determine the subject species, percent identity, length of match, number of mismatches, number of gaps, E-value, and overall score. Significant alignments were declared only if the length of the alignment was ≥150 bp for DNA or ≥50 bp for RNA. Only the best match for each aligned contig was reported. This output was summarized according to the total number of alignments per species, mean, median, and maximum percent identity, mean, median, and maximum length of match, and mean and median e-value (https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-015-2313-7/MediaObjects/12864_2015_2313_MOESM1_ESM.xlsx; Supplementary Tables 3 and 5).

Quantification and identification of coding regions within unmapped reads

Contigs from unmapped RNA-seq reads were aligned to contigs from unmapped DNA reads using NextGENe 2.4.1 requiring ≥98% overall match to declare a match.

Additionally, for the significant RNA alignments, the gene symbol corresponding to the GI accession number for the alignment was captured where possible and recorded using the db2db tool in bioDBnet [74]. A unique list of gene symbols was constructed and the number of significant alignments to each gene was tallied (https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-015-2313-7/MediaObjects/12864_2015_2313_MOESM1_ESM.xlsx; Supplementary Tables 7 and 8).

Availability of Data and Materials

The data set supporting the results of this article is available in the SRA repository, SRA accessions SRX1177177 through SRX1177278.

Competing interests

J.F.T. is on the scientific advisory boards (SABs) of Recombinetics, Inc and Neogen Corporation.

Author contributions

Author contributions are as follows: J.F.T., J.E.D., R.D.S., and L.K.W. designed the experiments and interpreted the results of all analyses. L.K.W. built the analysis pipeline and analyzed the sequence data. P.C.T. did alignments and *de novo* assemblies for RNA sequence data. L.J.A. collected tissues and extracted nucleic acids. J.W.K. extracted and quantitated RNA. T.S.S., S.G.S. and J.F.M. sequenced genomic DNA. J.F.T., J.E.D. and R.D.S. sequenced DNA and RNA. L.K.W. wrote the manuscript and J.F.T., J.E.D. and R.D.S. edited the manuscript. All authors read the final manuscript.

Acknowledgments

Funding for this study was provided in part from the bovine species coordinators of the USDA National Institute of Food and Agriculture supported NRSP-8 National Animal Genome Research Support Program and National Research Initiative Competitive Grants numbers 2011-68004-30214, 2011-68004-30367, 2012-67012-19743, 2013-68004-20364, 2015-67015-23183, MO-HAAS0027, and MO-MSAS0014 from the USDA National Institute of Food and Agriculture. The authors appreciate the contributions of the Beijing Genomics Institute in generating whole genome and tissue transcriptome sequence from the bovine reference animal.

Figures

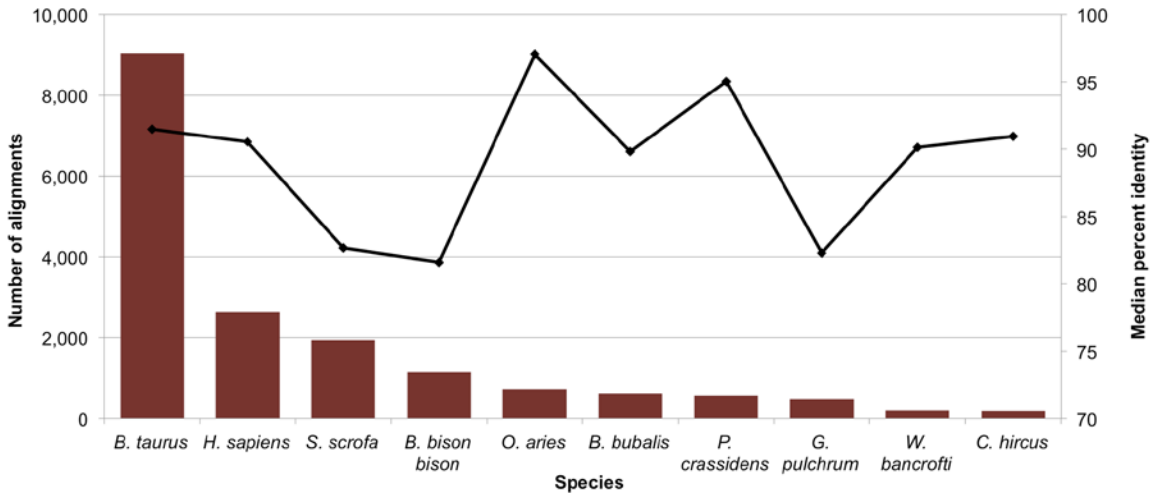

Figure 2.1. Most common alignments from DNA. Bar chart of the ten most common species with significant alignments from the pairwise alignment of *de novo* assembled contigs from unmapped DNA reads. Trend line represents the median percent identity for each species.

[Fig. 1 in publication.]

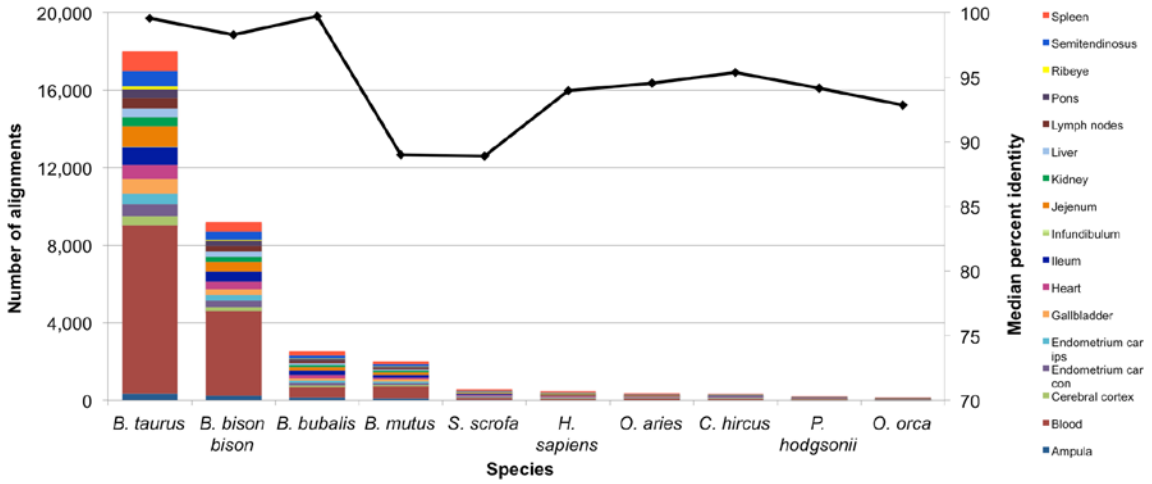

Figure 2.2. Most common alignments from RNA. Bar chart of the ten most common species with significant alignments from the pairwise alignment of *de novo* assembled contigs from unmapped RNA-seq reads by tissue. Trend line represents the overall median percent identity for each species across tissues.

[Fig. 2 in publication.]

Tables

Table 2.1. Top four non-vertebrate alignments to *de novo* assembled contigs from unmapped DNA sequence reads.

Species	No. of Alignments	Median Identity (%)	Max. Identity (%)	Median Length (bp)	Max. Length (bp)	Median E-Value
Gongylonema pulchrum	516	82.33	100	641.0	1,008	2.00E-134
Wuchereria bancrofti	273	81.35	98.82	640.0	1,607	1.53E-143
Babesia bigemina	206	91.10	100.00	505.0	2,078	3.50E-179
Parascaris equorum	11	81.17	100	206.0	1,008	4.00E-41

[Table 1 in publication.]

Table 2.2. Top four non-vertebrate alignments to *de novo* assembled contigs from unmapped RNA-seq reads.

Species	No. of Alignments	Median Identity (%)	Max. Identity (%)	Median Length (bp)	Max. Length (bp)	Median E-Value
Uncultured bacterium	34	97.11	100.00	274.0	1,381	2.01E-116
Bovine herpesvirus 6	22	99.18	100.00	294.5	933	1.00E-143
Onchocerca flexuosa	13	87.74	89.12	224.0	510	3.00E-57
Babesia bigemina	12	94.35	99.51	379.5	926	1.00E-151

[Table 2 in publication.]

Supplementary Material

Supplementary Note 1: The Onchocerca ochengi reference assembly is contaminated with bovine genomic sequence.

After detecting a significant number of alignments to *O. ochengi* and *B. bigemina*, both known to be bovine endosymbionts, we generated pseudo paired-end read data from the *O. ochengi* (GenBank accession number GCA_000950515.1) and *B. bigemina* (GenBank accession number GCA_000723445.1) reference assemblies deposited in NCBI's assembly database. One hundred bp paired-end reads were generated with a random library DNA fragment size distribution between 200 and 500 bases (mean of 350 bases) by stepping through the genome in 1 bp steps. A total of 100X coverage of each genome was simulated. The same procedure was used to generate pseudo reads for five other genomes that were not detected in the unmapped read contigs and also for *O. volvulus*, the human counterpart of *O. ochengi* (GCA_000499405.1, GCA_000001215.4, GCA_000002985.3, GCA_000469785.1, GCA_000612645.1, and GCA_000816705.1).

The alignment of the *O. ochengi* pseudo reads to the bovine reference genome produced significantly more mapped reads than did the alignments for the other species, including *O. volvulus*, a very closely related species (https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-015-2313-7/MediaObjects/12864_2015_2313_MOESM1_ESM.xlsx; Supplementary Table 9). Read mapping rates were significantly different (X^2= 16,714,393, df = 7, p-value ≈ 0) and read mapping was significantly different for *O. ochengi* (X^2 = 16,417,090, df = 1, p-value ≈ 0). Less than 0.004% of the *O. ochengi* reads that were aligned to the cattle reference assembly could be aligned to *O. volvulus*, while over 97% of the *O. ochengi*

reads that did not align to the cattle assembly could be aligned to *O. volvulus*. Furthermore, when the aligned reads were individually aligned to the *nt* database the most significant alignment was always to *O. ochengi*, but all other significant alignments were to *Bos taurus* or other mammals. The *O. ochengi* sample originated from cattle skin, and when we downloaded the raw FASTQ files used to generate this assembly and aligned them to the cow reference, over 33% of the reads aligned. While the majority of the cattle sequences were filtered from the dataset, the *O. ochengi* genome assembly contains contamination from cattle sequence. Therefore, we excluded this genome from further BLAST searches because we could not distinguish alignments to this genome being due to the presence of nematode sequence in our data or cattle sequence in the nematode assembly. This resulted in an increase in the number of alignments to other nematodes from the order Spirurida. All species that were subsequently detected are only known to infect humans. Thus, they were presumably detected due to the presence of a genetically similar, but unsequenced, nematode(s) in our sample.

Supplementary Note 2: Estimation of the number of protein coding genes missing or misassembled in the UMD3.1 bovine reference assembly.

To arrive at an estimate of the quantity and types of genes that may be missing or misassembled in the UMD3.1 reference assembly, we mapped the GI number from the most significant BLAST alignment of each RNA contig to a gene symbol. Based on this mapping, we estimate that 4,412 annotated bovine genes were represented in the 17,856 alignments to *Bos taurus* sequences. A complete list of the affected genes can be found in (https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-015-2313-7/MediaObjects/12864_2015_2313_MOESM1_ESM.xlsx; Supplementary Table 7). Of

the 13,769 significant alignments to bison (*Bison bison bison*), water buffalo (*Bubalis bubalus*), or yak (*Bos mutus*), 4,029 genes were represented. Only one gene was found in common between the 4,412 detected from alignments to *Bos taurus* sequences and the 4,029 detected from alignments to bison, water buffalo, or yak. A complete list of these genes is presented in (https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-015-2313-7/MediaObjects/12864_2015_2313_MOESM1_ESM.xlsx; Supplementary Table 8). We also aligned the RNA-seq contigs to our set of DNA contigs and found that approximately 21% aligned with greater than 98% identity. Of the 46,690,692 bases represented in the DNA contigs, the RNA contigs covered 1,232,501 bases. This is equivalent to 2.6% of the bases which is slightly greater than the genome-wide estimate of coding regions, which has been estimated to be less than 2% of the genome in mammals [75,76].

Based on these analyses, it is apparent that there is a large amount of coding sequence that is either absent or misassembled in the UMD3.1 bovine reference assembly, resulting in the matching of these unmapped contigs to other vertebrate or *Bos taurus* sequences. The extent to which the bovine coding sequences are affected by misassembly is difficult to determine. Florea *et al.* indicated that the UMD3 assembly, although generally superior, does not contain 660 genes that are present in UMD2 [77]. We detected significant alignments to 8,440 unique genes suggesting that as much as 42% of the bovine protein coding genes (http://useast.ensembl.org/Bos_taurus/Info/Annotation?redirect=no) are misassembled. In terms of total nucleotides, we assembled approximately 9.5 million bases of unmapped

reads from RNA sequencing data that aligned to cattle, bison, water buffalo, or yak. Given that the coding regions represent around 2% of the genome, the total number of coding nucleotides is approximately 53 million.

Based on the lengths of the transcripts assembled (median N50 of 334 bp across tissues), we predict that many of these may be minor misassemblies, such as a missing exon. However, transcripts as large as 6,591 bp were also assembled from the unmapped reads, indicating that there are large portions or even entire genes missing as well. Clearly, improvement of the bovine reference assembly is essential to correctly interpret genome resequencing and RNA-seq data and maximize progress from sequencing efforts. This conclusion has now been reached by several research efforts [78–83], and new long read sequencing technology will likely play a significant role in reference assembly improvement [23].

CHAPTER 3

ELUCIDATING THE GENETIC BASIS OF AN OLIGOGENIC BIRTH DEFECT USING WHOLE GENOME SEQUENCE DATA IN A NON-MODEL ORGANISM, *BUBALUS BUBALIS*

Lynsey K. Whitacre[1,2], Jesse L. Hoff[2], Robert D. Schnabel[1,2], Sara Albarella[3], Francesca Ciotola[3], Vincenzo Peretti[3], Francesco Strozzi[4], Chiara Ferrandi[4], Luigi Rammuno[5], Tad S. Sonstegard[6], John L. Williams[7*], Jeremy F. Taylor[2*], and Jared E. Decker[1,2*]

[1]Informatics Institute, University of Missouri, Columbia, Missouri, USA

[2]Division of Animal Sciences, University of Missouri, Columbia, Missouri, USA

[3]Department of Veterinary Medicine and Animal Production, University of Naples Federico II, Naples, Italy

[4]Parco Tecnologico Padano, Lodi, Italy

[5]Department of Agriculture, University of Naples Federico II, Portici, Napoli, Italy

[6]Recombinetics, St. Paul, Minnesota, USA

[7]Davies Research Centre, School of Animal and Veterinary Sciences, University of Adelaide, Roseworthy, Australia

*Corresponding authors

Abstract

Recent strong selection for dairy traits in water buffalo has been associated with higher levels of inbreeding, leading to an increase in the prevalence of genetic diseases such as transverse hemimelia (TH), a congenital developmental abnormality characterized by the absence of a variable distal portion of the hindlimbs. The limited genomic resources available for water buffalo required an original approach to identify genetic variants associated with the disease. The genomes of 4 bilateral and 7 unilateral affected cases and 14 controls were sequenced. A concordance analysis of SNPs and INDELs requiring homozygosity unique to all unilateral and bilateral cases revealed two genes, *WNT7A* and *SMARCA4*, known to play a role in embryonic hindlimb development. Additionally, SNP alleles in *NOTCH1* and *RARB* were homozygous exclusively in the bilateral cases, suggesting an oligogenic mode of inheritance. Homozygosity mapping by whole genome *de novo* assembly also supported oligogenic inheritance; implicating 13 genes involved in aberrant hindlimb development in bilateral cases and 11 in unilateral cases. A genome-wide association study (GWAS) predicted additional modifier genes. Although our data show that the inheritance of TH is complex, we predict that homozygous variants in *WNT7A* and *SMARCA4* are necessary for the expression of TH and selection against these variants should eradicate TH.

Introduction

Water buffalo were domesticated approximately 5,000 years ago on the Indian subcontinent [84]. Today, there are over 130 million domesticated water buffalo worldwide that serve as an important component of agriculture through both milk and meat production [85]. In many developing countries, water buffalo account for more than 50% of the milk production and are relied upon more than any other domesticated species [86,87]. Recently, a genetic disease called transverse hemimelia (TH) has appeared in Italian Mediterranean River buffalo, most likely as an indirect result of strong selection for dairy production traits and an accompanying increase in the rate of inbreeding. Transverse hemimelia causes unilateral or bilateral hindlimb malformation and is defined by the lack of development of distal hindlimb structures, which manifests as the loss of one or both hindlimbs at a distal point that is variable among cases [88] (Figure 3.1A, 3.1B) . It was first reported in water buffalo in 2008 after the conclusion of a study of buffalo with limb malformations from farms in the southern Italian region of Campania [89]. In severe cases with bilateral hindlimb malformation, one or both forelimbs may also be affected. Involved limbs appear as to be amputated with the exception of the fact that there are sketches of claws in the terminal part [90]. The prevalence of the disease has been estimated to be between two and five percent in some populations of Mediterranean Italian River buffalo [91]. Unfortunately, because record keeping is poor and pedigrees are often unknown or incomplete, the mode of inheritance of TH in water buffalo has not been established.

Other types of hemimelia, a generalized developmental anomaly resulting in the absence of the distal portions of one or more limbs, have been reported in domestic

species including goats, lambs, cattle, dogs, and cats [92–96]. In goats, lambs, and cattle, hemimelia has been shown to be heritable, but it can also be caused by environmental exposures to teratogenic plants, parasites, and drugs [97]. The increase in hemimelia in livestock species has been blamed on high levels of inbreeding due to selection for economically valuable traits resulting in increased homozygosity of recessive deleterious mutations. Pedigree analyses of dogs and Shorthorn cattle have revealed hemimelia to be inherited as an autosomal recessive disorder [94,95]. Hemimelia has also been reported in humans, but occurs either due to autosomal recessive inheritance or sporadically, suggesting a polygenic mode of inheritance [98].

Despite several recent research efforts to elucidate the molecular mechanisms involved in hemimelia, the causal mutations in water buffalo and many other species are currently unknown. However, the genetic mechanisms responsible for embryonic hindlimb morphogenesis have been extensively studied in model species and over 30 genes have been implicated in hindlimb development [99–102]. Furthermore, many genes involved in embryonic morphogenesis have been suggested to play roles in the manifestation and inheritance of TH and could be candidates for the causal mutation [103,104]. To elucidate the genes involved in the inheritance of TH in water buffaloes, we sequenced 11 affected buffaloes (4 with bilateral TH and 7 with unilateral TH) and obtained sequences for 14 control buffaloes from the International Water Buffalo Genome Consortium. Our analyses of these data suggest an oligogenic inheritance pattern, and implicate variants in *SMARCA4* and *WNT7A* as the main drivers necessary for the manifestation of TH. The accumulation of homozygous mutations in modifier genes appears to impact the severity of the TH phenotype, resulting in animals that vary

from the lack of a single transverse bone in one limb to the complete lack of both hindlimbs with malformation of one forelimb. The analyses leading to these conclusions describe a novel method for detecting the loci underlying an oligogenic disease in a non-model organism that lacks refined genomic resources such as a completed reference genome or annotated gene models.

Results

Alignment and variant calling

Alignment of DNA sequences from 11 cases and 14 controls to the UMD_CASPUR_WB_2.0 water buffalo reference assembly resulted in an average mapping rate of 99.17% and average coverage of 9.15X (http://www.nature.com/article-assets/npg/srep/2017/170103/srep39719/extref/srep39719-s1.pdf; Table S1). Initial variant calling identified approximately 21.7 million SNPs and 2.8 million INDELs. After filtering on quality, 19.8 million SNPs and 2.7 million INDELs remained for analysis. The overall genotype call rate was 98.02% in the cases and 90.99% in the controls; consistent with the reduced depth of sequence coverage and older Illumina chemistry version used to sequence the controls (http://www.nature.com/article-assets/npg/srep/2017/170103/srep39719/extref/srep39719-s1.pdf; Table S2).

Case versus control concordance analysis

SNP and INDEL concordance analyses were performed to identify variants for which all cases were homozygous for an allele that was never homozygous in the controls. In total, 1,741 SNPs and 793 INDELs met this criterion (http://www.nature.com/article-

assets/npg/srep/2017/170103/srep39719/extref/srep39719-s2.xls; http://www.nature.com/article-assets/npg/srep/2017/170103/srep39719/extref/srep39719-s3.xls). Nine hundred seventy-one of the SNPs were not in genes, but the 770 remaining SNPs were located in 451 unique genes. Two of the genes, *SMARCA4* and *WNT7A*, were associated with the GO term "embryonic hindlimb morphogenesis" (GO:0035116). Furthermore, when only the bilaterally affected cases were analyzed for SNP concordance two additional genes, *NOTCH1* and *RARB*, which are also associated with embryonic hindlimb morphogenesis, were detected. When only the three most severely affected bilateral cases, with the complete loss of both hindlimbs, were analyzed one additional hindlimb morphogenesis related gene, *TFAP2B*, was detected. These findings, along with the fact that no additional associated genes were detected when the unilaterally affected cases were analyzed, present the first genomic evidence for an oligogenic mode of inheritance for TH in water buffalo. However, by aligning the available cattle gene models for these genes to the water buffalo genome assembly, we predicted that all of the disease-associated mutations were located in introns.

