UNIVERSITY OF CALIFORNIA
SANTA CRUZ

LONG-RANGE ASSEMBLY AND TRANSCRIPTOMICS ELUCIDATE THE REGULATORY ARCHITECTURE OF THREE VERTEBRATE GENOMES

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOINFORMATICS

by

Edward Stallknecht Rice

June 2018

The Dissertation of Edward Stallknecht Rice
is approved:

Professor Richard E. Green, Chair

Professor Beth A. Shapiro

Professor Angela N. Brooks

Dean Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Long-range assembly and transcriptomics elucidate the regulatory architecture of three vertebrate genomes

by

Edward Stallknecht Rice

Technologies used to sequence and assemble genomes have developed rapidly in the past decade, such that the money and time required to sequence a genome have both fallen by a factor of 10,000. This has given scientists new tools to study a wide variety of questions in biology. In this thesis, I first discuss some of these new technologies and the scientific advances they have facilitated. During my graduate studies, I worked on three different projects that involved producing long-range genome assemblies with these new technologies and/or using these assemblies along with transcriptomic data to answer biological questions about the regulatory architecture of genomes.

The first project I discuss is the assembly of a new reference genome for the domestic horse. A reference assembly of the domestic horse genome was released in 2007 using the best genomic technologies available at the time. Along with collaborators, I used data from new technologies not available in 2007 to assemble a new genome with improved contiguity, completeness, and accuracy. This work provides a resource for horse geneticists studying regulation of gene expression, among other subjects.

The next project I discuss is about temperature-dependent sex determination in the American alligator. Unlike in humans, the sex of an alligator is determined by the temperature

at which its egg is incubated. I used a new long-range genome assembly, RNA sequencing, and differential expression analysis to test a hypothesis about the role of estrogen in regulating gene expression during temperature-dependent sex determination.

The final project I worked on as part of my dissertation research involves learning about how a genetic difference between modern humans and Neanderthals makes human brains unique. While Neanderthals were extremely genetically similar to modern humans, they had a different version of the gene NOVA1, which regulates splicing during brain development. I used RNA-seq transcriptomic data from cell lines with different versions of NOVA1 to determine which genes are spliced differently by these different versions and test whether this is a result of an incompatibility between the Neanderthal version of NOVA1 and the human genetic background.

Acknowledgments

The text of this dissertation includes reprints of previously published material of which I am the first or co-first author. Many co-authors contributed to these publications. Richard E. Green directed and supervised the research which forms the basis for the dissertation and co-authored these publications. Satomi Kohno and Louis J. Guillette Jr. were my primary collaborators on the research that comprises **Chapter 3**. Theodore S. Kalbfleisch was my primary collaborator on the research that comprises **Chapter 2**, and graciously hosted me in Kentucky for a week. Alysson Muotri and Cleber Trujillo were my primary collaborators on the research that comprises **Chapter 4**, and Angela N. Brooks provided important input into this chapter. I thank the following additional co-authors for their contributions to this research: John St. John, Son Pham, Jonathan Howard, Liana F. Lareau, Brendan L. O'Connell, Glenn Hickey, Joel Armstrong, Alden Deran, Ian Fiddes, Roy N. Platt II, Cathy Gresham, Fiona McCarthy, Colin Kern, David Haan, Tan Phan, Carl Schmidt, Jeremy R. Sanford, David A. Ray, Benedict Paten, Michael S. DePriest Jr., Brian P. Walenz, Matthew S. Hestand, Joris R. Vermeesch, Alisa O. Vershinina, Jessica L. Petersen, Carrie J. Finno, Rebecca R. Bellone, Molly E. McCue, Samantha A. Brooks, Ernest Bailey, Ludovic Orlando, Donald C. Miller, Douglas F. Antczak, James N. MacLeod, Ashley Byrne, Maximilian Marin, Jolene Draper.

The ARCS Foundation provided financial assistance for my graduate education.

Thank you to the members of my thesis committee: Richard E. Green, Beth A. Shapiro, and Angela N. Brooks.

Thank you to my thesis advisor Richard E. Green and my co-advisor Beth A. Shapiro.

Thank you to all members of the UC Santa Cruz Paleogenomics Lab, past and present.