Despite the challenge of calling INDEL genotypes with high accuracy from low coverage sequence data, we also analyzed the detected INDELs for their concordance with TH phenotype. Of the 793 concordant INDELs, 222 were located in genes, but only one was found in a gene associated with hindlimb morphogenesis. This INDEL was found in *WNT7A*, a gene also identified by the SNP analysis, and occurs as a 3 bp insertion (C -> CCCG). Based on aligning to the *WNT7A* gene annotation in bovine, this variant appears to be located in intron 3. Unlike the concordance analysis performed for SNPs, no additional concordant INDELs were detected as the cases were further

restricted according to the severity of the TH phenotype. We interpret the results of the INDEL analysis cautiously because, even after filtering for quality, a large proportion of the remaining INDELs appear to have been identified because of homopolymer repeat errors.

Homozygosity mapping by de novo assembly

Large runs of homozygosity (ROH) are common in inbred animals, which have an increased risk for genetic diseases because the deleterious effects of recessive alleles are expressed when they are found in a homozygous state [105]. However, with the exception of selective sweep regions that affect all animals, runs of homozygosity should not be shared across large numbers of unrelated individuals. Thus, homozygosity mapping is a powerful method to identify loci responsible for autosomal recessive traits [106]. Assuming a common origin for all cases, ROH should capture the loci that cause TH in distantly related affected animals (see Methods). However, because the water buffalo reference assembly currently exists as an early draft with more than 367,000 unplaced sequence scaffolds, we used a novel approach for homozygosity mapping that was not limited by the reference assembly scaffold lengths.

Three whole genome *de novo* assemblies were performed from the sequence data, which were pooled separately from four bilaterally affected cases, four unilaterally affected cases, and four controls. Regions of the genome with lower heterozygosity in sequence reads pooled from multiple individuals should be assembled into longer contigs, due to reducing forking in the assembly graph. Overall, the contig N50 statistics achieved for the bilaterally affected and unilaterally affected cases were much larger than for the controls (http://www.nature.com/article-

assets/npg/srep/2017/170103/srep39719/extref/srep39719-s1.pdf; Table S3) and contained contigs that were approximately 0.75 orders of magnitude larger than those assembled for the controls (Figure 3.2). We were also able to assemble nearly the entire water buffalo genome (~2.64 Gb) in 218,053 contigs in the bilaterally affected cases and in 221,020 contigs in the unilaterally affected cases, both significantly fewer than for the current reference assembly, compared to 541,203 contigs for the controls. Estimated from a negative binomial generalized linear model, the mean contig lengths of the bilateral (12,140.04 bp) and unilateral (11,957.47 bp) assemblies were significantly longer than the mean contig length from the assembly of the control samples (4,738.88 bp), which is most likely due to the higher coverage for these samples (http://www.nature.com/article-assets/npg/srep/2017/170103/srep39719/extref/srep39719-s1.pdf; Table S3). Further, the mean contig length from the assembly of the bilateral cases was significantly longer than from the assembly of the unilateral cases (Z-score = -4.08, *p*-value = 4.6e-05). Overall, these results clearly indicate an increase in genome-wide homozygosity in the TH affected buffaloes.

We annotated the gene content of the contigs from each assembly that were significantly larger than average by aligning them to the water buffalo reference genome. Following false discovery rate (FDR) correction with $q < 0.10$, the assembly produced for the controls had 354 contigs significantly larger than the average, the assembly for the unilateral cases had 194 contigs significantly larger than average, and the assembly for the bilateral cases had 365 contigs significantly larger than average. The large contigs identified following FDR correction contained 5, 2, and 0 hindlimb morphogenesis genes for the bilateral cases, the unilateral cases and the controls, respectively

(http://www.nature.com/article-assets/npg/srep/2017/170103/srep39719/extref/srep39719-s4.xls). None of these genes were in common with those detected from the SNP and INDEL concordance analyses. However, homozygosity mapping by *de novo* assembly is impacted by the presence of repetitive elements in the genome that are larger than the sequencing library fragment size and terminate contig assembly. The distribution of these elements in the water buffalo genome is unknown as the reference assembly is not of high quality. Therefore, the effect of repetitive elements on disrupting the assembly of long contigs could not be assessed. To compensate for this and include regions that may be largely homozygous but poorly assembled, contigs in the 99th percentile for size from each of the assemblies were also aligned to the water buffalo reference genome to assess their gene content. Analysis of the longest 1% of contigs assembled for the controls, unilaterally and bilaterally affected cases revealed 2, 11, and 13 genes associated with hindlimb morphogenesis, respectively (http://www.nature.com/article-assets/npg/srep/2017/170103/srep39719/extref/srep39719-s5.xls). Six of these genes – *SMARCA4, NOTCH1, CHD7, MSX1, SALL1,* and *TBX3* – were detected in both the bilaterally affected and unilaterally affected cases. The *SMARCA4* locus, which was also identified in the SNP and INDEL analyses, was of particular interest because this gene and its flanking regions were assembled into a single contiguous sequence approximately 140 kb in length in both the bilaterally and unilaterally affected cases, but in the controls was placed on over 40 disjoint contigs (Figure 3.3).

Genome-wide association study

Association analyses using both binary and semi-quantitative phenotypes were used to discover additional candidate loci (Figure 3.4). Binary phenotypes were simply coded as case *versus* control. Semi-quantitative phenotypes were calculated for each animal based on the number of major distal bones present in each limb ranging from 0 (complete loss of both hindlimbs) to 10 (unaffected control) (http://www.nature.com/article-assets/npg/srep/2017/170103/srep39719/extref/srep39719-s1.pdf; Table S1). While the binary trait GWAS primarily identified regions on small contigs with no nearby genes, markers near *CHAMP1* and three uncharacterized predicted coding regions exceeded the genome-wide significance threshold (Bonferroni correction) (http://www.nature.com/article-assets/npg/srep/2017/170103/srep39719/extref/srep39719-s1.pdf; Table S4). *CHAMP1* was also detected by the homozygosity mapping analysis and was on a contig significantly larger than average in both the bilateral and unilateral cases, however, no concordant SNPs or INDELs were identified. GWAS using semi-quantitative phenotypes revealed 15 additional significantly associated genes. These included *FZD4*, a Wnt receptor, and *FGFR1*, a fibroblast growth factor receptor (http://www.nature.com/article-assets/npg/srep/2017/170103/srep39719/extref/srep39719-s1.pdf; Table S5). The GWAS results also suggested an oligogenic mode of inheritance because numerous loci rose to the same level of significance, in contrast to a GWAS for a Mendelian trait, where one primary peak would be expected in a large sample case *versus* control analysis.

Although the GWAS failed to identify any genes related to hindlimb morphogenesis, several potential modifier genes were detected. This may be due to the nature of the GWAS, which assumed an additive model underlying the severity (expressivity) of the phenotype. The concordance and homozygosity mapping analyses, however, suggested that the phenotype is influenced by epistatic interactions among driver and modifier genes. A further limitation of the GWAS was that SNPs with one or more missing genotypes were either ignored by the analysis algorithm or had association effects estimated by assigning the mean genotype (2 × allele frequency) to missing genotypes. This was particularly problematic here, because of the higher rate of missing genotypes in the controls *versus* cases, due to the lower sequencing depth. Consequently, we filtered results for loci with one or more missing genotypes, resulting in only about 15% of the loci being analyzed for association with the TH phenotypes (http://www.nature.com/article-assets/npg/srep/2017/170103/srep39719/extref/srep39719-s1.pdf; Table S6).

Although several loci rose to genome-wide significance, we recognize that a larger sample size could provide greater power to detect associated variants. However, it is difficult to estimate the number of samples required to achieve a predetermined power when the trait is oligogenic or polygenic, because the power calculation requires the number of causal loci to be known beforehand. While the mode of inheritance of TH in Italian Mediterranean River buffalo was initially unknown, it was suspected to be a fully penetrant Mendelian defect. With a disease prevalence of ~4%, 11 cases are sufficient to detect a recessive Mendelian disease locus at a significance level of 0.001 with 80% power [107]. While future research will require obtaining and evaluating additional data

from affected animals and their parents, we analyzed sequence data for every available case.

Candidate region mapping

To overcome the challenge of harmonizing significant associations with TH from the variety of performed analyses considering that the water buffalo scaffolds are not assigned to chromosomes, we mapped all of the buffalo genomic regions containing candidate loci to the UMD3.1 bovine reference assembly. While this identified candidate regions on all 29 bovine chromosomes, several chromosomal regions were identified as being significantly associated with TH in all of the performed analyses (Figure 3.5). This again suggests an oligogenic mode of inheritance since loci involved in determining TH are spread throughout the buffalo genome and are not concentrated in a single region as would be expected if TH was inherited as a simple Mendelian.

Rudimentary gene enrichment analyses from the mapping of candidate regions to the bovine reference genome also indicated oligogenicity. From the SNP concordance mapping, a total of 769 candidate genes were discovered. This corresponds to approximately 3.85% of the total number of annotated genes in the genome (n = 19,994; http://useast.ensembl.org/Bos_taurus/Info/Annotation?redirect=no) but 6.06% of the total number of hindlimb morphogenesis genes (n = 31; GO:0035137). When only the bilaterally affected cases were analyzed for SNP concordance, two additional hindlimb morphogenesis genes were detected, but when only the unilaterally affected cases were analyzed no additional genes were associated with the GO term hindlimb morphogenesis. Similarly, from the homozygosity mapping by *de novo* assembly analyses, contigs in the 99th percentile for size from the unilaterally affected cases contained 20.35% of all

annotated bovine genes but 33.33% of all hindlimb morphogenesis genes. Large, homozygous contigs assembled from the pooled sequences from the bilaterally affected cases contained 21.36% of all annotated bovine genes but 39.39% of all hindlimb morphogenesis genes. Contigs assembled from the sequences for controls contained 9.70% of all annotated bovine genes but only 6.06% of all hindlimb morphogenesis genes. These results consistently demonstrate an enrichment of hindlimb morphogenesis genes in the larger contigs assembled for the cases compared to the controls with a greater enrichment in the bilaterally versus unilaterally affected cases. They also validate the oligogenic mode of inheritance of TH in Italian Mediterranean River buffalo.

Gene ontology enrichment

Taking into account all 3,988 loci identified as candidates for TH by the various performed analyses, we queried all GO terms to determine sets which may be enriched. The GO term embryonic hindlimb morphogenesis was not significantly enriched (p-value < 0.05). However, the collective list of genes identified by SNP concordance, GWAS, and homozygosity mapping was significantly enriched for embryo development (GO:0009795; p-value = 0.0163), developmental processes (GO:0032502; p-value = 7.04E-9), and anatomical structure development (GO:0048856; p-value = 1.57E-9) among others (http://www.nature.com/article-assets/npg/srep/2017/170103/srep39719/extref/srep39719-s1.pdf; Table S7).

Network analysis

Network analysis of all hindlimb morphogenesis genes (n = 16) detected by the SNP concordance and homozygosity mapping analyses and all potential modifier genes identified by the GWAS (n = 15) was performed to understand how these genes may

interact. Of these 31 genes, 23 formed an exclusive network based on functional association data including genetic interactions, pathways, co-expression, co-localization, and protein domain similarity (Figure 3.6). In this network, *SMARCA4* is directly associated with 12 other genes while *WNT7A* is directly associated with 13 other genes. The remaining 8 genes were not directly associated with any of the other detected genes, and they did not have biological functions that were consistent with the disease phenotype.

Discussion

We used several tactics to identify the genes involved in TH, a congenital limb abnormality resulting in the loss of transverse elements of the hindlimbs in Italian Mediterranean River buffalo. Although the inheritance pattern of TH was initially unknown, we present evidence for an oligogenic mode of inheritance and identify two primary driver genes and several modifier genes. While mutations in both of the driver genes, *SMARCA4* and *WNT7A*, appear to be necessary for the disease, mutations in the modifier genes contribute to the severity of the expressed phenotype. The *SMARCA4* chromatin remodeling factor is required for normal embryonic development and *SMARCA4* knockouts are lethal [108,109]. However, *SMARCA4* expression knockdowns in mice have a large effect on embryonic hindlimb and tail development [102]. These knockdowns produce a phenotype that is very similar to that of the TH affected water buffalo, where the development of the rest of the body and forelimbs is normal and the embryo is viable, but the hindlimbs are extremely underdeveloped.

Several signaling pathways are also required for normal hindlimb development. For example, Wnt signaling has been recognized as important for multiple aspects of

mammalian embryonic development [110], and mutations in *WNT7A* have been reported in hindlimb malformation studies in human and mouse [100,111,112]. In this research, the identification of two of the three known retinoic acid receptors (*RARG* and *RARB*) from the homozygosity mapping and SNP concordance analyses suggests that the retinoic acid signaling pathways also play a role in embryonic hindlimb development and, subsequently, TH. The role of retinoic acid in limb development is controversial: while previous research has reported an association between hindlimb development and retinoic acid levels [113–115], a recent study found that hindlimb budding and patterning do not explicitly require retinoic acid signaling [116]. Our data support a role for retinoic acid receptor genes in hindlimb development as the severity of the TH phenotype appears to increase when mutations are also present in these modifier genes. The data also support a role for Notch signaling, as *NOTCH1* was detected both in the SNP concordance analysis and the *de novo* assembly homozygosity mapping. *NOTCH1* and its ligand, *JAG2*, have repeatedly been implicated in hindlimb development [117,118].

We predict that modifier genes interact with *SMARCA4* and *WNT7A* and underlie the oligogenic inheritance pattern and subsequent variable phenotype. This hypothesis is supported by the structure of the network that was generated from the genes identified from the concordance mapping, GWAS, and homozygosity mapping analyses. However, the complex mode of inheritance of TH and the limited data available on both TH affected and control buffaloes preclude identification of the causal mutations, and the molecular mechanism by which the involved genes regulate hindlimb development remains unclear. The variants identified here in *SMARCA4* and *WNT7A* appear necessary for the expression of TH, but are located in introns based on computational predictions. It

is possible that these variants are in complete linkage disequilibrium with other variants on the same haplotype that are causal but were not detected in this study, either due to low sequence coverage or gaps in the reference assembly. For example, there are four gaps in *SMARCA4* in the current buffalo reference assembly and there is a large gap just upstream of the gene. The *WNT7A* gene is more completely assembled, but also has one gap within the gene and three gaps in the upstream region. These gaps may contain genomic variants that alter the protein encoded by each gene or that alter gene expression through enhancers, promoters, or transcription factor binding sites. Likewise, the identified intronic variants may themselves disrupt unidentified regulatory elements. Variation in regulatory elements likely contributes to the expression of TH as the down-regulation of the expression of *SMARCA4* produces a similar phenotype in mice [102].

We predict that selection against the variants found to be homozygous in *SMARCA4* and *WNT7A* in all cases but not in the controls and the avoidance of mating carriers of these variants would quickly eradicate the disease. This hypothesis could be tested by the targeted genotyping of these loci in buffalo affected by TH and their parents to confirm that the loci are reliably predictive of the disease phenotype. The molecular dissection of the effects of mutations in *SMARCA4* and *WNT7A* could then be performed. However, this will likely require the collection of tissues from developing fetuses and possibly also significant improvements in the water buffalo reference genome and its annotation. Nevertheless, we hypothesize that eradication of the disease is now possible by selecting against the disease associated alleles for any of the concordant SNPs detected in *SMARCA4* and/or *WNT7A*.

Methods

Sample collection

DNA was collected from 4 bilaterally affected and 7 unilaterally affected TH cases and sequenced on an Illumina HiSeq 2500 at the Parco Tecnologico Padano in Milan, Italy. DNA sequences from 14 control buffaloes were provided by the International Water Buffalo Genome Consortium and were sequenced on an Illumina Genome Analyzer at the USDA Beltsville Research Center, USA. Although the pedigree of all sampled individuals was unknown, a principal component analysis conducted with smartPCA from the EIGENSOFT package [119] subsequent to variant calling indicated that the cases were not more related to one another than they were to the controls (Figure 3.7). Disease phenotypes were recorded for each case and a semi-quantitative phenotype score was calculated based on the number of major distal bones present in each limb ranging from 0 (complete loss of both hindlimbs) to 10 (unaffected control) (http://www.nature.com/article-assets/npg/srep/2017/170103/srep39719/extref/srep39719-s1.pdf; Table S1).

Genome sequencing

All animals were sequenced using Illumina technologies and 2 x 100 bp paired end libraries. Sequence depth varied from 5X to 15X average genome coverage, however, the cases were sequenced to an average depth of 12X and controls to only 7X (http://www.nature.com/article-assets/npg/srep/2017/170103/srep39719/extref/srep39719-s1.pdf; Table S1). Raw FASTQ sequences have been deposited to NCBI Short Read Archive (SRA) under BioProject PRJNA350833. Supplementary Table 1 (http://www.nature.com/article-

assets/npg/srep/2017/170103/srep39719/extref/srep39719-s1.pdf) contains sample, experiment, and run accessions for each animal.

Genome alignment and variant detection

Raw sequences were trimmed for adaptors and quality using Trimmomatic-0.33 [120]. The reads were then aligned to the UMD_CASPUR_WB_2.0 water buffalo reference assembly (GCF_000471725.1) using the BWA-MEM algorithm, version 0.7.10-r789 [121]. Subsequently, we built a variant calling pipeline according to GATK Best Practices and optimized the pipeline for a scaffold level reference genome [122–124]. The pipeline included duplicate removal using Picard (http://broadinstitute.github.io/picard), INDEL realignment, SNP and INDEL discovery using HaplotypeCaller, and genotype calling with GenotypeGVCFs. Base quality score recalibration and variant quality score recalibration were not performed due to the lack of availability of a known reference set of polymorphic sites in water buffalo.

SNP and INDEL variant sites were independently filtered. SNPs were filtered based on the number of detected alleles < 3 (biallelic), QD (Variant Confidence/Quality by Depth) < 2.0, FS (Phred-scaled *p*-value using Fisher's exact test to detect strand bias) > 60.0, SOR (Symmetric Odds Ratio of 2x2 contingency table to detect strand bias) > 4.0, MQ (RMS Mapping Quality) < 40.0, MQRankSum (Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities) < -12.5, or ReadPosRankSum (Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias) < -8.0. INDELs were filtered based on QD < 2.0, FS > 200.0, SOR > 10.0, ReadPosRankSum < -20.0, or InbreedingCoeff (Inbreeding coefficient as estimated from the genotype likelihoods per-sample when compared against the Hardy-Weinberg expectation) < -0.8. Furthermore,

SNPs were filtered on an individual animal basis by setting genotypes with a Phred-scale genotype quality (GQ) < 10 to missing.

Case versus control concordance analysis

Filtered SNPs and INDELs were analyzed for concordance on a case *versus* control basis. This involved sorting variants such that all cases were homozygous for an allele for which none of the controls were homozygous. A missing genotype among the cases caused the variant to be rejected from the analysis, but a missing genotype among the controls was ignored due to the lower mean sequence coverage for the controls.

Homozygosity mapping by de novo assembly

Three *de novo* genome assemblies were generated: unilaterally affected TH cases, bilaterally affected TH cases, and controls. Each assembly was initiated by pooling sequence reads from four individuals with the respective phenotype. The reads were assembled using MaSuRCA-3.1.3 using default parameters [70]. We used a negative binomial generalized linear model with the glm.nb function from the MASS package [125] to estimate the mean and dispersion parameter for the contig lengths produced by each assembly. A *p*-value was calculated for each contig to test the hypothesis that the contig was significantly greater in size than the mean, and the *p*-values were corrected for multiple testing by estimating *q*-values [126]. As regions of the genome for which all of the individuals in each pool are homozygous for a single haplotype can be assembled into large contigs (because the assembly graph does not fork), we extracted contigs that were significantly larger than average and those in the 99th percentile for size from each assembly. These contigs were aligned to the UMD_CASPUR_WB_2.0 reference genome assembly and intersected with the water buffalo gene annotation. Finally, the lists of

genes in the largest contigs produced from each assembly were compared. Genes within regions that were homozygous in all TH cases, but not in controls, were identified as candidates for risk of TH.

Genome-wide association study (GWAS)

Given our initial uncertainty as to the mode of inheritance of TH, two GWAS analyses were run. The first was a mixed-model case *versus* control analysis while the second attempted to recover phenotypic information regarding disease severity by scoring the TH phenotypes according to the number of missing hindlimb bones, as previously described. Association tests for both models were performed using univariate linear mixed models and likelihood ratio tests implemented in GEMMA (version 0.94) with a centered genomic relationship matrix [127]. Each analysis was based on 2,990,419 SNPs and statistical significance was determined using a Bonferroni multiple testing correction (p-value < 0.05/2990419).

Candidate region mapping and annotation

Variants identified by the concordance analysis and GWAS were intersected with the water buffalo gene annotation. *De novo* assembled contigs were aligned to the UMD_CASPUR_WB_2.0 water buffalo reference genome assembly using MUMmer3.23 [128]. The resulting reference positions were next compared with the water buffalo gene annotation. Additionally, buffalo scaffolds including either a concordant SNP in the case *versus* control analysis or a significant GWAS association after Bonferroni correction and the top 1% of *de novo* contigs were aligned to the *Bos taurus* UMD3.1 reference genome assembly using MUMmer3.23 [128]. This allowed us

to interpret potential causal loci from the context of a genome as opposed to the 367,000+ unplaced scaffolds.

Candidate gene ontology and network analysis

Candidate genes identified from SNP concordance analyses, GWAS, and homozygosity mapping were uploaded to the BovineMine warehouse [129] to compare the list of candidate genes with the list of bovine genes associated with the GO term "hindlimb morphogenesis" (GO: 0035137). Network analyses were performed on the set of candidate genes associated with the GO term "hindlimb morphogenesis" and all candidate genes identified from GWAS using GeneMANIA [130]. We selected only these genes for network analysis because the entire list was too large and because we wanted to use the network analysis to investigate whether genes identified from the GWAS might be acting as modifier genes in conjunction with what we hypothesize to be the primary driver genes. gProfileR version 0.6.1 [131,132] was used to conduct GO term enrichment analyses using genes identified from SNP concordance analyses, GWAS from binary and quantitative phenotypes, and homozygosity mapping by *de novo* assembly (99^{th} percentile analysis). A Bonferroni multiple testing correcting and a *p*-value < 0.05 were used to determine statistical significance.

Acknowledgements

JFT was supported by grants 2011-68004-30214, 2011-68004-30367, 2013-68004-20364, and 2015-67015-23183 from the USDA NIFA AFRI. JED was supported by grants MO-HAAS0027 and 2016-68004-24827 from the USDA NIFA. The Regione Lombardia, Italy is acknowledged for the "BuffaloSNP" project (ref AGRO-09 ID

16978), which in part funded the sequencing of the control buffaloes, and the MIUR/CNR, FIRB project "GenHome" (GA B81J12002520001) is acknowledged for funding the sequencing of the affected animals.

Author Contributions

SA, FC, VP, FS, CF, and LR collected samples and described the case phenotypes. LKW, JLH, JLW, JFT, and JED designed the analyses. TSS sequenced the animals. LKW, JLH, and RDS completed bioinformatic analyses. LKW completed data analysis and interpreted the results. LKW, JED, JFT prepared the manuscript. All authors read and approved the final manuscript.

Figures

Figure 3.1. Water buffalo calves with transverse hemimelia (TH). (A) A bilaterally affected TH case with both hindlimbs completely absent at birth and hypoplasia of carpal bones and absence of medial bones starting from metacarpus and X-ray images. (B) A unilaterally affected TH case with one hindlimb truncated at the tarsus with X-ray image.

[Figure 1 in publication.]

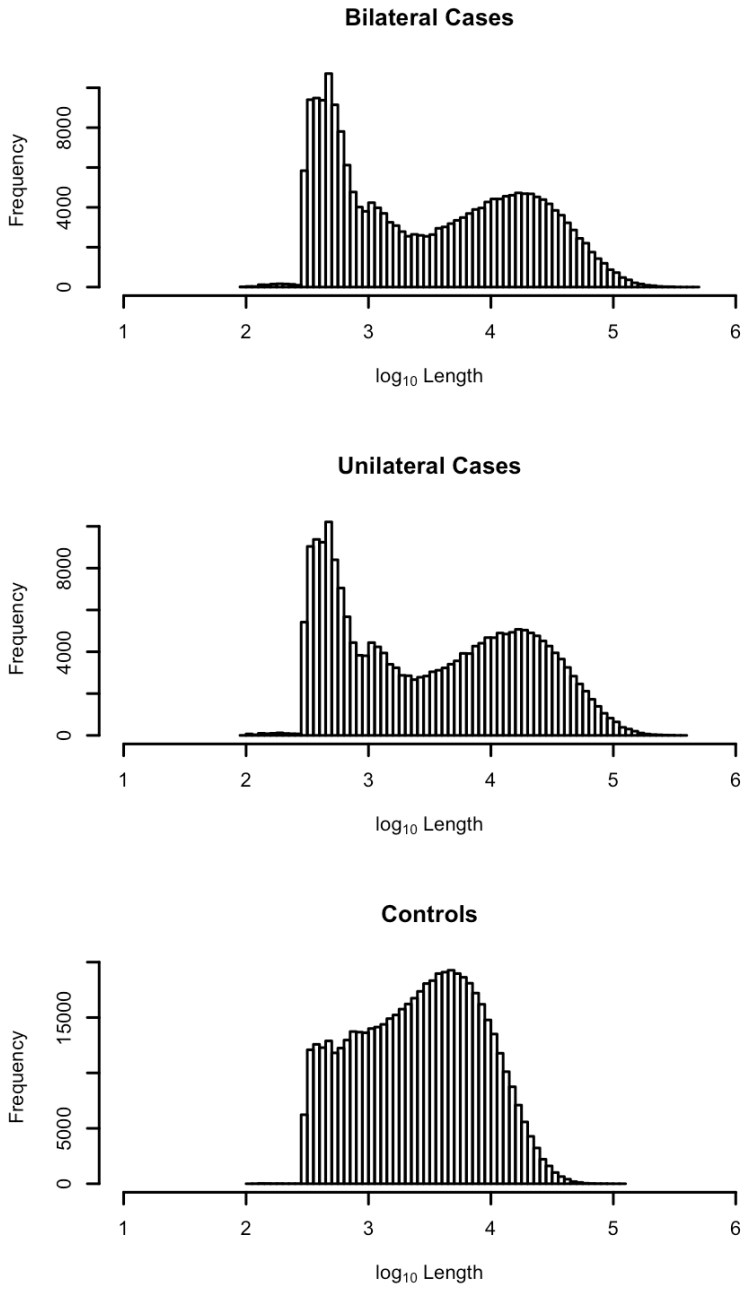

Figure 3.2. Log_{10}-transformed distribution of contig sizes from the *de novo* assembly of pooled sequences from the bilaterally affected cases, unilaterally affected cases and controls.

[Figure 2 in publication.]

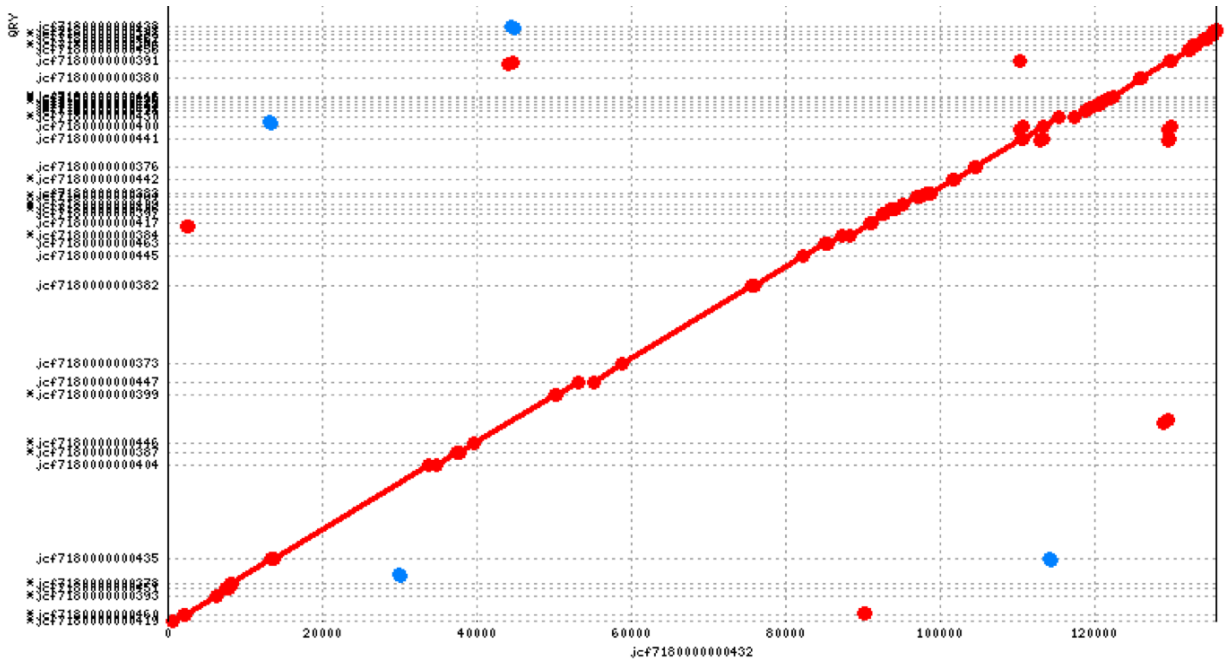

Figure 3.3. Dot plot of multiple contigs comprising the *SMARCA4* gene region in controls *versus* a single contig comprising the *SMARCA4* gene region in cases indicates increased homozygosity in *SMARCA4* in affected animals.

[Figure S1 in publication.]

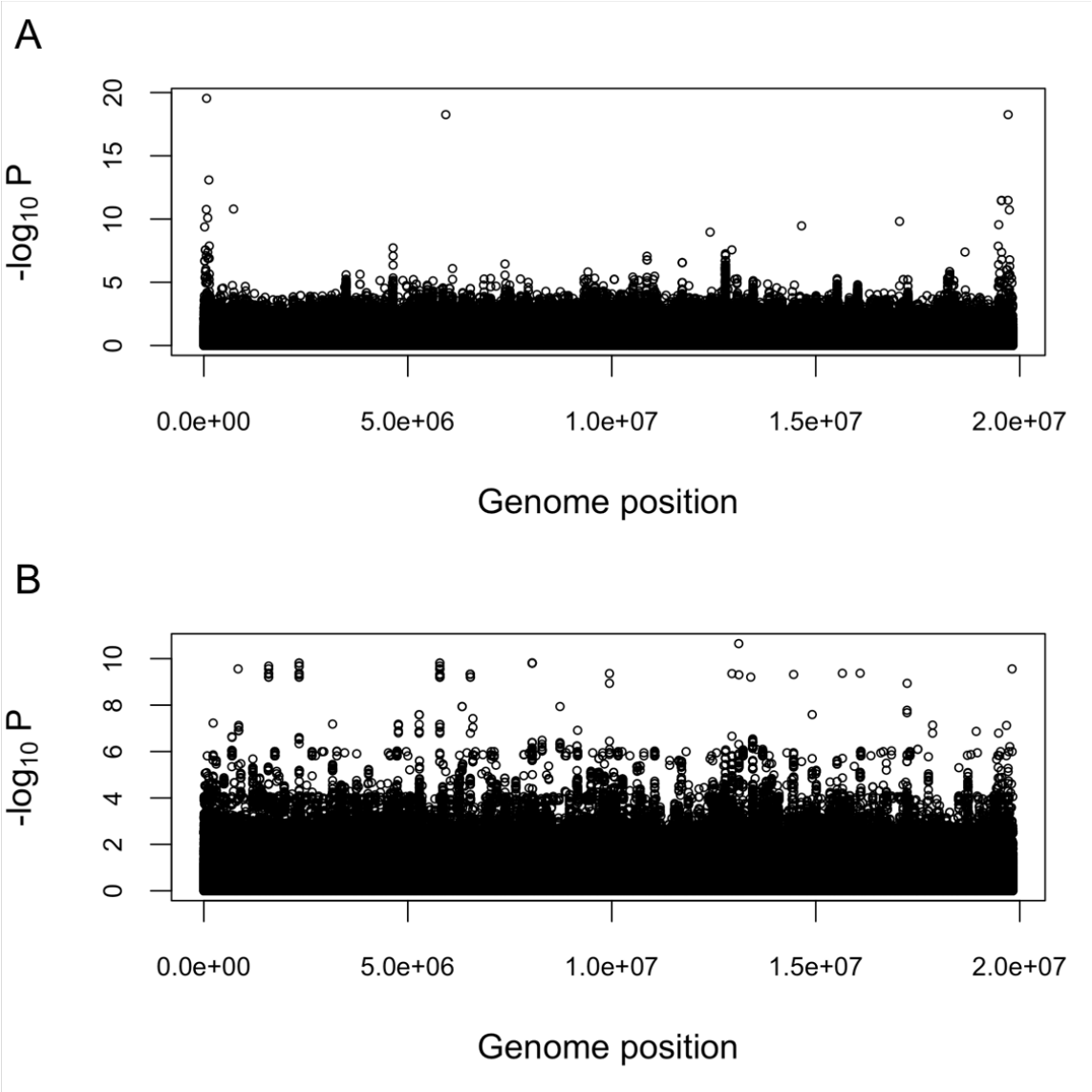

Figure 3.4. Manhattan plots of GWAS results. (A) GWAS results from association with a binary phenotype. (B) GWAS results from association with a semi-quantitative phenotype.

[Figure 3 in publication.]

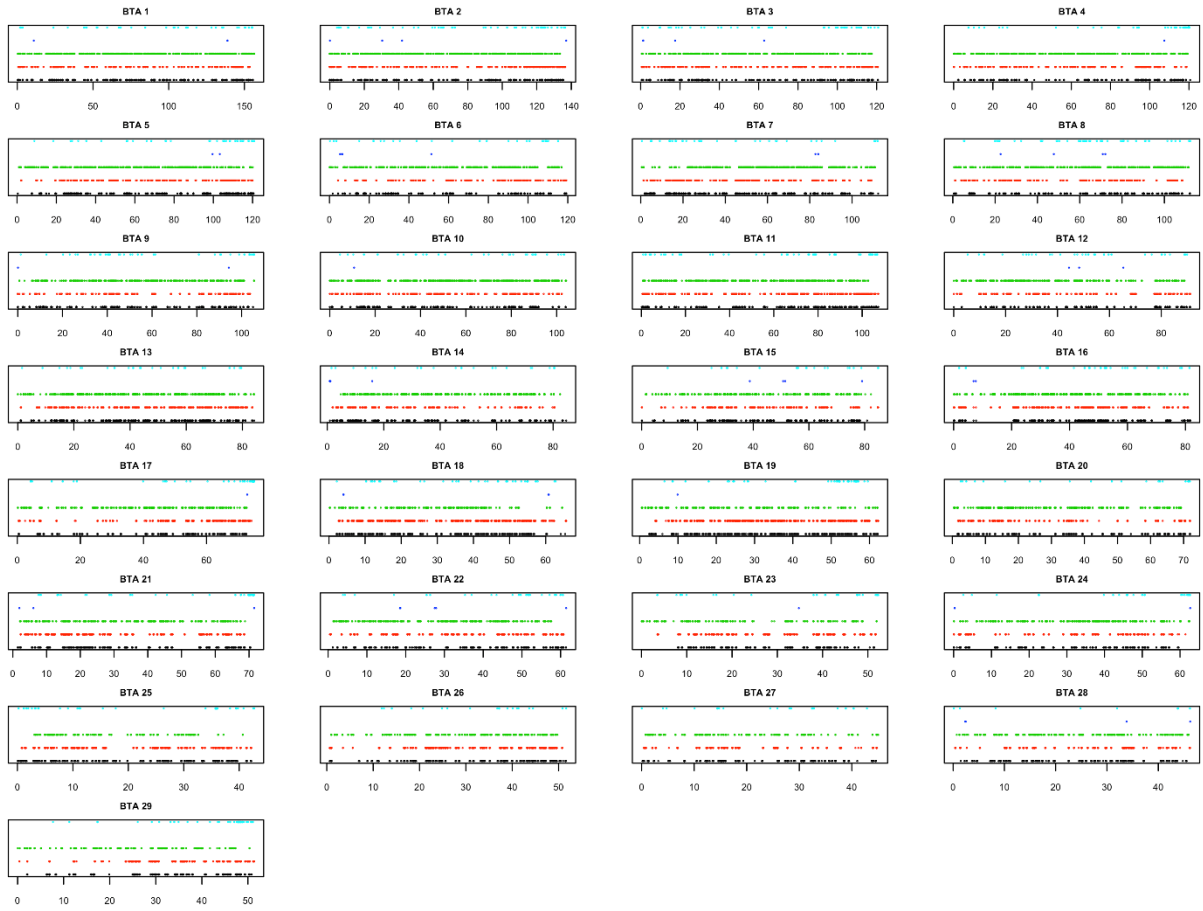

Figure 3.5. Mapping of regions significantly associated with TH to the *Bos taurus* UMD3.1 reference assembly from bilaterally affected case homozygosity mapping (black), unilaterally affected case homozygosity mapping (red), control homozygosity mapping (green), GWAS (dark blue), and SNP concordance analysis (light blue). [Figure S2 in publication.]

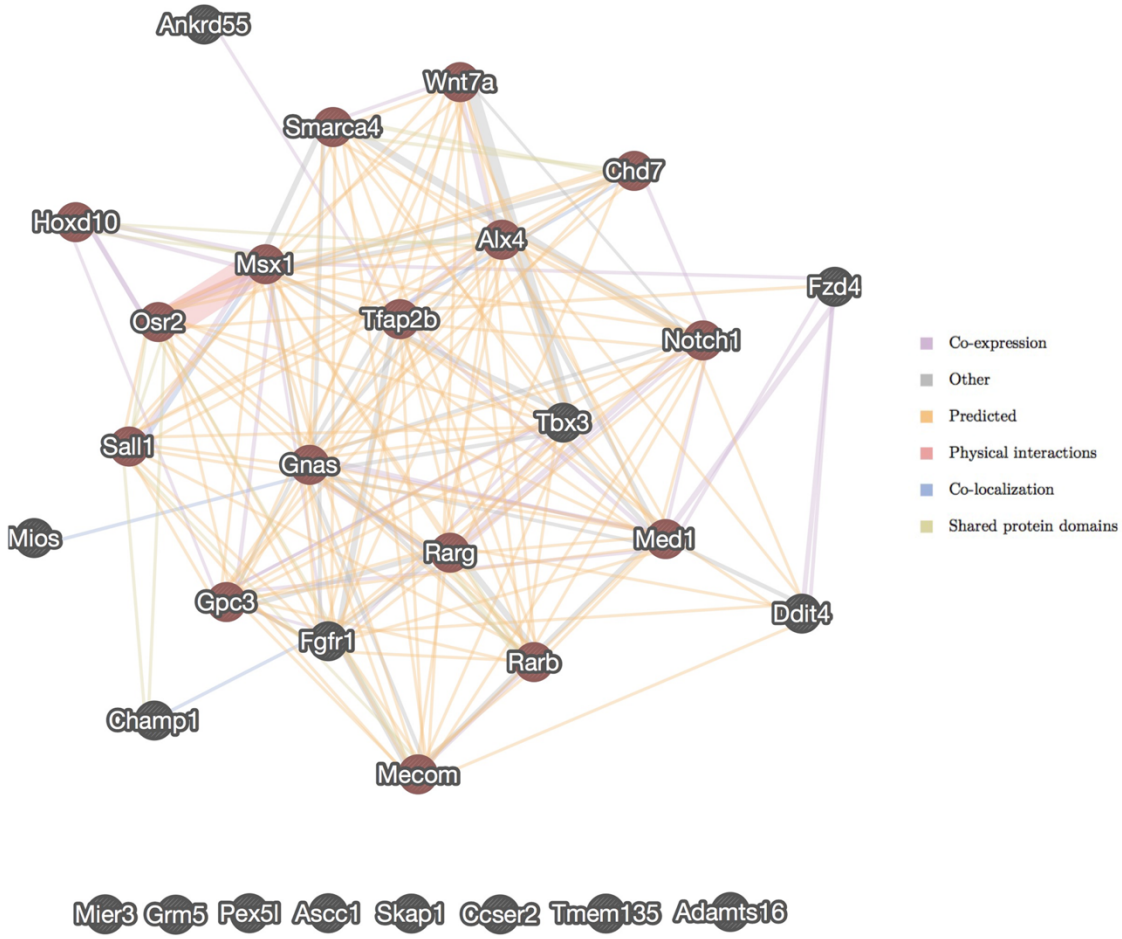

Figure 3.6. Network analysis of genes predicted to be associated with transverse hemimelia (TH) based on SNP concordance, homozygosity mapping by *de novo* assembly, and GWAS analyses. Genes associated with hindlimb morphogenesis (GO: 0035137) are shaded in red.

[Figure 4 in publication.]

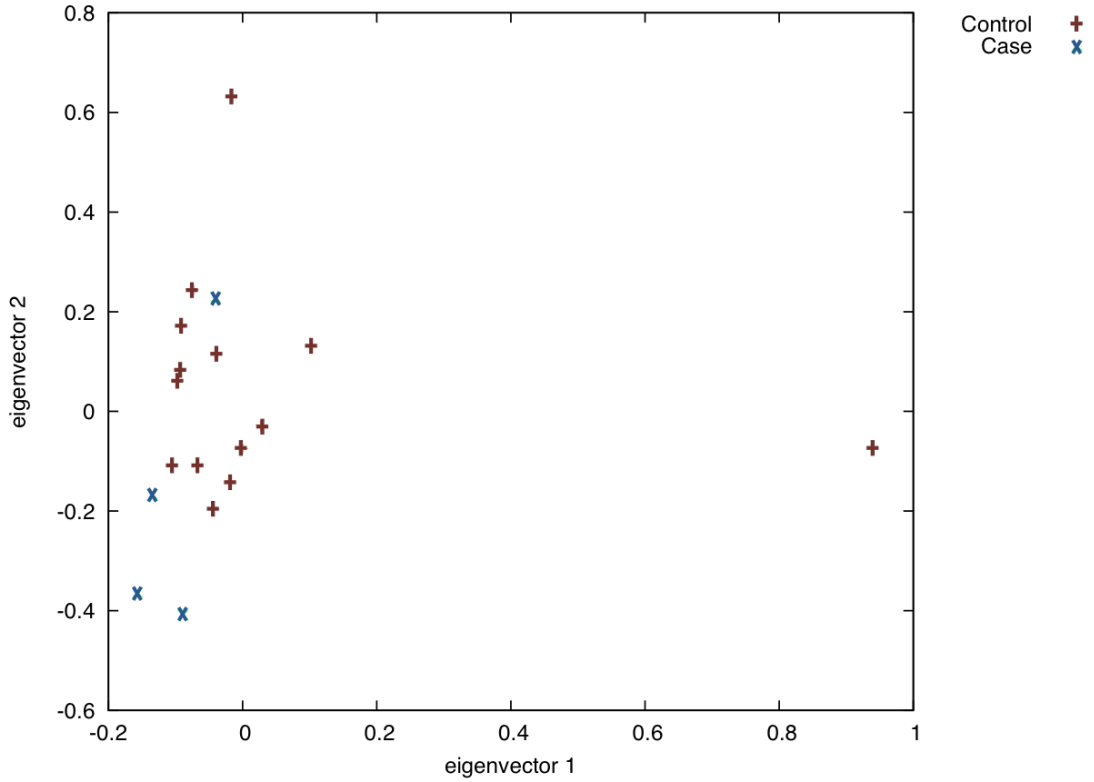

Figure 3.7. Principal component analysis of genotypes for 11 TH affected cases and 14 controls.