Thank you to my graduate school cohort, especially Nathan Schaefer, Ian Fiddes, Arjun Rao, and John Vivian.

Thank you to Alejandro Schaffer for his continued mentorship and interest in my career.

Thank you to Natural Bridges State Beach, the Calvary Choral Scholars program, the Community Music School, the Santa Cruz shape note singing community, and Ray and Alicia of Raymond's Catering.

Thank you to my family for their support: Leslie Ruth Stallknecht, Karl Milton Rice, Karl Rice Stallknecht, Lena Meyer, Uli Meyer, Wolfgang Meyer, Lilly Meyer, Max Meyer, and Nick Meyer.

Chapter 1

Introduction

2018 is an exciting year to finish a Ph.D. in Bioinformatics. Since I started graduate school in 2013, there have been enormous advances in genome assembly and gene editing technologies. In the field of genome assembly, molecular biologists have invented or optimized new molecular techniques that can be used to gain long-range information about genomes at a fraction of the price of traditional methods. Computational biologists have written and implemented algorithms to use these new sources of information to create chromosome-scale assemblies. These advances have allowed scientists to assemble highly contiguous reference genomes for many news species. In the field of genome editing, the CRISPR/Cas9 system has been continuously refined and improved over the past five years, putting fast and precise genome editing within the reach of smaller laboratories and allowing biologists to perform experiments that were not possible before.

I mention long-range genome assembly and CRISPR/Cas9 specifically because the three projects that comprise this dissertation would not have been possible without recent ad-

vances in these technologies. The first project in this dissertation is the assembly of a new reference genome for the domestic horse *Equus caballus*. The previous reference genome for this species was released in 2007 and constructed using the best genomic technologies available at the time, including Sanger sequencing, bacterial artificial chromosome end pair sequencing, radiation hybrid mapping, and fluorescence in situ hybridization mapping (Wade et al., 2009). This assembly has been used to study the genetics, health, and evolution of horses (Coleman et al., 2010; Bellone et al., 2013; Gaunitz et al., 2018). However, due to limitations in the technologies available in 2007, it contains many gaps where the sequence is not known as well as sequences that have not been placed on chromosomes.

My collaborators and I used several new technologies that were not available in 2007, including short-read sequencing, proximity ligation, long-read sequencing, and linked-read sequencing with molecular barcodes, to assemble a new reference genome for the species. This updated genome fills in many of the gaps present in the previous version, increasing the ungapped contiguity of the genome by a factor of 40. It also assigns more sequence to chromosomes, resulting in a 3% increase in assigned sequence. Furthermore, more sequencing reads from the Functional Annotation of Animal Genomes (FAANG) project map to this new genome than the previous reference, especially in GC-rich regulatory regions, which will better allow horse geneticists to study the regulation of gene expression. Many of the technologies we used to make these improvements either did not yet exist when I began my Ph.D. or had not yet been refined and optimized enough to reasonably use in a project like this. **Chapter 2** contains a preprint of a manuscript about this project (Kalbfleisch et al., 2018) that is currently in review at *Communications Biology*.

The next project in this thesis involves estrogen regulation of gene expression during temperature-dependent sex determination in the American alligator. The American alligator *Alligator mississippiensis*, like all other crocodilians and many other reptiles, does not have sex chromosomes in its genome to determine the sex of an individual like mammals do. Instead, male and female alligators are genetically identical, and the temperature at which an egg is incubated after it is laid determines whether an embryo develops as a male or a female (Ferguson and Joanen, 1983). This sex determination system, called "temperature-dependent sex determination," is not well-understood.

However, in the presence of excess exogenous estrogen, embryos develop into females regardless of incubation temperature (Bull et al., 1988). This has led many scientists to hypothesize that estrogenic regulation of gene expression plays an important role in temperature-dependent sex determination (Lance, 2009). However, estrogenic regulation of gene expression involves long-range interactions between different parts of the genome (Chan and Song, 2008; Fullwood et al., 2009; Zhang et al., 2010), and at the time I began the project, the available reference genome for the American alligator was not contiguous enough to use to study such long-range interactions (St. John et al., 2012; Green et al., 2014). My collaborators and I used Chicago (Putnam et al., 2016), a new technique for preparing long insert-size sequencing libraries, to assemble an improved alligator reference genome, and then used this genome to provide the first evidence that estrogen regulates genes involved in temperature-dependent sex determination. **Chapter 3** contains a paper I wrote with these collaborators and published in *Genome Research* (Rice et al., 2017).

The third and final project in this dissertation is about brain development in Nean-

derthals. Neanderthals were a group of archaic humans who went extinct 40,000 years ago (Higham et al., 2014). Neanderthals shared genetic variation due to incomplete lineage sorting and admixture, leaving most humans today with some amount of Neanderthal ancestry (Green et al., 2010). In most places in the genome where the average human differs from Neanderthals, there are some humans alive today with the Neanderthal version. However, in a small number of these places, there are no humans alive today with the Neanderthal version, suggesting that the Neanderthal allele may have been too deleterious to humans to persist into the present day. One of these places where an allele specific to modern humans reached fixation is a coding change in the gene NOVA1, a splicing factor responsible for regulating other genes during brain development. This raises the possibility that this difference is part of what made modern humans unique from archaic humans and other primates.

While sequencing ancient DNA can allow scientists to answer questions about the past, such as how prehistoric humans evolved and spread out across the world, determining how individual genetic differences between Neanderthals and modern humans made them different is not possible with ancient DNA alone, but requires performing a controlled experiment with a single independent variable. CRISPR/Cas9 makes such experiments possible. In **Chapter 4**, I describe how my collaborators and I used CRISPR/Cas9 to grow neural organoids with the Neanderthal version of NOVA1 along with controls with the human version of NOVA1, and then sequenced RNA from these organoids and analyzed it to determine how genes are spliced differently based on what version of NOVA1 is regulating them. The ability to edit genomes to perform an experiment to learn how a genetic difference between Neanderthals and modern humans caused differences in brain development represents the beginning of a new era

in evolutionary biology, in which we can learn about the past not just through observing old things that have survived into the present, but by recreating parts of the past and performing controlled experiments on them.

I worked on these three projects with many collaborators, without whom none of the work presented in this dissertation would have been possible. In prefaces to each of the proceeding chapters, I discuss what parts of the work described in that chapter I performed myself and what work was performed by collaborators. My adviser Richard E. Green directed and supervised the research which forms the basis for the dissertation and co-authored these publications. **Chapter 2** and **Chapter 3** are comprised primarily of published or submitted papers written with co-authors, whom I thank for giving me permission to reproduce our manuscripts in this dissertation.

Chapter 2

A new reference genome for the domestic horse using first-, second-, and third-generation sequencing

This chapter consists of a preprint of a paper in review at *Nature Communications* about a new assembly of the domestic horse genome. I am co-first author on this manuscript along with Theodore S. Kalbfleisch. My specific contributions to this project include testing various assembly approaches to determine the best one for this project, preparing the Hi-C library using a protocol written by Brendan C. O'Connell, running the HiRise step of the assembly, performing quality control and assessment on the assembly, assigning scaffolds to chromosomes based on a physical map, and submitting the genome to NCBI. For the manuscript, I co-wrote the Introduction with T.S.K., wrote the subsections of Results titled "Agreement with existing RH map" and "Protein set completeness and comparative annotation," wrote the Methods

subsections pertaining to the steps in the assembly I performed, and created Figure 3. I also edited other authors' contributions. T.S.K. ran the Masurca and PBJelly steps of the assembly, aligned reads to the reference genome for quality control and assessment purposes, and wrote the sections of the manuscript pertaining to these steps. James N. MacLeod designed and led the project with critical input from other authors, including my adviser Richard E. Green, who also extensively advised me in this project and edited the manuscript. Other coauthors assisted with data generation, quality control and assessment, and writing and/or editing. I thank my coauthors for their work, without which this project would not have been possible, and for allowing me to reproduce our manuscript here.