[Figure S3 in publication.]

CHAPTER 4

GENOME-WIDE VARIATION AND POPULATION STRUCTURE IN NEOSHO MADTOM CATFISH

Lynsey K. Whitacre[1,2], Mark L. Wildhaber[3], Gary S. Johnson[4], Robert D. Schnabel[1,2], Justin M. Downs[5], Tendai Mhlanga-Mutangadura[4], Vernon M. Tabor[6], Daniel Fenner[7], Jared E. Decker[1,2*]

[1] Informatics Institute, University of Missouri, Columbia, Missouri 65211, USA

[2] Division of Animal Sciences, University of Missouri, Columbia, Missouri 65211, USA

[3] U.S. Geological Survey, Columbia Environmental Research Center, Columbia, Missouri 65201, USA

[4] Department of Veterinary Pathobiology, College of Veterinary Medicine, University of Missouri, Columbia, Missouri 65211, USA

[5]The Peoria Tribe of Indians of Oklahoma, Miami, Oklahoma 74354, USA

[6]U.S. Fish and Wildlife Service, Kansas Ecological Services Field Office, Manhattan, Kansas 66502, USA

[7]U.S. Fish and Wildlife Service, Oklahoma Ecological Services Field Office, Tulsa, Oklahoma 74129, USA

[*]Corresponding author

Abstract

The Neosho madtom (*Noturus placidus*) is a small catfish, generally less than 3 inches in length, unique to the Neosho-Spring River system within the Arkansas River Basin. It was Federally listed as threatened in 1990, largely due to habitat loss. As part of conservation efforts, we generated whole genome sequence data from ten Neosho madtom originating from three geographically separated populations to evaluate genetic diversity and population structure. Single nucleotide polymorphisms (SNPs) were assessed *de novo* and via reference alignment with the closely related channel catfish (*Ictalurus punctatus)* reference genome. Principal component analysis and structure analysis indicated weak population structure, suggesting fish from the three locations represent one panmictic population. Sequence data were also used for whole-genome *de novo* assembly and assessment of genome size and content. Genome-wide divergence between the Neosho madtom and channel catfish was assessed by pair-wise scaffold alignment and demonstrated that genes important to embryonic development were largely conserved in sequence. This research in a threatened species with no previous genomic research or technologies provides novel genetic information to guide current and future conservation efforts and demonstrates the feasibility of using whole genome sequencing in other ongoing conservation efforts.

Introduction

The Neosho madtom (*Noturus placidus*) is a small ictalurid, generally less than 3 inches in length, unique to parts of Kansas, Missouri, and Oklahoma in the Neosho, Cottonwood, and Spring Rivers [133–137] (Figure 4.1). The small fish inhabit loosely compacted gravel in high to moderate velocity water that is generally associated with riffles [138,139]. Due to factors such as dam construction, agricultural runoff, and lead-zinc mining, much of the species' historical habitat has been destroyed [140–142]. The Neosho madtom was listed as threatened in 1990 by the U.S. Fish and Wildlife service (55 FR 21148) and a recovery plan was put in place in 1991 [143]. Over the last 17 years, several research efforts have focused on elucidating the habitat and behaviors of these fish to aid in their conservation [138,144–146]. In addition, conservation efforts have focused on the relationships between population density to water quality and nutrients and competition from other species [142,146].

Extant populations of Neosho madtom exist in regions of the Neosho, Cottonwood, and Spring Rivers with the largest populations residing in the Neosho and Cottonwood Rivers above the John Redmond Reservoir. The Spring River population is particularly sparse, most likely due to the presence of cadmium, lead, and zinc and limited food and habitats [134,147,148]. Sampling of Neosho madtom over the last 26 years has indicated an overall decline in population density (fish per square meter), with density declining steadily from 1991 to 2008 [149]. Populations on opposing sides of the reservoir may have been geographically isolated since the dam was constructed nearly 60 years ago [136]. Additional isolation of populations may have also occurred because of the presence of low-water dams found throughout the Neosho and Cottonwood Rivers

[136]. These potential sources of population isolation prompted research to determine the extent to which these populations have diverged from one another [136,137]. The objective of this study was to determine whether differences in selective pressures or genetic drift, which can have a significant impact on genetic divergence between small populations [150,151], has resulted in genetically unique populations within the current range of the Neosho madtom.

Significant advances in DNA sequencing technology have made it possible for next-generation sequencing (NGS) methods to efficiently produce genomic data that can be used to reveal genetic variation on a base-by-base level and aid in conservation efforts of threatened and endangered species such as the Iberian lynx, giant panda, scarlet macaw, and Tasmanian devil [152–155]. In this paper, we describe the application of NGS technology for assessing genetic variability and population structure in Neosho madtom from the upper and lower Neosho River and the Cottonwood River. We also completed a draft *de novo* assembly of the Neosho madtom genome and determine divergence from the channel catfish (*Ictalurus punctatus*) on a genome-wide scale.

Methods

Sample collection

Neosho madtom samples were collected from the upper Neosho River above the John Redmond Reservoir (UNR), the lower Neosho River below the John Redmond Reservoir (LNR), and from the Cottonwood River (CR) (Figure 4.2; Table 4.1). An additional sample from a Stonecat (*Noturus flavus*) was also collected to serve as an

outgroup for phylogenetic analyses. Genomic DNA was extracted from tissue using phenol-chloroform extraction.

Genome sequencing

Genomic DNA was used to construct one small insert paired end library per fish with an approximate insert size of 300 bp. Each library was sequenced using Illumina technology to an average coverage of 39X (Table 4.2). Additionally, two mate-pair libraries with insert sizes of 2,000 and 3,000 bp were constructed using genomic DNA from one fish to improve the scaffolding of the *de novo* assembly. All sequences have been deposited in NCBI's SRA database (Table 4.2).

De novo variant calling and filtering

The cortex_var algorithm was used for *de novo* variant discovery [156]. Cortex calls variants by constructing multicolor *de Bruijn* graphs from the DNA sequence reads from all samples. We constructed multicolor *de Bruijn* graphs using *k*-mer sizes of 31 and 63 bp from individual graphs corrected for low coverage supernodes. Variants were subsequently called as the algorithm searched for motifs within the graph, referred to as bubbles, that are created by polymorphisms or repeats. *K*-mers of 31 and 63 were used jointly in variant calling to maximize the likelihood of detection of variants that may only be visible at low or high *k*.

Variants were first filtered based on variant type such that only biallelic single nucleotide polymorphisms (SNPs) were retained. These were then filtered based on genotype confidence, which is defined as the natural log probability of the maximum likelihood genotype – log probability of the second mostly likely genotype. Finally, SNPs

with a genotype confidence < 5.54 (meaning the remaining genotypes called are less than $e^{5.54}$ or 254.5 times more likely than the alternative) were filtered from the data set.

Reference variant calling and filtering

During the course of the project, the channel catfish (*Ictalurus punctatus*) reference genome became publicly available [157] and was subsequently used to call variants. Raw sequences were trimmed for adaptors and base quality using Trimmomatic-0.33 [120]. The reads were then aligned to the IpCoco_1.2 channel catfish reference assembly (GCA_001660625.1) using the BWA-MEM algorithm, version 0.7.10-r789 [121]. We then built a pipeline according to GATK Best Practices to call variants from the alignment files [122–124]. The pipeline included modules for duplicate removal using Picard (http://broadinstitute.github.io/picard), INDEL realignment, SNP and INDEL discovery using HaplotypeCaller, and genotype calling with GenotypeGVCFs. Base quality score recalibration and variant quality score recalibration was not performed due to the lack of availability of a reference set of validated polymorphic sites in catfish.

SNPs were filtered based on the number of detected alleles < 3 (biallelic), QD (Variant Confidence/Quality by Depth) < 2.0, FS (Phred-scaled *p*-value using Fisher's exact test to detect strand bias) > 60.0, SOR (Symmetric Odds Ratio of 2x2 contingency table to detect strand bias) > 4.0, MQ (RMS Mapping Quality) < 40.0, MQRankSum (Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities) < -12.5, or ReadPosRankSum (Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias) < -8.0. The remaining set of SNPs was deposited in NCBI's dbSNP database (accessions TBD).

Principal component and structure analysis

Principal component analysis (PCA) was conducted using the smartpca module from the EIGENSOFT 5.0 population genetics package [119] using the high confidence and filtered genotype sets from *de novo* and reference variant calling for all samples (including the outgroup) and for only Neosho madtom. PCA excluding the outgroup was conducted after removing variants that were fixed for different alleles between the Neosho madtom and the outgroup. Tracy-Widom staistics were calculated to determine the significance of eigenvalues.

Structure analysis was performed using fastSTRUCTURE, an algorithm designed to use high density SNP data and a variational Bayesian framework for posterior inference to infer the ancestry of individuals [158]. FastSTRUCTURE was run using K = 1 to K = 4 with simple priors on SNP data from all individuals (Neosho madtom and Stonecat) and from K = 1 to K = 3 on SNP data from Neosho madtom only. These K values were tested as they represent the total number of possible madtom populations and outgroup species. Both analyses were run using *de novo* and reference called variants separately.

Estimation of historical effective population size

The historical effective population size of both Neosho madtom and Stonecat populations was estimated using SMC++ [159] and genotypes from unphased whole genome sequence data. SMC++ utilizes variant calls and genome position information to estimate recombination rates across varying block sizes. Therefore, only variants called from reference alignment were used for the estimation of historical effective population size. The divergence time between Neosho madtom and Stonecat was also inferred using

SMC++. All estimates were calculated and interpreted using a generation interval of 1 year, as suggested by Bulger and Edds [160], and a Watterson estimation of mutation rate. The Watterson estimate is based on coalescent theory and is calculated by dividing the number of segregating variants by the $(n - 1)^{th}$ harmonic number [161].

Genome size and de novo assembly

Two paired-end and two mate-pair libraries were sequenced from a single Neosho madtom and were assembled using MaSuRCA-2.3.2 [162]. This assembler uses the QuorUM error correction method [25] and automatically chooses an appropriate *K*-mer size. It relies on both *de Bruijn* graph and overlap layout consensus (OLC) methods for accurate assembly. Parameters were adjusted to limit jump coverage to 150, set the maximum error rate for the scaffolder to 0.10, and use a Jellyfish [163] hash size of 10^{10}. The resulting assembly has been submitted to NCBI's Genome database (pending accession number).

We also used *K*-mers from whole genome sequence reads to estimate the genome size of the Neosho madtom. Assuming *K*-mers are uniquely mapped to the genome, their frequency reflects the depth of coverage (the sequencing of each base in the genome multiple times). Therefore, the genome size can be estimated as the total number of *K*-mers divided by the average frequency or coverage. We determined the *K*-mer distribution using Jellyfish [163] to count *K*-mers using $K = 31$. Counts were summed to determine the total number of *K*-mers analyzed and visualized as a histogram. *K*-mers with coverage < 5 were truncated as they likely represent sequencing errors. Using the peak of the histogram, genome size is estimated as the total number of *K*-mers analyzed divided by the value of the most frequent *K*-mer (coverage). We can also estimate the

portion of the genome that is single copy by calculating the total number of *K*-mers represented by the bell curve without considering the tails of the distribution ($5 < K$-mer count < 45).

Whole genome analysis of divergence

Scaffolds from the *de novo* assembly of Neosho madtom via MaSuRCA [162] were aligned to the channel catfish (GenBank accession GCA_001660625.1) and zebrafish (GenBank accession GCA_000002035.3) reference genomes using NUCmer 3.1 [128] with a break length of 200 bp, a minimum length for a maximal exact match equal to 20, and a minimum cluster length of 65. The alignment was then used to perform a whole genome analyses for divergence. Orthologous regions of the channel catfish and zebrafish genomes to the Neosho madtom genome were determined by observing the best alignment and selecting additional secondary, non-overlapping alignments on the same chromosome. All other spurious alignments to a chromosome that differed from the identified orthologous region for each scaffold were filtered from the alignment data set.

For each valid alignment, the number of identical aligned bases was calculated as the percent identity × alignment length. A generalized linear model was then fit analyzing the number of identical aligned bases ÷ the total alignment length using a quasi-binomial distribution weighted by the alignment length. Outliers from the model were determined using a Bonferroni corrected *p*-value for the Studentized residuals from the generalized linear model. Outliers were determined to be significantly conserved if the Bonferroni p-value < 0.05 and Studentized residual > 0 or significantly diverged if the Bonferroni p-value < 0.05 and Studentized residual < 0. The calculation of this statistic was motivated by the correction for alignment length in an analysis previously reported by Seabury et al.

[155]. Regions determined to be significantly conserved or diverged were next intersected with the channel catfish or zebrafish GFF file, respectively, to determine the genes located within the region.

Results

Genetic variation

The *de novo* variant discovery algorithm was more successful in calling variants when sequences from Neosho madtom and Stonecat were not combined (Supplementary Note 1). When variants were discovered *de novo* within Neosho madtom only, 967,681 SNPs were detected and 709,790 SNPs (73.5%) passed quality filtering. The channel catfish reference genome became publicly available during our research and we also called variants by directly aligning reads to this assembly. Despite the divergence time between channel catfish and madtom, on average, 80.5% of the reads aligned to the assembly (Table 4.3). There were 40.4 million SNPs detected, with approximately 31.1 million (77%) being fixed different (homozygous alternate) in Neosho madtom and Stonecat. Thus, 9.3 million SNPs were variable across the madtom species. Within those SNPs, 2.1 million SNPs were determined to be variable within Neosho madtom and a proportion were uniquely variable to each population (Table 4.4). The most genetic variation was detected in LNR Neosho madtom, however, this result is confounded by having one additional sample from the lower Neosho population.

Population structure

Using variants called from the alignment to the channel catfish reference, principal component analysis was conducted to analyze the relationships between the

three populations of Neosho madtom as well as between the Neosho madtom and the Stonecat. When all 11 fish were considered, the primary source of variation was, as expected, between the two species of madtom (Figure 4.3A). Eigenvector 1 for this PCA accounted for 89.21% of the variation (Table 4.5). Eigenvector 2 seems to separate the Neosho madtom populations, but the clusters are not well defined. By computing Tracy-Widom statistics for eigenvector 1, we conclude that the eigenvector is not significant (*p*-value = 0.1962), and that the population structure is weak.

When only the Neosho madtom are considered, the populations, and some individuals, separate from one another (Figure 4.3B), but, unlike the first PCA, their eigenvalues are nearly equal (Figure 4.4). In this case, eigenvector 1 only explains 13.56% of the variance and has a Tracy-Widom *p*-value = 0.9847 (Table 4.5). This indicates that there is no evidence of population structure among the different populations of Neosho madtom. The ancestry analysis recapitulated this result, giving an optimal value of K (the number of ancestral populations) of two if Neosho madtom and Stonecat SNP genotypes were provided (Figure 4.5) or one if only Neosho SNP genotypes were provided. *De novo* variant calls reiterated these findings (Figures 4.6 and 4.7). This suggests that the Stonecat and Neosho madtom populations have differentiable ancestry, but the UNR, CR, and LNR Neosho madtom populations do not.

Estimation of historical effective population size

Estimation of the historical effective population size of Neosho madtom and Stonecat suggests that the Stonecat population has historically been a larger population than the Neosho madtom (Figure 4.8A). This was also the case for the estimates of current effective population size. As expected, the historical effective population size

estimates for each Neosho madtom population were similar (Figure 4.8A). The most recent effective population size estimates for Neosho madtom show the largest effective population size in the LNR population and the smallest effective population size in the UNR population. It is important to note that, due to the large parameter space associated with the estimation of effective population size, the precision of the inferences is low. We also estimated the divergence time between Neosho madtom and Stonecat to be approximately 10,000 years ago (Figure 4.8B) assuming a clean split and a generation interval of one year.

Genome size and de novo assembly

The decreased genetic diversity and small genome size for the Neosho madtom allowed us to generate a quality whole genome *de novo* assembly. Whole genome *de novo* assembly utilized sequence reads from two paired-end and two mate-pair libraries. The assembly resulted in 56,087 scaffolds with an N50 of 108,346 bp. A total of 899,438,561 bases were assembled (Table 4.6). Before assembly, the genome size was estimated to be 981,199,801 bases using *K*-mer counts from the sequence data. Assuming the Nesoho madtom genome is similar or equal in size to the 1.0 Gb approximate genome size of the channel catfish [164,165], we were able to assemble ~91.7% of the genome and hypothesize that some regions of the genome comprising repetitive sequences were not able to be resolved by the assembler.

The portion of the genome that is single copy was also estimated using *K*-mer counts. The single copy regions can be identified by the total number of *K*-mers in the bell curve portion of the *K*-mer distribution divided by the total estimated genome size. For this sequencing library (sample 93670, paired end library abbreviation aa; Table 4.2)

, coverage for the single copy regions was between 5 and 45 (Figure 4.9). Using the *K*-mer counts within this range of coverage, approximately 72.9% of the genome was estimated to be single copy. The remaining portion of the genome that harbors duplicated sequences likely contributed to the inability to assemble the entire genome.

Whole genome analysis of divergence

Scaffolds produced in the *de novo* assembly of the Neosho madtom were compared to the channel catfish and zebrafish reference assemblies to assess divergence on a genome-wide scale. Scaffolds were aligned to each reference assembly and the alignment lengths and percent identity were used to determine significantly diverged and conserved regions (see methods). NUCmer alignments against the channel catfish reference genome produced at least one valid alignment for 60.1% of the *de novo* assembled scaffolds. The average percent identity and length of all alignments was 84.68% and 2,477 bases, respectively (Figure 4.10A-B). Of these alignments, 32 were statistical outliers identified as being highly conserved and 5 were statistical outliers identified as highly diverged between the species (Figure 4.10C). Highly conserved regions contained 47 unique genes annotated in the channel catfish reference (Table 4.7). This included 19 homeobox or homeobox-like genes, known to encode transcription factors that are involved in the development of the vertebrate body plan [166]. As expected, highly diverged regions were mostly found in non-coding regions of the genome and overlapped with only 3 genes (Table 4.8).

NUCmer alignments of Neosho madtom scaffolds against the zebrafish reference genome resulted in at least one valid alignment for only 14.4% of the *de novo* assembled scaffolds. The average percent identity of alignments was 87.74% and the average length

of alignments was 3,766 bases (Figure 4.11A-B). No significantly conserved regions were identified using our method. However, 26 significantly diverged regions were identified (Figure 4.11C). Those regions overlapped 12 unique genes annotated in the zebrafish reference (Table 4.9). It is important to note that our method inherently fails to identify some significantly diverged regions of the genome when scaffolds cannot be aligned at the required stringency due to the high divergence. However, high divergence is not the only reason why a scaffold might not align to a reference genome. Other reasons include misassemblies or regions that are absent from one or both genomes and low complexity or repetitive sequence. Scaffolds falling into each of these classes are difficult to identify with any confidence.

Discussion

Neosho madtom were placed on the Federal Threatened and Endangered Species list in 1990 and have subsequently become a focus of several research efforts. Along with the removal of small dams to improve habitat range for the species, a primary need identified in conservation and recovery efforts associated with Neosho madtom has been population genetic information in support of reintroduction efforts into areas of their historic range from which they have been extirpated (e.g., sections of the Spring River in Oklahoma) [136,137]. As the first attempt to address genetic variation and population and genome structure in three of the remaining populations, we collected and sequenced DNA from 11 fish. Using these data, we demonstrate that although *de novo* variant detection is a viable option, additional variation can be detected by aligning to a reference genome of a closely related species. Alignment to the reference genome revealed genetic variation at a density of approximately 1 SNP per 5 kb of sequence in the Neosho

madtom genome. This is much lower than the extent of genetic diversity in humans who have approximately 4 SNPs per 5 kb of sequence [167]. Even greater genetic diversity is found in cattle, chicken, and swine [168–171].

The whole genome sequencing approach used in this research provides more strength for population genetic conclusions than the more commonly used microsatellite analyses, which yield only a small fraction of the genetic data that is generated by sequencing projects. Analysis of population structure based on SNPs indicates that Neosho madtom from the upper and lower Neosho River and the Cottonwood River represent a single panmictic population. Although these populations were not thought to interbreed due to geographical isolation, it is possible that eggs or fish from the UNR and CR populations are moving downstream and contributing genetic material to the LNR population. This would explain the increased genetic variation and effective population size seen in the LNR Neosho madtom. The overwhelming support for a panmictic population is useful for future conservation efforts, which may utilize translocation.