EquCab3, an Updated Reference Genome for the Domestic Horse

Theodore S. Kalbfleisch*†[1], Edward S. Rice[2]†, Michael S. DePriest Jr.[1], Brian P. Walenz[3], Matthew S. Hestand[4], Joris R. Vermeesch[4], Brendan L. O'Connell[2]‡, Ian T. Fiddes[2]□, Alisa O. Vershinina[5], Jessica L. Petersen[6], Carrie J. Finno[7], Rebecca R. Bellone[8], Molly E. McCue[9], Samantha A. Brooks[10], Ernest Bailey[11], Ludovic Orlando[12,13], Richard E. Green[2], Donald C. Miller[14], Douglas F. Antczak[14], James N. MacLeod[11]

[1] Department of Biochemistry and Molecular Genetics, School of Medicine, University of Louisville, Louisville, KY 40292
[2] Department of Biomolecular Engineering, UC Santa Cruz, CA 95064
[3] Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892
[4] Center for Human Genetics, Katholieke University Leuven (KU Leuven), 3000 Leuven, Belgium
[5] Department of Ecology and Evolutionary Biology, UC Santa Cruz, CA 95064
[6] Department of Animal Science, University of Nebraska - Lincoln, Lincoln, NE, 68583-0908
[7] Department of Population Health and Reproduction, University of California, Davis, CA 95616
[8] Veterinary Genetics Laboratory, University of California, Davis, CA 95616
[9] Department of Veterinary Population Medicine, University of Minnesota, St. Paul, MN 55108
[10] Department of Animal Sciences, University of Florida, Gainesville, FL 32611
[11] Gluck Equine Research Center, Department of Veterinary Science, University of Kentucky, Lexington, KY 40546
[12] Centre for GeoGenetics, Natural History Museum of Denmark, 1350K Copenhagen, Denmark.
[13] Laboratoire d'Anthropobiologie Moléculaire et d'Imagerie de Synthèse UMR 5288, Université de Toulouse, CNRS, Université Paul Sabatier, France.
[14] Baker Institute for Animal Health, College of Veterinary Medicine, Cornell University, Ithaca, New York, NY 14853

* to whom correspondence should be addressed: ted.kalbfleisch@louisville.edu
† These authors contributed equally to this work.
‡ Present address: Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR 97239
□ Present address: 10x Genomics, Inc. Pleasanton, CA 94566

Abstract

EquCab2, a high-quality reference genome for the domestic horse, was released in 2007. Since then, it has served as the foundation for nearly all genomic work done in equids. Recent advances in genomic sequencing technology and computational assembly methods have allowed scientists to improve reference assemblies of large animal and plant genomes in terms of contiguity and composition. In 2014, the equine genomics research community began a project to improve the reference sequence for the horse, building upon the solid foundation of EquCab2 and incorporating new short-read data, long-read data, and proximity ligation data. The result, EquCab3, is presented here. The count of non-N bases in the incorporated chromosomes is improved from 2.33Gb in EquCab2 to 2.41Gb from EquCab3. Contiguity has also been improved nearly 40-fold with a contig N50 of 4.5Mb and scaffold contiguity enhanced to where all but one of the 32 chromosomes is comprised of a single scaffold.

The domestic horse *Equus caballus* is a culturally, economically, and historically important domesticated animal. Since horses were domesticated ~5kya in central Asia[1], humans have used them extensively for agriculture, transportation, military conflict, and sport. Horses have been selectively bred for speed, strength, endurance, size, appearance traits, and temperament.

EquCab2, a high-quality reference genome assembly of the domestic horse, was released in 2007[2]. This assembly was generated using the best genomic sequencing and assembly technologies available at the time, namely: Sanger sequencing, bacterial artificial chromosome (BAC) end pairs, radiation hybrid (RH) mapping, and fluorescence in situ hybridization (FISH) mapping. Since then, many researchers have used this reference genome to study the genetics of various traits in horses[3–9], as well as their health[10–13] and evolution[14–17]. However, EquCab2 contains numerous gaps in scaffolds as well as sequences unassigned to chromosomes, and genomic DNA resequencing[18] and gene annotation[19] studies have found inconsistencies in this genome. Therefore, new genomic technologies present an opportunity to improve the equine reference genome.