Historic effective population sizes were also estimated using SNPs and provide some insight into the bottlenecks that have occurred in both Neosho madtom and Stonecat. Neosho madtom populations experienced steep declines since approximately 10,000 years ago. This could be due to the colonization of the Americas by humans, which is estimated to have occurred nearly 13,000 years ago [172,173]. The Stonecat has also shown an overall decline in effective population size. The estimation of its decline may be less accurate than that for the Neosho madtom due to the small sample size used in this study.

De novo assembly allowed whole genome divergence analyses between Neosho madtom and channel catfish and provides a novel resource for future madtom genomics research. Many conserved genes encode transcription factors, which have previously been observed to be conserved across species [174]. Extreme conservation of the regions of the genome containing developmental regulatory genes such as the homeobox clusters emphasizes the importance of the sequence and, subsequently, structure of these proteins. The homeobox genes and their regulatory elements have previously been reported to be highly conserved across vertebrates [175–178], indicating the importance of developing an apt body plan. Genes identified as significantly diverged between Neosho madtom and either channel catfish or zebrafish do not appear to have any coordinating theme. However, both genes identified as significantly diverged between Neosho madtom and zebrafish (*PLCH1* and *PTPRD*) have functions related to metabolism. It is also interesting to note that several genes identified as significantly conserved between Neosho madtom and channel catfish were identified as significantly diverged in the analysis between Neosho madtom and zebrafish.

Overall, our research characterizes the genetic variability in Neosho madtom and suggests that the geographically isolated populations have not drifted substantially from one another based on whole genome comparisons. This information and the resources we have developed, including a genome assembly and a catalog of genetic variants, can be used to advise future conservation efforts where the objective is to minimize the chances of accumulating detrimental variants that lower fitness (genetic rescue) [179]. A detailed genetic history and robust analysis of population structure are also important prerequisites for management strategies such as genetic restoration, a more

comprehensive approach that aims to eliminate the effects of detrimental variation while maintaining advantageous variation and neutral variation that may be adaptive in the future [180–184]. In addition, we have demonstrated the feasibility of utilizing next generation sequencing technologies in the population genetic analysis of any threatened or endangered species that does not possess a reference genome or other genomic resources. Whole genome comparisons were crucial to reinforcing the validity of our conclusions due to the small sample size, which is often a common feature of conservation genetic studies. To further assist conservation efforts for the Neosho madtom the next step would be to collect and sequence samples from populations in the Spring River in order to determine if Neosho and Cottonwood River populations are sufficiently genetically similar to be used for the reintroduction of the species into sections of the Spring River where they are currently not found.

Acknowledgements

We would like to thank The Genome Analysis Centre, Norwich, UK for providing additional sequence for one of the Neosho madtom samples. Funding was provided by The Peoria Tribe of Indians of Oklahoma through U.S. Fish and Wildlife Service grants. Thanks to Janice Albers for help with the culturing of wild Neosho madtoms to a size large enough to obtain sufficient DNA for sequencing. Thanks to Diana Papoulias for help in sample preparation. Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Figures

Figure 4.1. The Neosho madtom catfish (photo credit: Janice Albers).

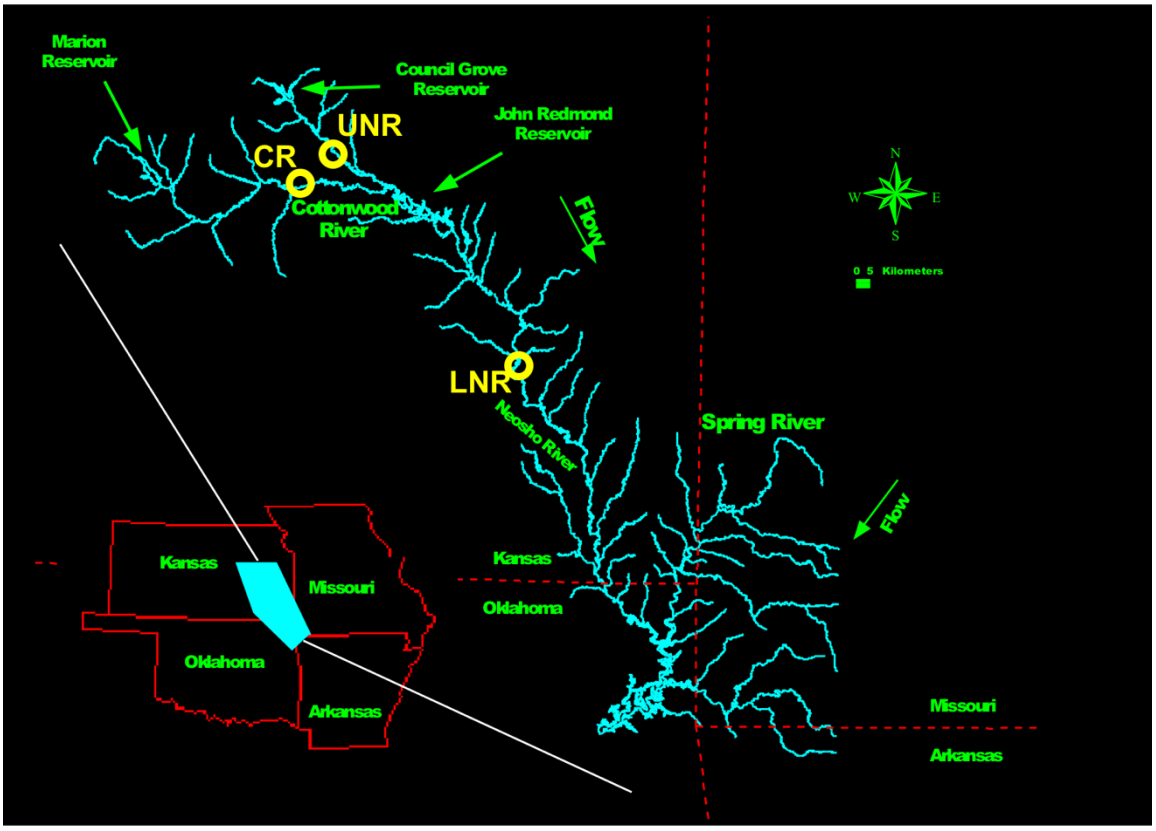

Figure 4.2. Sampling locations of Neosho madtom for sequencing. Locations include the Cottonwood River (CR), the upper Neosho River (UNR), and the lower Neosho River (LNR).

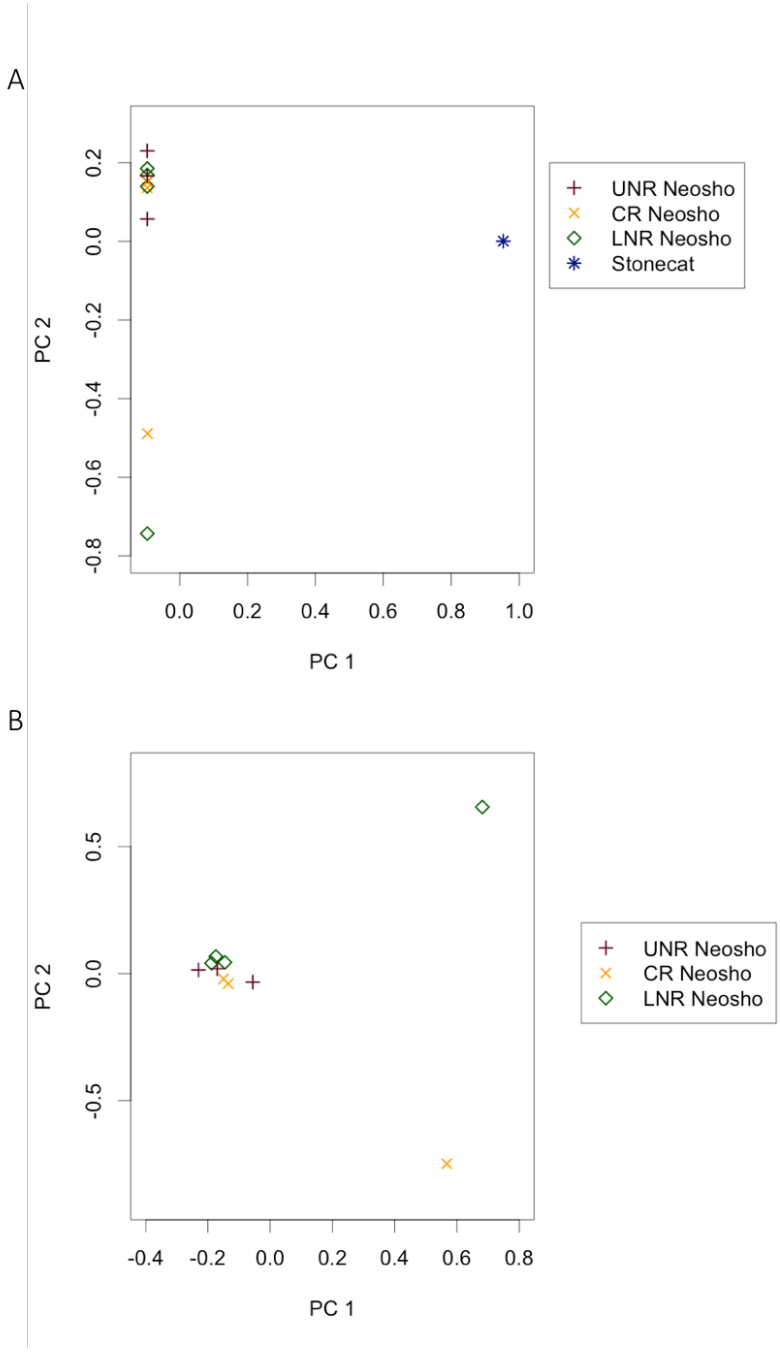

Figure 4.3. Eigenvectors 1 and 2 from principal component analysis of Neosho madtom and Stonecat (A) and Neosho madtom (B) based on SNPs discovered from reference alignment.

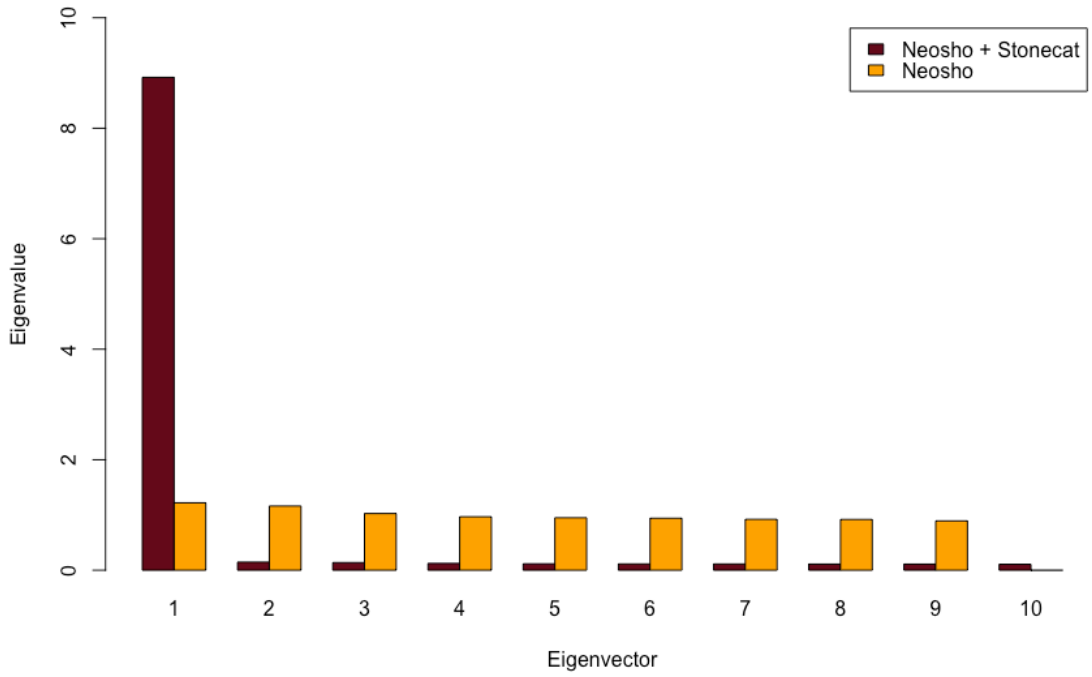

Figure 4.4. Eigenvalues of eigenvectors from principal component analysis of Neosho madtom and Stonecat and Neosho madtom.

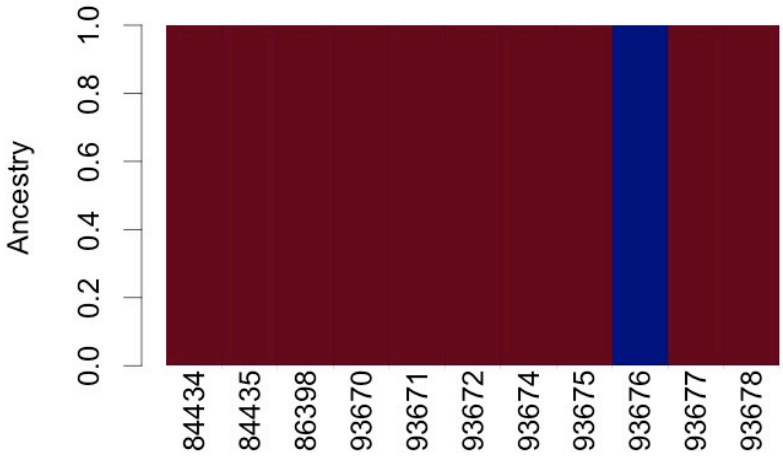

Figure 4.5. Structure analysis of Neosho madtom and Stonecat with an optimal value of $K = 2$ shows pure and identical ancestry of all Neosho madtom individuals with SNPs discovered from reference alignment.

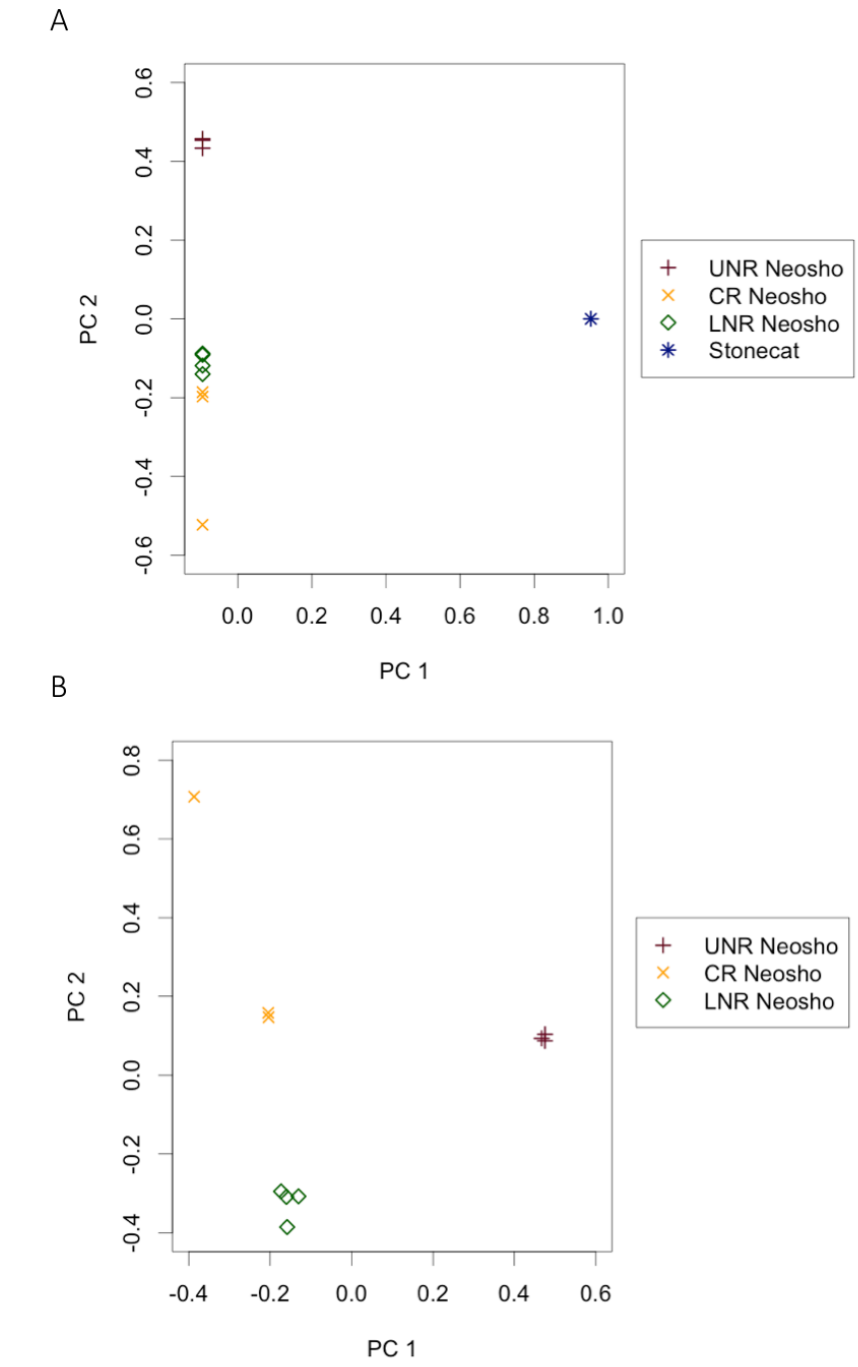

Figure 4.6. Eigenvectors 1 and 2 from principal component analysis of Neosho madtom and Stonecat (A) and Neosho madtom (B) based on *de novo* discovered SNPs.

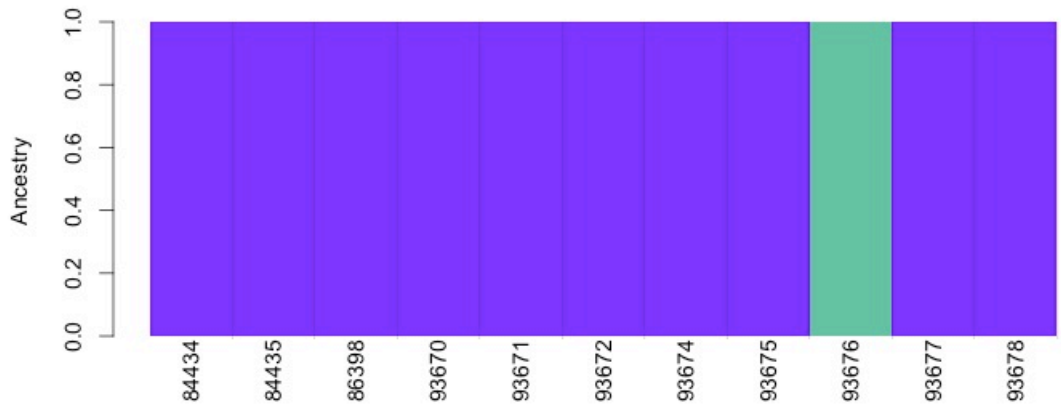

Figure 4.7. Structure analysis of Neosho madtom and Stonecat with an optimal value of $K = 2$ shows pure and identical ancestry of all Neosho madtom individuals with SNPs discovered *de novo*.

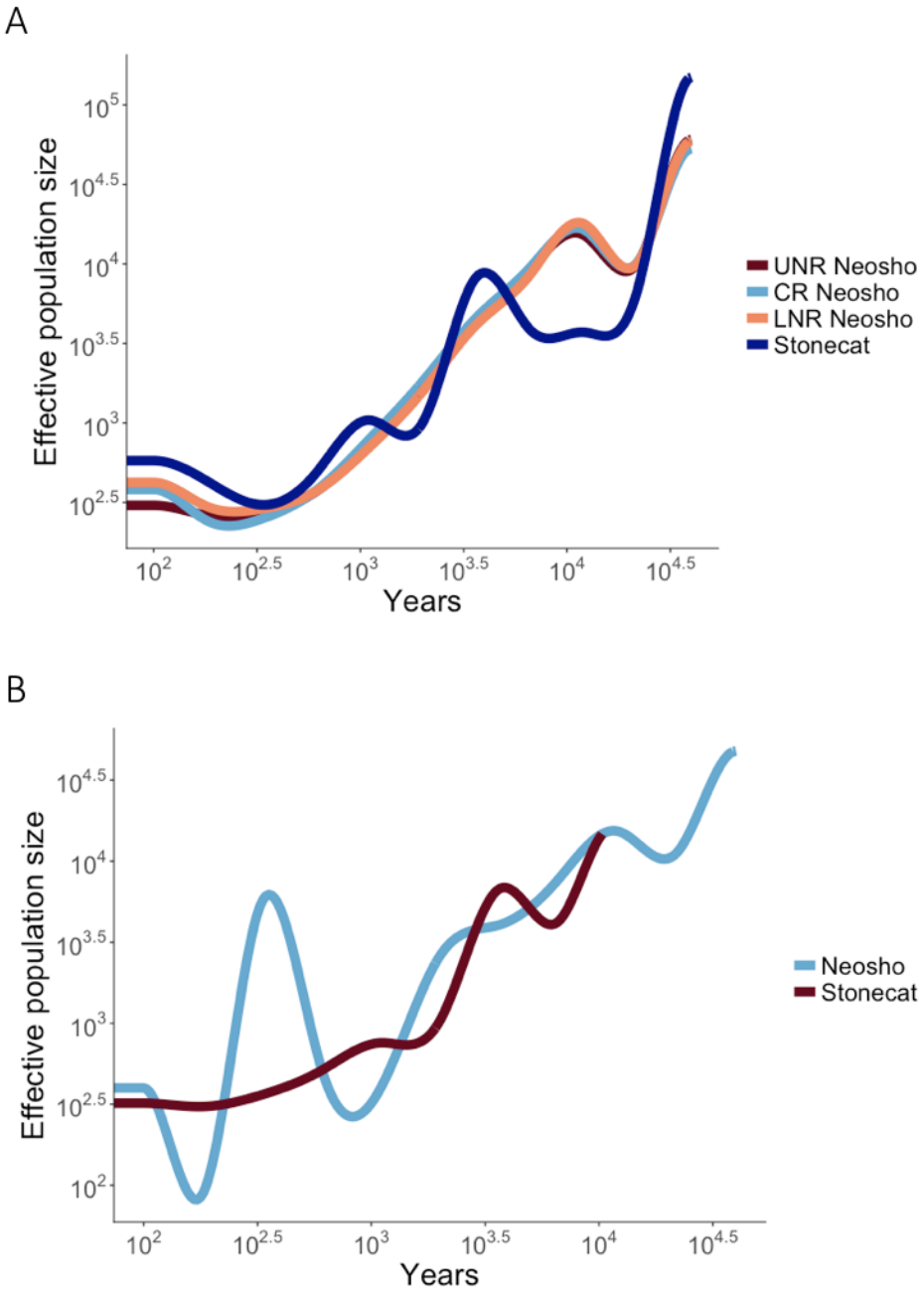

Figure 4.8. Estimation of historical effective population size of Neosho madtom and Stonecat (A) and the time to a clean split between the two species (B).

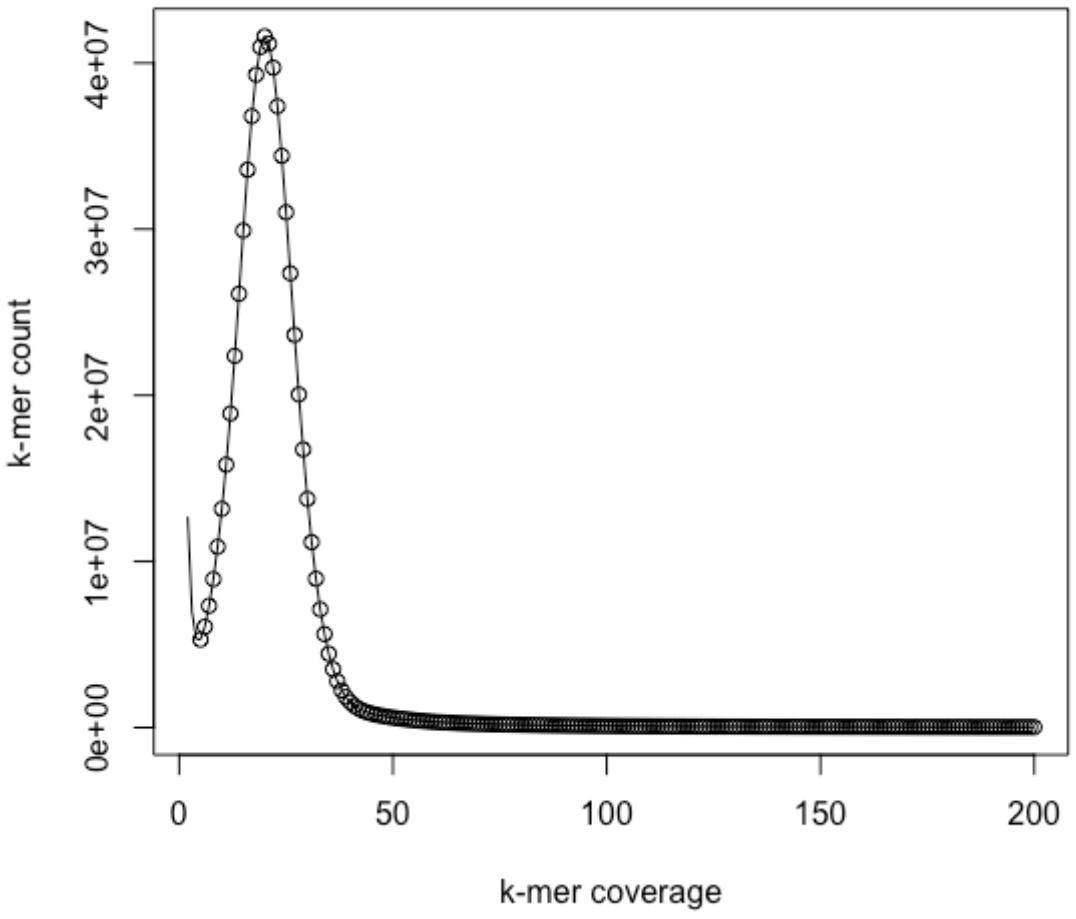

Figure 4.9. Distribution of *K*-mer frequencies from whole genome sequencing of one individual.

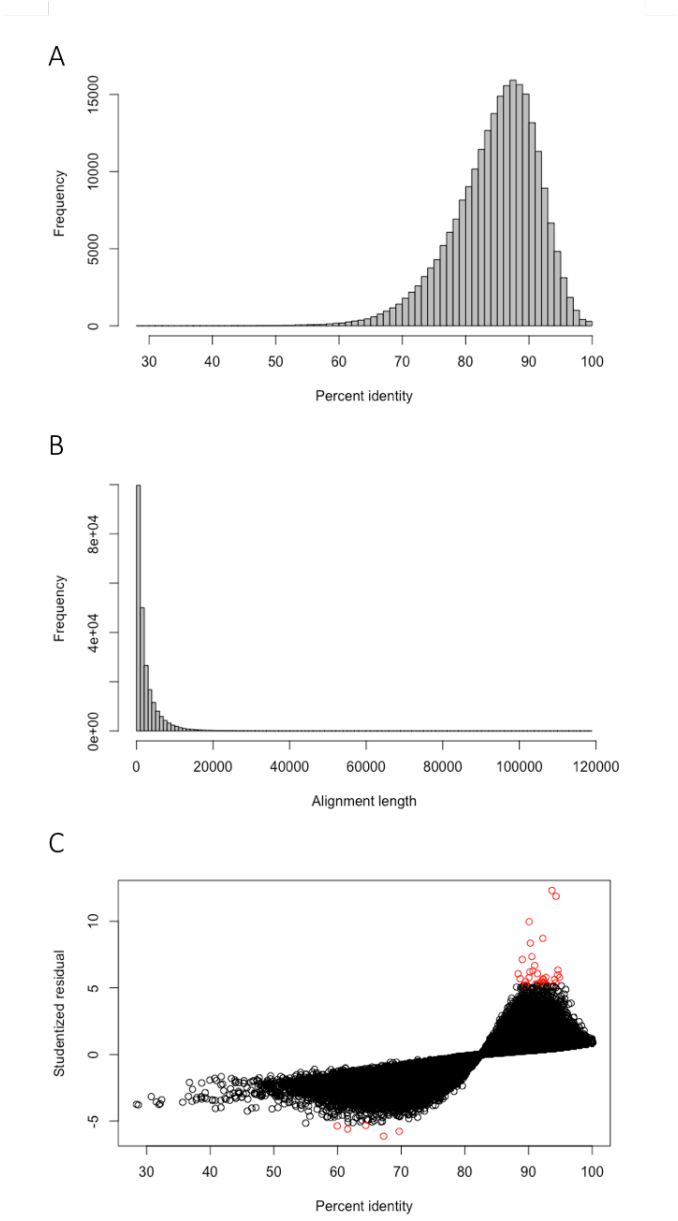

Figure 4.10. Statistics from alignments of Neosho madtom *de novo* assembled scaffolds to the channel catfish reference genome via NUCmer alignment. (A) Distribution of the percent identity of alignments. (B) Distribution of the length of alignments. (C) Distribution of Studentized residuals *versus* percent identity with significant alignments shown in red and non-significant alignments shown in black.

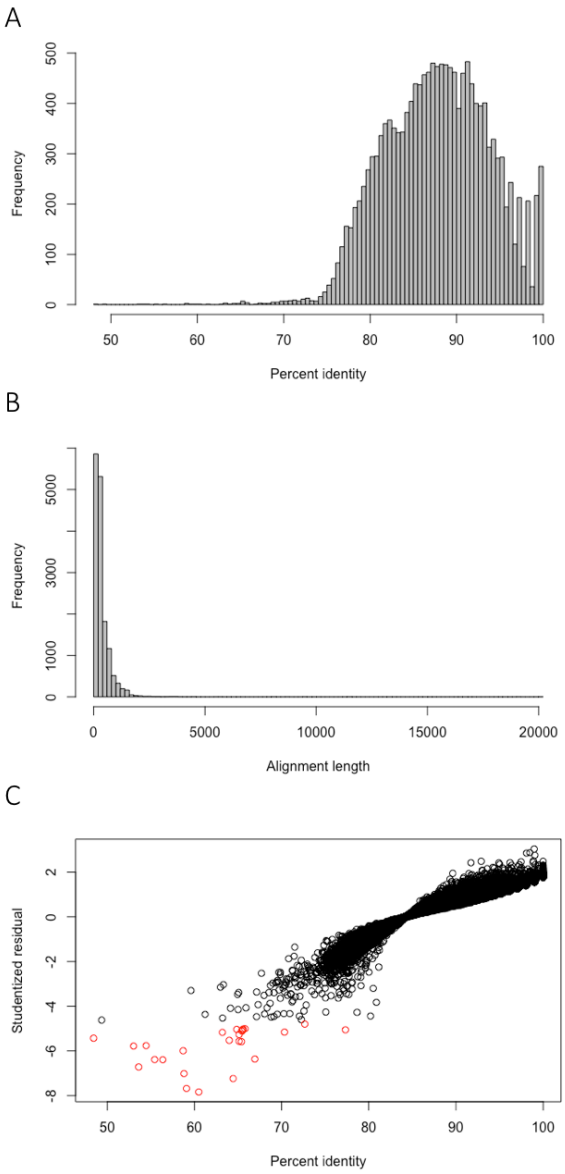

Figure 4.11. Statistics from alignments of Neosho madtom *de novo* scaffolds to the zebrafish reference genome via NUCmer alignment. (A) Distribution of the percent identity of alignments. (B) Distribution of the length of alignments. (C) Distribution of Studentized residuals *versus* percent identity with significant alignments shown in red and non-significant alignments shown in black.

Tables

Table 4.1. Sampling locations of Neosho madtom and Stonecat.

Fish ID	Species name	Common name	Sampling location	River	Population
84434	*Noturus placidus*	Neosho madtom	Americus, KS	Neosho River	UNR
84435	*Noturus placidus*	Neosho madtom	Cottonwood Falls, KS	Cottonwood River	CR
86398	*Noturus placidus*	Neosho madtom	Humboldt, KS	Neosho River	LNR
93670	*Noturus placidus*	Neosho madtom	Americus, KS	Neosho River	UNR
93671	*Noturus placidus*	Neosho madtom	Humboldt, KS	Neosho River	LNR
93672	*Noturus placidus*	Neosho madtom	Humboldt, KS	Neosho River	LNR
93674	*Noturus placidus*	Neosho madtom	Cottonwood Falls, KS	Cottonwood River	CR
93675	*Noturus placidus*	Neosho madtom	Humboldt, KS	Neosho River	LNR
93676	*Noturus flavus*	Stonecat	Grand River Dam, OK	Neosho River	Outgroup
93677	*Noturus placidus*	Neosho madtom	Cottonwood Falls, KS	Cottonwood River	CR
93678	*Noturus placidus*	Neosho madtom	Americus, KS	Neosho River	UNR

Table 4.2. Whole genome sequencing statistics for Neosho madtom and Stonecat.

Fish ID	Sample type	Library abbreviation	Mean insert size	Number of reads	Read length (bp)	Est. raw coverage*
84434	muscle	aa	271	256,484,994	100	25.65
84435	muscle	aa	224	328,701,858	100	32.87
86398	muscle	aa	245	274,665,938	100	27.47
93670	muscle	aa	336	432,376,174	100	43.24
93670	muscle	ba	406	381,863,838	250	
93670	muscle	ca	2000	187,522,278	100	18.75
93670	muscle	da	3000	215,913,390	100	21.59
93671	muscle	aa	337	455,916,970	100	45.59
93672	muscle	aa	320	475,853,862	100	47.59
93674	muscle	aa	336	394,229,316	100	39.42
93675	muscle	aa	320	484,838,720	100	48.48
93676	muscle	aa	339	450,976,652	100	45.10
93677	muscle	aa	334	401,332,076	100	40.13
93678	muscle	aa	326	318,100,540	100	31.81

*Based on genome size of 1 Gb

Table 4.3. Alignment statistics for Neosho madtom sequences to channel catfish reference.

Fish ID	Total Reads After QC	Total Reads Aligned	Proportion aligned
84434	233,456,403	195,384,483	0.8369
84435	559,694,518	458,145,137	0.8186
86398	248,653,169	202,410,073	0.8140
93670	350,505,554	281,634,420	0.8035
93671	361,475,118	288,446,421	0.7980
93672	363,505,237	289,009,895	0.7951
93674	304,190,856	241,617,882	0.7943
93675	348,672,910	273,244,224	0.7837
93676	356,455,627	286,450,417	0.8036
93677	328,383,024	258,760,008	0.7880
93678	278,475,624	227,856,740	0.8182

Table 4.4. Number of variable SNPs discovered from reference alignment.

	Neosho + Stonecat	**Neosho - all populations**	**Neosho - UNR**	**Neosho - CR**	**Neosho - LNR**
Number of variable SNPs	9,325,538	2,082,811	1,271,208	1,332,961	1,558,034

Table 4.5. Results from principal component analysis.

	With outgroup				Without outgroup			
Eigen-vector	**Eigen-value**	**Cumulative variance explained**	**Tracy-Widom statistic**	**Tracy-Widom *p*-value**	**Eigen-value**	**Cumulative variance explained**	**Tracy-Widom statistic**	**Tracy-Widom *p*-value**
1	8.92	89.21%	-0.1459	0.1962	1.22	13.56%	-3.728	0.9847
2	0.15	90.67%	--	--	1.16	26.47%	--	--
3	0.14	92.06%	--	--	1.03	37.87%	--	--
4	0.12	93.29%	--	--	0.97	48.63%	--	--
5	0.12	94.45%	--	--	0.95	59.18%	--	--
6	0.11	95.59%	--	--	0.94	69.63%	--	--
7	0.11	96.72%	--	--	0.92	79.87%	--	--
8	0.11	97.83%	--	--	0.92	90.08%	--	--
9	0.11	98.93%	--	--	0.89	100.00%	--	--
10	0.11	100.00%	--	--	--	--	--	--

Table 4.6. Statistics from whole genome *de novo* assembly for a single Neosho madtom.

Fish ID	Total bases assembled	Number of scaffolds	Scaffold N25	Scaffold N50	Scaffold N75	Contig N50
93670	899,438,561	56,807	217,732	108,346	43,999	6,113

Table 4.7. Significantly conserved regions between Neosho madtom and channel catfish genomes.

Scaffold ID	Mapped chromsome	Mapped start	Mapped end	Predicted gene symbol	Predicted gene description
jcf7180004898874	NC_030416.1	23465490	23478732	*LOC108270367*	plectin-like
jcf7180004891330	NC_030417.1	9636277	9720456	*HOXB2*	homeobox B2
jcf7180004895857	NC_030418.1	14759101	14819013	*BCL11A*	B-cell CLL/lymphoma 11A
jcf7180004894661	NC_030418.1	16152537	16174525	--	--
jcf7180004898871	NC_030418.1	16905534	16963603	*LOC108262760*	carbohydrate sulfotransferase 3-like
jcf7180004898871	NC_030418.1	16999920	17041288	*LOC108263455*	uncharacterized LOC108263455
jcf7180004893994	NC_030418.1	17375262	17397050	*ZMIZ1*	zinc finger MIZ-type containing 1
jcf7180004856844	NC_030421.1	12073004	12104205	*EPHA3*	EPH receptor a3
jcf7180004885960	NC_030421.1	13087483	13106594	--	--
jcf7180004885960	NC_030421.1	13106596	13146261	*SHOX*	short stature homeobox
jcf7180004891860	NC_030421.1	24943819	24979927	*TTN*	titin
jcf7180004895816	NC_030421.1	25038753	25101307	*TTN*	titin
jcf7180004895816	NC_030421.1	25189670	25264932	*LOC108266290*	titin-like
jcf7180004891534	NC_030423.1	14888561	14931104	*LOC108269150*	slit homolog 3 protein-like
jcf7180004893229	NC_030423.1	16176338	16201506	*LOC108268792*	microtubule-associated protein 1A-like
jcf7180004892919	NC_030424.1	16673261	16688020	*NPAS3*	neuronal PAS domain protein 3
jcf7180004892563	NC_030424.1	19651356	19770443	*C9H15ORF41*	chromosome 9 open reading frame, human C15orf41
jcf7180004892563	NC_030424.1	19800104	19819460	--	--
jcf7180004848114	NC_030426.1	10946402	10968214	*LOC108272157*	pre-B-cell leukemia transcription factor 1
jcf7180004893231	NC_030427.1	7815754	7828683	*LOC108273067*	zinc finger E-box-binding homeobox 2-like
jcf7180004897758	NC_030427.1	8865132	8904209	*LOC108273058*	multidrug resistance-associated protein 1-like
jcf7180004899247	NC_030428.1	13597775	13637447	*TCF7L2*	transcription factor 7 like 2
jcf7180004899234	NC_030428.1	18046471	18063008	*LOC108273946*	E3 ubiquitin-protein ligase rnf213-alpha-like
jcf7180004898327	NC_030429.1	15720970	15737720	*ZNF469*	zinc finger protein 469
jcf7180004891458	NC_030430.1	4354034	4430554	*HOXC10*	homeobox c10
jcf7180004894009	NC_030431.1	19506824	19528651	*LOC108277102*	uncharacterized LOC108277102

Scaffold ID	Mapped chromsome	Mapped start	Mapped end	Predicted gene symbol	Predicted gene description
jcf7180004894017	NC_030434.1	13639804	13718730	*FOXP2*	forkhead box P2
jcf7180004899882	NC_030436.1	1069656	1084510	*LOC108254756*	protein bassoon-like
jcf7180004892526	NC_030437.1	10017498	10062295	*MEF2C*	myocyte enhancer factor 2C
jcf7180004898966	NC_030439.1	12993991	13014604	*LOC108256925*	plectin-like
jcf7180004895339	NC_030442.1	7575835	7599426	*ZNF536*	zinc finger protein 536
jcf7180004895339	NC_030442.1	7663454	7715290	*ZNF536*	zinc finger protein 536

Table 4.8. Significantly diverged regions between Neosho madtom and channel catfish genomes.

Scaffold ID	Mapped chromsome	Mapped start	Mapped end	Predicted gene symbol	Predicted gene description
jcf7180004900536	NC_030419.1	31623390	31638329	*PLCH1*	phospholipase C eta 1
jcf7180004898244	NC_030422.1	27313758	27319548	*LOC108268315*	uncharacterized LOC108268315
jcf7180004894658	NC_030429.1	6397724	6402950	--	--
jcf7180004898264	NC_030440.1	2063138	2070585	--	--
jcf7180004898679	NC_030444.1	10984669	11002470	*PTPRD*	protein tyrosine phosphatase, receptor type D

Table 4.9. Significantly diverged regions between Neosho madtom and zebrafish genomes.