We present here a new reference assembly for the domestic horse, EquCab3. This assembly benefited from rapidly evolving high-throughput sequencing technologies and new algorithms used to assemble data from these platforms. Specifically, this project began from the solid foundation of 6.8-fold coverage Sanger sequence data[2], as well as a radiation-hybrid map and FISH data[20]. These data were augmented with 45-fold coverage Illumina short-read data that improved the characterization and accuracy of unique regions of the genome, increasing the contig N50 values 10-fold. Two different proximity ligation library preparation protocols made it

possible to order these contigs and generate chromosome length scaffolds. In EquCab3, only chr6 is comprised of more than one scaffold. Finally ~16X PacBio long reads made it possible to close many of the gaps between the ordered contigs, thereby improving the contig N50 values 4-fold again. The resulting assembly is enhanced not only in contiguity but also in composition. This new version of reference sequence for the domestic horse reduces the number of gaps 10-fold and increases the number of assembled bases by 3% in the incorporated chromosomes over EquCab2.

Results

A new reference assembly of the domestic horse genome

We generated a new reference assembly of the domestic horse using first-, second-, and third-generation sequencing data as well as physical chromosome maps. This new reference is derived from the same female Thoroughbred horse, Twilight, that produced EquCab2. We did not attempt to derive a new mitochondrial genome sequence, and instead relied on the work done by Xu and Arnason[21].

We used both previously published data and newly generated data to generate this reference assembly. The previously published data sets are comprised of the data used to construct EquCab2: Sanger sequencing data, BAC-end pairs[2], and a physical map containing radiation hybrid and FISH markers[20]. For this assembly, we generated shotgun Illumina short reads, Chicago and Hi-C proximity ligation libraries, PacBio long reads, and 10x Chromium linked reads. As there is no existing software or method for creating an assembly from this

combination of data types, we developed a custom pipeline to leverage the strengths of each of these data sets.

First, we used the high coverage (45x) and accuracy of Illumina short reads to generate "super-reads" with MaSuRCA[22]. We assembled these super-reads together with the long and accurate but lower coverage (6.8x) Sanger reads to create an initial assembly with Celera Assembler[23]. We scaffolded this initial assembly with the long insert-size Chicago and Hi-C proximity ligation libraries using the HiRise scaffolder[24]. To identify and correct misassemblies, we mapped all physical markers and sequence data, including BAC-end sequences, to the resulting scaffolds. We filled gaps in the corrected scaffolds with PacBio reads, which are longer but lower-accuracy than Sanger reads, using PBJelly. We phased the genome using 10x Chromium linked reads and the longranger pipeline. We aligned the high-identity and coverage Illumina short reads to the genome and used these alignments to correct errors. Finally, we used the physical map to assign scaffolds to chromosomes. The resulting assembly, EquCab3, is an improvement over EquCab2 in terms of contiguity, completeness, read mappability, and agreement with the physical map.

Improved Contiguity

EquCab3 has improved N50 values for both contigs and scaffolds over those reported for EquCab2. For the contigs, an N50 value of 4.5 Mb vs 112 kb, and for scaffolds, 86 Mb vs 46 Mb (Table 1). At each phase of the assembly process (described in Methods), there is an improvement in either the contig or scaffold N50 over the values achieved in EquCab2. The one exception is the scaffold N50 of the Sanger + MaSuRCa Super Reads. Our scaffold N50 is 6.6Mb, less than the final value of 46Mb reported in Wade et al. (2009). The EquCab2 value

incorporated additional long range data such as BAC-end reads from a library derived from Twilight's half-brother Bravo, as well as radiation hybrid map data. With all PacBio and proximity ligation data from Twilight included, the contig N50 is increased 40-fold, and the scaffold N50 is increased from a chromosome arm-limited 46Mb to a chromosome length-limited 86Mb. Further, the total count of gaps in the ordered chromosomes is decreased more than 90%, from 42,304 in EquCab2 to 3,771 in EquCab3.

Table 1

EquCab3 Sequence Composition	Contig N50	Scaffold N50
Sanger + MaSuRCa Super Reads	1.2Mb	6.6Mb
Sanger + MaSuRCa Super Reads + Chicago + HiC	1.2Mb	86Mb
Sanger + MaSuRCa Super Reads + Chicago + HiC + PacBio	4.5Mb	86Mb

EquCab2	Contig N50	Scaffold N50
Sanger Fosmid + BAC + RH Map data	112kb	46Mb

Read mapping

The equine genome community is participating in the Functional Annotation of Animal Genomes (FAANG) project. The initial phase of this project has produced RNA-seq and whole genome shotgun sequence data from two Thoroughbred mares that are not the subject of the reference assembly. Data from both horses have been mapped to both EquCab2 and EquCab3[25]. In the first phase of the equine FAANG effort, the RNA-seq data are comprised of samples from eight tissues. As shown in Figure 1, for RNA-seq, unique mappings of the reads are increased by an average of 2.15% over EquCab2, and WGS paired reads improved by 0.44%. In the WGS

dataset, more reads mapped to EquCab3 than EquCab2 (Figure 2) and the count of reads mapping in a proper pair, i.e., with both ends mapping with correct orientation, increased from a value of 811,622,501 to 814,804,213, an increase of 0.38% of the total read count.

This increase in read mapping is a function of several ways in which EquCab3 is an improvement over EquCab2. EquCab3 is more accurate due to the high-coverage high-identity Illumina data used both in the initial assembly and polishing steps, and contains fewer gaps than EquCab2 due to the long read gap-filling step, resulting in fewer dips in alignment coverage, shown in Figure 3a. In addition, EquCab3 has more sequence assigned to chromosomes, giving reads more total sequence to map to, also demonstrated by Figure 3a from the length increase in chr31 from EquCab2 to EquCab3. Finally, EquCab3 improves the characterization of GC-rich regions.

The GC content of EquCab3 is roughly equivalent to that of EquCab2 (both near 41.6%). However, the GC fraction of the WGS reads for the two FAANG horses that mapped to EquCab3 but failed to map to EquCab2 is 48.9%. The GC content for the entire WGS dataset is 41.8%. This demonstrates an improvement in the characterization of GC-rich regions of the equine genome, and is largely attributable to the PCR-free library preparations now in common use.

We also assessed the quality of EquCab3 by aligning ancient DNA (aDNA) reads to it. EquCab2 has been used in many studies as a reference for DNA recovered from paleontological samples, giving insight into the evolution and domestication of horses[14–18]. We compared mapping statistics between EquCab3 and EquCab2 for 13 previously sequenced ancient horses[17] (Supplementary Table S2). A paired Wilcoxon test showed a significant improvement in

mapping (p=0.0017), with all 13 samples having more reads mapped to EquCab3 than to EquCab2.

Agreement with existing RH map

We used a radiation hybrid map of the horse genome to assign scaffolds to chromosomes[20]. EquCab3 agrees with the radiation hybrid map more often than EquCab2. Of the 4,103 markers on the physical map, 2,982 map to EquCab2 while 3,039 map to EquCab3. In addition, EquCab2 contains 391 marker pairs that are oriented differently on the assembly than on the map, whereas EquCab3 contains 395, despite the 57 additional markers mapping to EquCab3. This improvement can be attributed to the lower rate of misassemblies from the use of proximity ligation data for scaffolding. An example of a misassembly in EquCab2 corrected in EquCab3 is shown in Figure 3b-e.

Of the 395 misoriented marker pairs on EquCab2, 352 are oriented the same way on both EquCab2 and EquCab3, but differently on the map. Given the multiple, orthogonal data types and differing assembly strategies used to construct EquCab2 and EquCab3, we suggest that some or all of these 352 marker pairs are oriented correctly in both assemblies but incorrectly on the RH map. Of the remaining 43 marker pairs that are misoriented on EquCab3 but not on EquCab2, 36 of these pairs do not have both markers mapping to EquCab2, leaving only 7 marker pairs agreeing with EquCab2 but not EquCab3. Given that the RH map was used to guide the assembly of EquCab2, we find this level of disagreement acceptable.

Protein set completeness and comparative annotation

We used two methods to evaluate the completeness of our genome: universal ortholog analysis and comparative annotation. For universal ortholog analysis, we used BUSCO[26] and the mammalian universal ortholog set. Out of 4,104 mammalian universal orthologs, BUSCO found 4,092 (99.7%) as complete orthologs in EquCab3 with 5 fragmented and 7 missing, compared to 4,064 (99.0%) complete orthologs in EquCab2 with 27 fragmented and 13 missing. EquCab3's higher BUSCO score indicates that it is more complete than EquCab2.