Scaffold ID	Mapped chromosome	Mapped start	Mapped end	Predicted gene symbol	Predicted gene description
jcf7180004899048	NC_007112.6	20740780	20742374	*GLRBA*	glycine receptor, beta a
jcf7180004897174	NC_007112.6	20740780	20742374	*GLRBA*	glycine receptor, beta a
jcf7180004894518	NC_007112.6	20740780	20742374	*GLRBA*	glycine receptor, beta a
jcf7180004897753	NC_007112.6	20740780	20742374	*GLRBA*	glycine receptor, beta a
jcf7180004897462	NC_007112.6	20740780	20742374	*GLRBA*	glycine receptor, beta a
jcf7180004896630	NC_007112.6	20740803	20742457	*GLRBA*	glycine receptor, beta a
jcf7180004892430	NC_007112.6	20740903	20742374	*GLRBA*	glycine receptor, beta a
jcf7180004890896	NC_007112.6	20740972	20742356	*GLRBA*	glycine receptor, beta a
jcf7180004893596	NC_007116.6	51063024	51063710	--	--
jcf7180004893231	NC_007117.6	1217912	1218485	--	--
jcf7180004886895	NC_007117.6	59888356	59888933	*CASKB*	calcium/calmodulin-dependent serine protein kinase b
jcf7180004862073	NC_007117.6	59888552	59888933	*CASKB*	calcium/calmodulin-dependent serine protein kinase b
jcf7180004895339	NC_007118.6	46204173	46206247	*ZNF536*	zinc finger protein 536
jcf7180004898913	NC_007119.6	39272299	39273393	--	--
jcf7180004895816	NC_007120.6	42946908	42955608	*TTNB*	titin b
jcf7180004898444	NC_007123.6	16990431	16991353	*ACTA2*	actin, alpha 2, smooth muscle, aorta
jcf7180004892010	NC_007123.6	19161775	19163832	*TNRC6B*	trinucleotide repeat containing 6b
jcf7180004897698	NC_007124.6	19196475	19198068	--	--
jcf7180004899253	NC_007125.6	32299200	32301778	--	--
jcf7180004858513	NC_007127.6	54861583	54862015	*RGS22*	regulator of G-protein signaling 22
jcf7180004891058	NC_007130.6	41878818	41881131	*DLX6A*	distal-less homeobox 6a
jcf7180004873001	NC_007130.6	48403460	48404074	*ZGC:85936*	zgc:85936
jcf7180004897713	NC_007132.6	11400114	11401969	*GRIN1A*	glutamate

Scaffold ID	Mapped chromosome	Mapped start	Mapped end	Predicted gene symbol	Predicted gene description
jcf7180004894315	NC_007132.6	39525089	39526430	*SRSF1B*	receptor, ionotropic, N-methyl D-aspartate 1a serine/arginine -rich splicing factor 1b
jcf7180004895067	NC_007134.6	45834581	45835634	*CYP17A2*	cytochrome P450, family 17, subfamily A, polypeptide 2
jcf7180004872030	NC_007135.6	35154885	35155908	--	--

Supplementary Materials

Supplementary Note 1: De novo variant calling

Cortex var [156] was used for *de novo* variant calling. The program was run twice, once with all Neosho madtom and Stonecat and a second time with only the Neosho madtom. Interestingly, the algorithm's performance was improved when only the Neosho madtom were included. Quality was measured using genotype confidence scores, which are calculated as the log probability of the maximum likelihood estimate of genotype minus the log probability of the second most likely genotype. As described in the methods, a high confidence set of genotypes was defined as having a genotype confidence score > 5.54, which indicates that the genotype called is $e^{5.54}$, or 254.5, times more likely than the alternative.

When both Neosho madtom and Stonecat were included in the *de novo* variant calling process, 1,792,687 SNPs were detected. However, 673,452 of these had low genotype confidence scores and were removed from subsequent analyses. Of the remaining 1,119,235 SNPs, only 59,787 were variable in Neosho madtom. Thus, most SNPs with high confidence scores represented putatively fixed differences between Neosho madtom and Stonecat. Utilizing the Cortex var algorithm with only Neosho madtom resulted in additional variable sites within the species being discovered (967,681 SNPs before genotype confidence filtering and 709,790 SNPs after filtering) in Neosho madtom, and substantially less variants having low genotype confidence scores. For this reason, we used *de novo* variants called from the analysis of Neosho madtom without Stonecat whenever possible (i.e. when the Stonecat was not included in the analysis under consideration).

REFERENCES

1. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. 1977. Biotechnology. 1992;24: 104–8.

2. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. Nature Publishing Group; 2008;26: 1135–1145.

3. Reinert K, Langmead B, Weese D, Evers DJ. Alignment of Next-Generation Sequencing Reads. Annu Rev Genomics Hum Genet. 2015;16: 133–51. doi:10.1146/annurev-genom-090413-025358

4. Nagarajan N, Pop M. Sequence assembly demystified. Nat Rev Genet. Nature Publishing Group; 2013;14: 157–167. doi:10.1038/nrg3367

5. Sohn J, Nam J-W. The present and future of de novo whole-genome assembly. Brief Bioinform. 2016; 1–18. doi:10.1093/bib/bbw096

6. Ekblom R, Wolf JBW. A field guide to whole-genome sequencing, assembly and annotation. Evol Appl. 2014;7: 1026–1042. doi:10.1111/eva.12178

7. Vezzi F, Narzisi G, Mishra B, Nagarajan N, Pop M, Vezzi F, et al. Reevaluating Assembly Evaluations with Feature Response Curves: GAGE and Assemblathons. Rzhetsky A, editor. PLoS One. Public Library of Science; 2012;7: e52210. doi:10.1371/journal.pone.0052210

8. Chain PSG, Grafham D V., Fulton RS, FitzGerald MG, Hostetler J, Muzny D, et al. Genome Project Standards in a New Era of Sequencing. Science (80-).

2009;326: 236–237. doi:10.1126/science.1180614

9. Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. REAPR: a universal tool for genome assembly evaluation. Genome Biol. 2013;14: R47. doi:10.1186/gb-2013-14-5-r47

10. Narzisi G, Mishra B. Comparing de novo genome assembly: the long and short of it. Aerts S, editor. PLoS One. Public Library of Science; 2011;6: e19175. doi:10.1371/journal.pone.0019175

11. Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. Genomics. 2010;95: 315–327. doi:10.1016/j.ygeno.2010.03.001

12. Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nat Methods. Nature Publishing Group; 2009;6: 291–295. doi:10.1038/nmeth.1311

13. Compeau PEC, Pevzner PA, Tesler G. How to apply de Bruijn graphs to genome assembly. Nat Biotechnol. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2011;29: 987–91. doi:10.1038/nbt.2023

14. Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. Genomics. Elsevier Inc.; 2010;95: 315–27. doi:10.1016/j.ygeno.2010.03.001

15. Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, et al. Library construction for next-generation sequencing: overviews and

challenges. Biotechniques. NIH Public Access; 2014;56: 61–4, 66, 68, passim. doi:10.2144/000114133

16. Roe BA. Shotgun Library Construction for DNA Sequencing. Methods in Molecular Biology. New Jersey: Humana Press; 2004. pp. 171–188. doi:10.1385/1-59259-752-1:171

17. Quail M, Smith ME, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. BMC Genomics. 2012;13: 341. doi:10.1186/1471-2164-13-341

18. Collins FS, Weissmant SM. Directional cloning of DNA fragments at a large distance from an initial probe: A circularization method. Proc Natl Acad Sci. 1984;81: 6812–6816.

19. Poustka A, Pohl TM, Barlow DP, Frischauf A-M, Lehrach H. Construction and use of human chromosome jumping libraries from NotI-digested DNA. Nature. Nature Publishing Group; 1987;325: 353–355. doi:10.1038/325353a0

20. Human Genome Sequencing Consortium I. Finishing the euchromatic sequence of the human genome. Nature. Nature Publishing Group; 2004;431: 931–945. doi:10.1038/nature03001

21. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, et al. Fine-scale structural variation of the human genome. Nat Genet Publ online 15 May 2005; | doi101038/ng1562. Nature Publishing Group; 2005;37: 727. doi:10.1038/ng1562

22. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics. 2015;13: 278–289. doi:10.1016/j.gpb.2015.08.002

23. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One. 2012;7: e47768. doi:10.1371/journal.pone.0047768

24. Fuller CW, Middendorf LR, Benner SA, Church GM, Harris T, Huang X, et al. The challenges of sequencing by synthesis. Nat Biotechnol. 2009;27: 1013–1023. doi:10.1038/nbt.1585

25. Marçais G, Yorke JA, Zimin A. QuorUM: an error corrector for Illumina reads. arXiv.org. 2013;

26. Paszkiewicz K, Studholme DJ. De novo assembly of short sequence reads. Brief Bioinform. Oxford University Press; 2010;11: 457–472. doi:10.1093/bib/bbq020

27. Batzer MA, Deininger PL. Alu repeats and human genomic diversity. Nat Rev Genet. 2002;3: 370–379. doi:10.1038/nrg798

28. Schmid CW, Deininger PL. Sequence organization of the human genome. Cell. 1975;6: 345–58.

29. Pop M, Salzberg SL. Bioinformatics challenges of new sequencing technology. Trends Genet. 2008;24: 142–149. doi:10.1016/j.tig.2007.12.006

30. Phillippy AM, Schatz MC, Pop M. Genome assembly forensics: finding the elusive mis-assembly. Genome Biol. BioMed Central; 2008;9: R55. doi:10.1186/gb-2008-9-3-r55

31. Taylor JF, Whitacre LK, Hoff JL, Tizioto PC, Kim J, Decker JE, et al. Lessons for livestock genomics from genome and transcriptome sequencing in cattle and other mammals. Genet Sel Evol. 2016;48: 59. doi:10.1186/s12711-016-0237-6

32. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, et al. Characterizing and measuring bias in sequence data. Genome Biol. BioMed Central; 2013;14: R51. doi:10.1186/gb-2013-14-5-r51

33. Desai A, Marwah VS, Yadav A, Jha V, Dhaygude K, Bangar U, et al. Identification of Optimum Sequencing Depth Especially for De Novo Genome Assembly of Small Genomes Using Next Generation Sequencing Data. Zhang S-D, editor. PLoS One. Public Library of Science; 2013;8: e60204. doi:10.1371/journal.pone.0060204

34. Haiminen N, Kuhn DN, Parida L, Rigoutsos I, Lipman D. Evaluation of Methods for De Novo Genome Assembly from High-Throughput Sequencing Reads Reveals Dependencies That Affect the Quality of the Results. Rzhetsky A, editor. PLoS One. Public Library of Science; 2011;6: e24182. doi:10.1371/journal.pone.0024182

35. Whitacre LK, Tizioto PC, Kim J, Sonstegard TS, Schroeder SG, Alexander LJ, et al. What's in your next-generation sequence data? An exploration of unmapped DNA and RNA sequence reads from the bovine reference individual. BMC Genomics. BioMed Central; 2015;16: 1114. doi:10.1186/s12864-015-2313-7

36. Whitacre LK, Hoff JL, Schnabel RD, Albarella S, Ciotola F, Peretti V, et al. Elucidating the genetic basis of an oligogenic birth defect using whole genome

sequence data in a non-model organism, Bubalus bubalis. Sci Rep. Nature Publishing Group; 2017;7: 39719. doi:10.1038/srep39719

37. Whitacre LK, Wildhaber ML, Schnabel RD, Johnson GS, Downs JM, Mutangadura-Mhlanga T, et al. Genome-wide variation and population structure in Neosho madtom catfish. Prep. 2017;

38. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. Nature Publishing Group; 2010;11: 31–46.

39. Bhaduri A, Qu K, Lee CS, Ungewickell A, Khavari PA. Rapid identification of non-human sequences in high-throughput sequencing datasets. Bioinformatics. 2012;28: 1174–5. doi:10.1093/bioinformatics/bts100

40. Kostic AD, Ojesina AI, Pedamallu CS, Jung J, Verhaak RGW, Getz G, et al. PathSeq: software to identify or discover microbes by deep sequencing of human tissue. Nat Biotechnol. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2011;29: 393–6. doi:10.1038/nbt.1868

41. Isakov O, Modai S, Shomron N. Pathogen detection using short-RNA deep sequencing subtraction and assembly. Bioinformatics. 2011;27: 2027–30. doi:10.1093/bioinformatics/btr349

42. Tae H, Karunasena E, Bavarva JH, McIver LJ, Garner HR. Large scale comparison of non-human sequences in human sequencing data. Genomics. 2014;104: 453–8. doi:10.1016/j.ygeno.2014.08.009

43. Gouin A, Legeai F, Nouhaud P, Whibley A, Simon J-C, Lemaitre C. Whole-

genome re-sequencing of non-model organisms: lessons from unmapped reads. Heredity (Edinb). 2014; doi:10.1038/hdy.2014.85

44. Zimin A V, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, et al. A whole-genome assembly of the domestic cow, Bos taurus. Genome Biol. 2009;10: R42. doi:10.1186/gb-2009-10-4-r42

45. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2015. Nucleic Acids Res. 2014;43: D662-669. doi:10.1093/nar/gku1010

46. Merchant S, Wood DE, Salzberg SL. Unexpected cross-species contamination in genome sequencing projects. PeerJ. 2014;2: e675. doi:10.7717/peerj.675

47. Eberhard ML, Stilesi OFO, Eberhardt ML. Studies on the Onchocerca (Nematoda: Filarioidea) Found in Cattle in the United States. I. Systematics of O. gutturosa O. lienalis with a Description of O. stilesi sp. n. J Parasitol. 1979;65: 379–388.

48. Gill LL, Hardman N, Chappell L, Hu Qu L, Nicoloso M, Bachellerie J-P. Phylogeny of Onchocerca volvulus and related species deduced from rRNA sequence comparisons. Mol Biochem Parasitol. 1988;28: 69–76. doi:10.1016/0166-6851(88)90182-X

49. Casiraghi M, Anderson TJC, Bandi C, Bazzocchi C, Genchi C. A phylogenetic analysis of filarial nematodes: comparison with the phylogeny of Wolbachia endosymbionts. Parasitology. Cambridge University Press; 2001;122: 93–103. doi:10.1017/S0031182000007149

50. Casiraghi M, Bain O, Guerrero R, Martin C, Pocacqua V, Gardner SL, et al.

Mapping the presence of Wolbachia pipientis on the phylogeny of filarial nematodes: evidence for symbiont loss during evolution. Int J Parasitol. 2004;34: 191–203. doi:10.1016/j.ijpara.2003.10.004

51. Xie H, Bain O, Williams SA. Molecular phylogenetic studies on filarial parasites based on 5S ribosomal spacer sequences. Parasite. EDP Sciences; 2014;1: 141–151. doi:10.1051/parasite/1994012141

52. Krueger A, Fischer P, Morales-Hojas R. Molecular phylogeny of the filaria genus Onchocerca with special emphasis on Afrotropical human and bovine parasites. Acta Trop. 2007;101: 1–14. doi:10.1016/j.actatropica.2006.11.004

53. Garofalo A, Kläger SL, Rowlinson M-C, Nirmalan N, Klion A, Allen JE, et al. The FAR proteins of filarial nematodes: secretion, glycosylation and lipid binding characteristics. Mol Biochem Parasitol. 2002;122: 161–170. doi:10.1016/S0166-6851(02)00097-X

54. Morales-Hojas R, Cheke RA, Post RJ. Molecular systematics of five Onchocerca species (Nematoda: Filarioidea) including the human parasite, O. volvulus, suggest sympatric speciation. J Helminthol. Cambridge University Press; 2006;80: 281–290. doi:10.1079/JOH2006331

55. Morales-Hojas R, Cheke RA, Post RJ. A preliminary analysis of the population genetics and molecular phylogenetics of Onchocerca volvulus (Nematoda: Filarioidea) using nuclear ribosomal second internal transcribed spacer sequences. Mem Inst Oswaldo Cruz. Fundação Oswaldo Cruz; 2007;102: 879–882. doi:10.1590/S0074-02762007005000114

56. Kulke D, von Samson-Himmelstjerna G, Miltsch SM, Wolstenholme AJ, Jex AR, Gasser RB, et al. Characterization of the Ca2+-gated and voltage-dependent K+-channel Slo-1 of nematodes and its interaction with emodepside. PLoS Negl Trop Dis. Public Library of Science; 2014;8: e3401. doi:10.1371/journal.pntd.0003401

57. Bock R, Jackson L, de Vos A, Jorgensen W. Babesiosis of cattle. Parasitology. 2004;129 Suppl: S247-69.

58. Altay K, Aydin MF, Dumanli N, Aktas M. Molecular detection of Theileria and Babesia infections in cattle. Vet Parasitol. 2008;158: 295–301. doi:10.1016/j.vetpar.2008.09.025

59. Terkawi MA, Alhasan H, Huyen NX, Sabagh A, Awier K, Cao S, et al. Molecular and serological prevalence of Babesia bovis and Babesia bigemina in cattle from central region of Syria. Vet Parasitol. 2012;187: 307–311. doi:10.1016/j.vetpar.2011.12.038

60. Simking P, Saengow S, Bangphoomi K, Sarataphan N, Wongnarkpet S, Inpankaew T, et al. The molecular prevalence and MSA-2b gene-based genetic diversity of Babesia bovis in dairy cattle in Thailand. Vet Parasitol. 2013;197: 642–648. doi:10.1016/j.vetpar.2013.07.015

61. Corwin RM. Economics of gastrointestinal parasitism of cattle. Vet Parasitol. 1997;72: 451-7-60.

62. Hawkins JA. Economic benefits of parasite control in cattle. Vet Parasitol. 1993;46: 159–173. doi:10.1016/0304-4017(93)90056-S

63. Gunn A, Irvine RJ. Subclinical parasitism and ruminant foraging strategies-a review. Wildl Soc Bull. 2003;31: 117–126.

64. Dunning Hotopp JC. Horizontal gene transfer between bacteria and animals. Trends Genet. 2011;27: 157–63. doi:10.1016/j.tig.2011.01.005

65. Syvanen M. Evolutionary implications of horizontal gene transfer. Annu Rev Genet. Annual Reviews; 2012;46: 341–58. doi:10.1146/annurev-genet-110711-155529

66. Crisp A, Boschetti C, Perry M, Tunnacliffe A, Micklem G. Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. Genome Biol. BioMed Central Ltd; 2015;16: 50. doi:10.1186/s13059-015-0607-3

67. Mittra I, Khare NK, Raghuram GV, Chaubal R, Khambatti F, Gupta D, et al. Circulating nucleic acids damage DNA of healthy cells by integrating into their genomes. J Biosci. 2015;40: 91–111.

68. Chapple RH, Tizioto PC, Wells KD, Givan SA, Kim J, McKay SD, et al. Characterization of the rat developmental liver transcriptome. Physiol Genomics. 2013;45: 301–11. doi:10.1152/physiolgenomics.00128.2012

69. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25: 1105–11. doi:10.1093/bioinformatics/btp120

70. Zimin A, Marcais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome Assembler. Bioinformatics. 2013; btt476-.

doi:10.1093/bioinformatics/btt476

71. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8: 1494–512. doi:10.1038/nprot.2013.084

72. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215: 403–10. doi:10.1016/S0022-2836(05)80360-2

73. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10: 421. doi:10.1186/1471-2105-10-421

74. Mudunuri U, Che A, Yi M, Stephens RM. bioDBnet: the biological database network. Bioinformatics. 2009;25: 555–6. doi:10.1093/bioinformatics/btn654

75. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409: 860–921. doi:10.1038/35057062

76. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 2001;291: 1304–51. doi:10.1126/science.1058040

77. Florea L, Souvorov A, Kalbfleisch TS, Salzberg SL. Genome assembly has a major impact on gene content: a comparison of annotation in two Bos taurus assemblies. PLoS One. 2011;6: e21400. doi:10.1371/journal.pone.0021400

78. Kelley DR, Salzberg SL. Detection and correction of false segmental duplications caused by genome mis-assembly. Genome Biol. 2010;11: R28. doi:10.1186/gb-2010-11-3-r28

79. Zimin A V, Kelley DR, Roberts M, Marçais G, Salzberg SL, Yorke JA. Mis-assembled "segmental duplications" in two versions of the Bos taurus genome. PLoS One. 2012;7: e42680. doi:10.1371/journal.pone.0042680

80. Salzberg SL, Yorke JA. Beware of mis-assembled genomes. Bioinformatics. 2005;21: 4320–1. doi:10.1093/bioinformatics/bti769

81. Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, et al. Modernizing reference genome assemblies. PLoS Biol. 2011;9: e1001091. doi:10.1371/journal.pbio.1001091

82. Church DM, Schneider VA, Steinberg KM, Schatz MC, Quinlan AR, Chin C-S, et al. Extending reference assembly models. Genome Biol. 2015;16: 13. doi:10.1186/s13059-015-0587-3

83. Genovese G, Handsaker RE, Li H, Kenny EE, McCarroll SA. Mapping the human reference genome's missing sequence by three-way admixture in Latino genomes. Am J Hum Genet. 2013;93: 411–21. doi:10.1016/j.ajhg.2013.07.002

84. Cockrill WR. The water buffalo: a review. Br Vet J. 1981;137: 8–16.

85. Kierstein G, Vallinoto M, Silva A, Schneider MP, Iannuzzi L, Brenig B. Analysis of mitochondrial D-loop region casts new light on domestic water buffalo (Bubalus bubalis) phylogeny. Mol Phylogenet Evol. 2004;30: 308–324.

doi:10.1016/S1055-7903(03)00221-5

86. Zicarelli L. Enhancing reproductive performance in domestic dairy water buffalo (Bubalus bubalis). Soc Reprod Fertil Suppl. 2010;67: 443–55.

87. Michelizzi VN, Dodson M V, Pan Z, Amaral MEJ, Michal JJ, McLean DJ, et al. Water buffalo genome science comes of age. Int J Biol Sci. 2010;6: 333–49.

88. Vegad JL, Swamy M. A Textbook of Veterinary Systemic Pathology. 2nd Editio. India: IBDC Publishers; 2010.

89. Peretti V, Ciotola F, Albarella S, Restucci B, Meomartino L, Ferretti L, et al. Increased SCE levels in Mediterranean Italian buffaloes affected by limb malformation (transversal hemimelia). Cytogenet Genome Res. Karger Publishers; 2008;120: 183–7. doi:10.1159/000118761

90. Albarella S, Ciotola F, Dario C, Iannuzzi L, Barbieri V, Peretti V. Chromosome instability in Mediterranean Italian buffaloes affected by limb malformation (transversal hemimelia). Mutagenesis. 2009;24: 471–474. doi:10.1093/mutage/gep030

91. Taylor JF. Using sequencing data to localise developmental mutations. Plant and Animal Genome XXIII Conference. 2015.

92. Radiological Findings in Three Cases of Paraxial Radial Hemimelia in Goats [Internet]. [cited 8 May 2015]. Available: https://www.jstage.jst.go.jp/article/jvms/64/9/64_9_843/_pdf

93. Allen JG, Fenny RE, Buckman PG, Hunt BR, Morcombe PW. Hemimelia in

lambs. Aust Vet J. 1983;60: 283–4.

94. Lapointe J-M, Lachance S, Steffen DJ. Tibial Hemimelia, Meningocele, and Abdominal Hernia in Shorthorn Cattle. Vet Pathol. 2000;37: 508–511. doi:10.1354/vp.37-5-508

95. Alonso RA, Hernández A, Díaz P, Cantú JM. An autosomal recessive form of hemimelia in dogs. Vet Rec. 1982;110: 128–9.

96. Lockwood A, Montgomery R, McEwen V. Bilateral radial hemimelia, polydactyly and cardiomegaly in two cats. Vet Comp Orthop Traumatol. 2009;22: 511–3. doi:10.3415/VCOT-08-12-0124

97. Kochhar DM. Skeletal morphogenesis: comparative effects of a mutant gene and a teratogen. Prog Clin Biol Res. 1985;171: 267–81.

98. McKay M, Clarren SK, Zorn R. Isolated tibial hemimelia in sibs: an autosomal-recessive disorder? Am J Med Genet. 1984;17: 603–7. doi:10.1002/ajmg.1320170308

99. Niemann S, Zhao C, Pascu F, Stahl U, Aulepp U, Niswander L, et al. Homozygous WNT3 mutation causes tetra-amelia in a large consanguineous family. Am J Hum Genet. 2004;74: 558–63. doi:10.1086/382196

100. Parr BA, Avery EJ, Cygan JA, McMahon AP. The classical mouse mutant postaxial hemimelia results from a mutation in the Wnt 7a gene. Dev Biol. 1998;202: 228–34. doi:10.1006/dbio.1998.9007

101. Schüle B, Oviedo A, Johnston K, Pai S, Francke U. Inactivating mutations in

ESCO2 cause SC phocomelia and Roberts syndrome: no phenotype-genotype correlation. Am J Hum Genet. 2005;77: 1117–28. doi:10.1086/498695

102. Indra AK, Dupé V, Bornert J-M, Messaddeq N, Yaniv M, Mark M, et al. Temporally controlled targeted somatic mutagenesis in embryonic surface ectoderm and fetal epidermal keratinocytes unveils two distinct developmental functions of BRG1 in limb morphogenesis and skin barrier formation. Development. 2005;132: 4533–44. doi:10.1242/dev.02019

103. Chiang C, Litingtung Y, Harris MP, Simandl BK, Li Y, Beachy PA, et al. Manifestation of the limb prepattern: limb development in the absence of sonic hedgehog function. Dev Biol. 2001;236: 421–35. doi:10.1006/dbio.2001.0346

104. Chen H, Johnson RL. Interactions between dorsal-ventral patterning genes lmx1b, engrailed-1 and wnt-7a in the vertebrate limb. Int J Dev Biol. 2002;46: 937–41.

105. Charlesworth D, Willis JH. The genetics of inbreeding depression. Nat Rev Genet. Nature Publishing Group; 2009;10: 783–796. doi:10.1038/nrg2664

106. Lander ES, Botstein D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science. 1987;236: 1567–1570. doi:10.1126/science.2884728

107. Purcell S, Cherny SS, Sham PC. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics. 2003;19: 149–50.

108. Attanasio C, Nord AS, Zhu Y, Blow MJ, Biddie SC, Mendenhall EM, et al.

Tissue-specific SMARCA4 binding at active and repressed regulatory elements during embryogenesis. Genome Res. 2014;24: 920–9. doi:10.1101/gr.168930.113

109. Bultman S, Gebuhr T, Yee D, La Mantia C, Nicholson J, Gilliam A, et al. A Brg1 Null Mutation in the Mouse Reveals Functional Differences among Mammalian SWI/SNF Complexes. Mol Cell. 2000;6: 1287–1295. doi:10.1016/S1097-2765(00)00127-1

110. Wang J, Sinha T, Wynshaw-Boris A. Wnt signaling in mammalian development: lessons from mouse genetics. Cold Spring Harb Perspect Biol. Cold Spring Harbor Laboratory Press; 2012;4: a007963. doi:10.1101/cshperspect.a007963

111. Parr BA, McMahon AP. Dorsalizing signal Wnt-7a required for normal polarity of D-V and A-P axes of mouse limb. Nature. 1995;374: 350–3. doi:10.1038/374350a0

112. Woods CG, Stricker S, Seemann P, Stern R, Cox J, Sherridan E, et al. Mutations in WNT7A cause a range of limb malformations, including Fuhrmann syndrome and Al-Awadi/Raas-Rothschild/Schinzel phocomelia syndrome. Am J Hum Genet. Elsevier; 2006;79: 402–8. doi:10.1086/506332

113. Lohnes D, Mark M, Mendelsohn C, Dolle P, Dierich A, Gorry P, et al. Function of the retinoic acid receptors (RARs) during development (I). Craniofacial and skeletal abnormalities in RAR double mutants. Development. 1994;120: 2723–2748.

114. Maden M. Retinoic acid in development and regeneration. J Biosci. 1996;21: 299–

312. doi:10.1007/BF02703090

115. Abu-Hijleh G, Padmanabhan R. Retinoic acid-induced abnormal development of hindlimb joints in the mouse. Eur J Morphol. 1997;35: 327–36.

116. Zhao X, Sirbu IO, Mic FA, Molotkova N, Molotkov A, Kumar S, et al. Retinoic acid promotes limb induction through effects on body axis extension but is unnecessary for limb patterning. Curr Biol. 2009;19: 1050–7. doi:10.1016/j.cub.2009.04.059

117. Francis JC, Radtke F, Logan MPO. Notch1 signals through Jagged2 to regulate apoptosis in the apical ectodermal ridge of the developing limb bud. Dev Dyn. 2005;234: 1006–15. doi:10.1002/dvdy.20590

118. Pan Y, Liu Z, Shen J, Kopan R. Notch1 and 2 cooperate in limb ectoderm to receive an early Jagged2 signal regulating interdigital apoptosis. Dev Biol. 2005;286: 472–82. doi:10.1016/j.ydbio.2005.08.037

119. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. Public Library of Science; 2006;2: e190. doi:10.1371/journal.pgen.0020190

120. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30: 2114–20. doi:10.1093/bioinformatics/btu170

121. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25: 1754–60. doi:10.1093/bioinformatics/btp324

122. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al.

The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20: 1297–303. doi:10.1101/gr.107524.110

123. DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2011;43: 491–8. doi:10.1038/ng.806

124. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Bateman A, Pearson WR, Stein LD, Stormo GD, Yates JR, editors. Curr Protoc Bioinforma. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2013;11: 11.10.1-11.10.33. doi:10.1002/0471250953

125. Venables WN, Ripley BD. Modern Applied Statistics with S. New York, NY: Springer New York; 2002. doi:10.1007/978-0-387-21706-2

126. Storey JD, Bass AJ, Dabney A, Robinson D. qvalue: Q-value estimation for false discovery rate control. 2015.

127. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012;44: 821–4. doi:10.1038/ng.2310

128. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. BioMed Central; 2004;5: R12. doi:10.1186/gb-2004-5-2-r12

129. Elsik CG, Unni DR, Diesh CM, Tayal A, Emery ML, Nguyen HN, et al. Bovine Genome Database: new tools for gleaning function from the Bos taurus genome. Nucleic Acids Res. 2016;44: D834-9. doi:10.1093/nar/gkv1077

130. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res. 2010;38: W214-20. doi:10.1093/nar/gkq537

131. Reimand J, Kull M, Peterson H, Hansen J, Vilo J. G:Profiler-a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic Acids Res. 2007;35. doi:10.1093/nar/gkm226

132. Reimand J, Kolde R, Arak T. gProfileR: Interface to the "g:Profiler" Toolkit. 2016.

133. Luttrell GR, Larson RD, Stark WJ, Ashbaugh NA, Echelle AA, Zale A V. Status and Distribution of the Neosho Madtom (Noturus placidus) in Oklahoma. Proc Okla Acad Sci. 1992;72: 5–6.

134. Wilkinson C, Edds D, Dorlac J, Wildhaber ML, Schmitt CJ, Allert A. Neosho madtom distribution and abundance in the Spring River on JSTOR. Southwest Nat. 1996;41: 78–81.

135. Taylor WR. A revision of the catfish genus Noturus Rafinesque, with an analysis of higher groups in the Ictaluridae. U.S. Museu. Washington, D.C.: Smithsonian Institution Press; 1969.

136. Wildhaber ML. The Neosho madtom (Noturus placidus) and the multifaceted nature of population limiting factors. Catfish 2010 Conserv Ecol Manag Worldw Catfish Popul Habitats. 2011;77: 281–294.

137. Wildhaber. Neosho Madtom, Noturus placidus. Kansas Fishes Committee. Lawrence: University Press of Kansas; 2014. pp. 303–305.

138. Fuselier L, Edds D. Seasonal Variation in Habitat Use by the Neosho Madtom (Teleostei: Ictaluridae: Noturus placidus). Southwest Nat. 1994;39: 217. doi:10.2307/3671585

139. Wenke TL, Eberle ME, Ernsting GW, Stark WJ. Winter Collections of the Neosho Madtom (Noturus placidus). Southwest Nat. 1992;37: 330. doi:10.2307/3671884

140. Allen GT, Blackford SH, Tabor VM, Cringan MS. Metals, Boron, and Selenium in Neosho Madtom Habitats in the Neosho River in Kansas, U.S.A. Environ Monit Assess. Kluwer Academic Publishers; 2001;66: 1–21. doi:10.1023/A:1026433229820

141. Tiemann JS, Gillette DP, Wildhaber ML, Edds DR. Effects of Lowhead Dams on Riffle-Dwelling Fishes and Macroinvertebrates in a Midwestern River. Trans Am Fish Soc. Taylor & Francis Group ; 2004;133: 705–717. doi:10.1577/T03-058.1

142. Wildhaber ML, Tabor VM, Whitaker JAE, Allert AL, Mulhern DW, Lamberson PJ, et al. Ictalurid Populations in Relation to the Presence of a Main-Stem Reservoir in a Midwestern Warmwater Stream with Emphasis on the Threatened Neosho Madtom. Trans Am Fish Soc. Taylor & Francis Group ; 2000;129: 1264–

1280. doi:10.1577/1548-8659(2000)129<1264:IPIRTT>2.0.CO;2

143. Wenke TL, Eberle ME, U.S. Fish and Wildlife Service. Neosho madtom recovery plan. Denver: U.S. FIsh and Wildlife Service; 1991.

144. Bryan J, Wildhaber M, Noltie D. Examining Neosho Madtom Reproductive Biology Using Ultrasound and Artificial Photothermal Cycles. N Am J Aquac. 2005;67: 221–230. doi:10.1577/A04-020.1

145. Pfingsten DG, Edds DR. Reproductive Traits of the Neosho Madtom, Noturus placidus (Pisces: Ictaluridae). Trans Kansas Acad Sci. 1994;97: 82. doi:10.2307/3627774

146. Wildhaber ML, Allert AL, Schmitt CJ. Potential effects of interspecific competition on Neosho madtom (Noturus placidus) populations. J Freshw Ecol. 1999;14: 19–30.

147. Wildhaber ML, Allert AL, Schmitt CJ, Tabor VM, Mulhern DW, Powell KL. Both contaminants and habitat limit Neosho madtom (Noturus placidus) numbers in the Spring River, a midwestern warmwater stream effected by runoff from historic zinc and lead mining. Fish Response to Toxic Environments. 1998. pp. 9–14.

148. Wildhaber ML, Allert AL, Schmitt CJ, Tabor VM, Mulhern D, Powell KL, et al. Natural and Anthropogenic Influences on the Distribution of the Threatened Neosho Madtom in a Midwestern Warmwater Stream. Trans Am Fish Soc. 2000;129: 243–261. doi:10.1577/1548-8659(2000)129<0243:NAAIOT>2.0.CO;2

149. Bryan JL, Wildhaber ML, Leeds WB, Dey R. Neosho Madtom and Other Ictalurid

Populations in Relation to Hydrologic Characteristics of an Impounded Midwestern Warmwater Stream: Update. US Geol Surv Open-File Rep 2010-1109. 2010;

150. Ellstrand NC, Elam DR. Population Genetic Consequences of Small Population Size: Implications for Plant Conservation. Annu Rev Ecol Syst. Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA ; 1993;24: 217–242. doi:10.1146/annurev.es.24.110193.001245

151. Nei M, Maruyama T, Chakraborty R. The Bottleneck Effect and Genetic Variability in Populations. Evolution (N Y). 1975;29: 1. doi:10.2307/2407137

152. Abascal F, Corvelo A, Cruz F, Villanueva-Cañas JL, Vlasova A, Marcet-Houben M, et al. Extreme genomic erosion after recurrent demographic bottlenecks in the highly endangered Iberian lynx. Genome Biol. 2016;17: 251. doi:10.1186/s13059-016-1090-1

153. Li R, Fan W, Tian G, Zhu H, He L, Cai J, et al. The sequence and de novo assembly of the giant panda genome. Nature. Nature Publishing Group; 2010;463: 311–317. doi:10.1038/nature08696

154. Miller W, Hayes VM, Ratan A, Petersen DC, Wittekindt NE, Miller J, et al. Genetic diversity and population structure of the endangered marsupial Sarcophilus harrisii (Tasmanian devil). Proc Natl Acad Sci U S A. National Academy of Sciences; 2011;108: 12348–53. doi:10.1073/pnas.1102838108

155. Seabury CM, Dowd SE, Seabury PM, Raudsepp T, Brightsmith DJ, Liboriussen P,

et al. A Multi-Platform Draft de novo Genome Assembly and Comparative Analysis for the Scarlet Macaw (Ara macao). Janke A, editor. PLoS One. Public Library of Science; 2013;8: e62415. doi:10.1371/journal.pone.0062415

156. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat Genet. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2012;44: 226–32. doi:10.1038/ng.1028

157. Liu Z, Liu S, Yao J, Bao L, Zhang J, Li Y, et al. The channel catfish genome sequence provides insights into the evolution of scale formation in teleosts. Nat Commun. Nature Publishing Group; 2016;7: 11757. doi:10.1038/ncomms11757

158. Raj A, Stephens M, Pritchard JK. fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. Genetics. 2014;197: 573–89. doi:10.1534/genetics.114.164350

159. Terhorst J, Kamm JA, Song YS. Robust and scalable inference of population history from hundreds of unphased whole genomes. Nat Genet. Nature Research; 2016;49: 303–309. doi:10.1038/ng.3748

160. Bulger AG, Edds DR. Population Structure and Habitat Use in Neosho Madtom (Noturus placidus). Southwest Nat. 2001;46: 8. doi:10.2307/3672368

161. Watterson GA. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 1975;7: 256–276. doi:10.1016/0040-5809(75)90020-9

162. Zimin A V, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. Bioinformatics. 2013;29: 2669–77. doi:10.1093/bioinformatics/btt476

163. Marcais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. Oxford University Press; 2011;27: 764–770. doi:10.1093/bioinformatics/btr011

164. LeGrande WH, Dunham RA, Smitherman RO. Karyology of Three Species of Catfishes (Ictaluridae: Ictalurus) and Four Hybrid Combinations. Copeia. 1984;1984: 873. doi:10.2307/1445331

165. Tiersch TR, Goudie CA. Inheritance and Variation of Genome Size in Half-Sib Families of Hybrid Catfishes. J Hered. Oxford University Press; 1993;84: 122–125. doi:10.1093/oxfordjournals.jhered.a111292

166. Lewis EB. A gene complex controlling segmentation in Drosophila. Nature. Nature Publishing Group; 1978;276: 565–570. doi:10.1038/276565a0

167. Zhao Z, Fu Y-X, Hewett-Emmett D, Boerwinkle E. Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. Gene. 2003;312: 207–13.

168. Rubin C-J, Zody MC, Eriksson J, Meadows JRS, Sherwood E, Webster MT, et al. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature. Nature Publishing Group; 2010;464: 587–591. doi:10.1038/nature08832

169. Ka-Shu Wong G, Liu B, Wang J, Zhang Y, Yang X, Zhang Z, et al. A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. Nature. Nature Publishing Group; 2004;432: 717–722. doi:10.1038/nature03156

170. Bianco E, Nevado B, Ramos-Onsins SE, Pérez-Enciso M, Paudel Y, Crooijmans R. A Deep Catalog of Autosomal Single Nucleotide Variation in the Pig. Yao Y-G, editor. PLoS One. HELM Information Ltd; 2015;10: e0118867. doi:10.1371/journal.pone.0118867

171. Gibbs RA, Taylor JF, Van Tassell CP, Barendse W, Eversole KA, Gill CA, et al. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science. 2009;324: 528–32. doi:10.1126/science.1167936

172. Waters MR, Stafford TW, Kooyman B, Hills L V. Late Pleistocene horse and camel hunting at the southern margin of the ice-free corridor: reassessing the age of Wally's Beach, Canada. Proc Natl Acad Sci U S A. National Academy of Sciences; 2015;112: 4263–7. doi:10.1073/pnas.1420650112

173. Bourgeon L, Burke A, Higham T, Soares A, Zazula G, Letts B. Earliest Human Presence in North America Dated to the Last Glacial Maximum: New Radiocarbon Dates from Bluefish Caves, Canada. Hart JP, editor. PLoS One. Academic Press; 2017;12: e0169486. doi:10.1371/journal.pone.0169486

174. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: function, expression and evolution. Nat Rev Genet. 2009;10: 252–63. doi:10.1038/nrg2538

175. Gehring WJ. Exploring the homeobox. Gene. 1993;135: 215–221. doi:10.1016/0378-1119(93)90068-E

176. Lutz B, Lu HC, Eichele G, Miller D, Kaufman TC. Rescue of Drosophila labial null mutant by the chicken ortholog Hoxb-1 demonstrates that the function of Hox genes is phylogenetically conserved. Genes Dev. Cold Spring Harbor Laboratory Press; 1996;10: 176–84. doi:10.1101/GAD.10.2.176

177. McGinnis W, Krumlauf R, Gehring WJ, Paro R, DeRobertis EM, Stott D, et al. Homeobox genes and axial patterning. Cell. Yale University Press, New Haven; 1992;68: 283–302. doi:10.1016/0092-8674(92)90471-N

178. Lee AP, Koh EGL, Tay A, Brenner S, Venkatesh B. Highly conserved syntenic blocks at the vertebrate Hox loci and conserved regulatory elements within and outside Hox gene clusters. Proc Natl Acad Sci. 2006;103: 6994–6999. doi:10.1073/pnas.0601492103

179. Tallmon DA, Luikart G, Waples RS. The alluring simplicity and complex reality of genetic rescue. Trends Ecol Evol. 2004;19: 489–496. doi:10.1016/j.tree.2004.07.003

180. Johnson WE, Onorato DP, Roelke ME, Land ED, Cunningham M, Belden RC, et al. Genetic Restoration of the Florida Panther. Science (80-). 2010;329.

181. Hostetler JA, Onorato DP, Jansen D, Oli MK. A cat's tale: the impact of genetic restoration on Florida panther population dynamics and persistence. J Anim Ecol. 2013;82: 608–620. doi:10.1111/1365-2656.12033

182. Hansen MM, Bekkevold D, Jansen LF, Mensberg K-LD, Nielsen EE. Genetic restoration of a stocked brown trout Salmo trutta population using microsatellite DNA analysis of historical and contemporary samples. J Appl Ecol. Blackwell Publishing Ltd; 2006;43: 669–679. doi:10.1111/j.1365-2664.2006.01185.x

183. Bouzat JL, Johnson JA, Toepfer JE, Simpson SA, Esker TL, Westemeier RL. Beyond the beneficial effects of translocations as an effective tool for the genetic restoration of isolated populations. Conserv Genet. Springer Netherlands; 2009;10: 191–201. doi:10.1007/s10592-008-9547-8

184. Hedrick P. "Genetic restoration:" a more comprehensive perspective than "genetic rescue." Trends in Ecology & Evolution. 2005. doi:10.1016/j.tree.2005.01.006

VITA

Lynsey Katherine Whitacre was born February 9, 1990 in Saint Joseph, Missouri. During grade school, Lynsey participated in several state and national level science fairs and problem solving contests. Lynsey grew up with two parents heavily involved in the horse and cattle industries. Her competitive nature led her to continued success in horse showing and sparked an interest in animal genetics.

After graduating from Central High School in 2008, Lynsey attended the University of Nebraska-Lincoln to pursue a degree in Biological Engineering. After two semesters, Lynsey transferred from the College of Engineering to the College of Agricultural Sciences and Natural Resources. She would go on to receive a Bachelor's degree in Animal Science with minors in Mathematics and Chemistry. During the last year of her Bachelor's degree Lynsey was accepted into the Undergraduate Creative Activities and Research Experience (UCARE) program. Through this program, Lynsey was paired with Animal Geneticist Dr. Matthew Spangler to conduct research in bovine genetics.

After an initial interest in animal genetics and being submerged into ongoing bovine genetics research, Lynsey decided to apply to graduate school and started a Master's program at the University of Missouri with Dr. Jerry Taylor in 2012. Lynsey completed her Master's degree in 2014 and continued on at the University of Missouri with Dr. Jared Decker to pursue a PhD. During this time, Lynsey published numerous peer-reviewed journal articles and presented at several international research conferences.