Comparative Annotation Toolkit (CAT) is a software pipeline that leverages whole genome alignments, existing annotations, and comparative gene prediction tools to simultaneously annotate multiple genomes, defining orthologous relationships and discovering gene family expansion and contraction[27]. CAT also diagnoses assembly quality by investigating the rate of gene model-breaking indels seen in transcript projections from a reference as well as looking at the rate of transcript projections that map in a disjoint fashion. We performed comparative annotation of EquCab2 and EquCab3 using the genomes of pig, cow, white rhinoceros, elephant, and human. Comparative annotation of EquCab3 and EquCab2 found that more orthologs of genes in the other genomes were found in EquCab3 than in EquCab2 (Figure 4a), fewer predicted genes were split between contigs in EquCab3 than in EquCab2 (Figure 4b), and the distribution of gene coverage is significantly better in EquCab3 than in EquCab2 (Figure 4c). These results indicate that EquCab3 is a more complete and contiguous assembly than EquCab2.

Phasing

Most published assemblies of diploid organisms are pseudo-haploidizations produced by arbitrarily choosing between the two alleles at each heterozygous site in the genome. The 10X Chromium platform is useful for haplotype phasing, as each set of linked reads it produces comes from the same haplotype. We took advantage of this by using 10X reads and the longranger pipeline to phase Twilight's variants in EquCab3. For each phase block inferred by longranger, rather than arbitrarily choosing which haplotype to include in the final assembly, we chose the allele which is most common among 4 Thoroughbreds, the 2 FAANG horses, and data from two other Thoroughbreds from an earlier study by Sarkar et al.[12] This makes the reference pseudo-haploidization more similar to the population average and thus more likely to contain the ancestral allele at each heterozygous site in Twilight's genome. For analyses which would be adversely affected by this ancestral reference bias, we provide the phased 10X variant calls as supplemental data.

Discussion

This new genome represents an improvement for the horse reference in terms of both composition and contiguity. It is also more consistent with the existing RH map and FISH data for the horse than was EquCab2. Going forward, the lens through which this reference will be viewed will be as an alignment target for the vast amount of high throughput sequence data that will continue to be generated for the horse and other related species. The assembly process described here was guided and informed by data that included not only high quality short reads but long reads and proximity ligation data. All equine data produced by any of these technologies should be well served going forward. The most common data types for genetic and genomic

studies, Illumina short reads, have been demonstrated to map to the new reference for two non-related Thoroughbreds at an improved average rate of 2.15% for RNA-Seq and 0.44% for WGS libraries. In a comparative genomics analysis, more gene orthologs were found, and for those that were found, the coverage of the homologous transcript sequence was more complete. The new long-range sequence data not only improved the contiguity of the genome, but allowed us to phase the genomic data for Twilight. Finally, the regions added for the genome were higher in GC content, which will enable a better characterization of both genetic variation and epigenetic status in GC-rich regulatory regions for the horse.

This represents a culmination of a project conceived and begun in 2014 with the support of the equine genomics community. Although it will certainly not be the last reference genome for the domestic horse produced for public annotation, it should foster genetic and genomic discoveries for years to come.

Methods

Sequence data Generation

Sanger Data: The Sanger sequence data for the Thoroughbred mare Twilight, produced for and used to build EquCab2[2], were downloaded from the NCBI Trace Archive as described in Rebolledo-Mendez et al.[18]

Illumina PE HiSeq and MiSeq: Construction of a PCR-free shotgun genomic library and sequencing on MiSeq and HiSeq2500 instruments were carried out at the Roy J. Carver Biotechnology Center, University of Illinois at Urbana-Champaign (UIUC).

A shotgun genomic DNA library with an insert size of 500bp (range 300 - 650bp) was constructed from 2µg of Twilight's genomic DNA after sonication with a Covaris M220 (Covaris, MA) with the Hyper Library Preparation Kit from Kapa Biosystems (Roche) with no PCR amplification. The adaptored DNA library was loaded onto a 2% agarose gel and fragments 450bp to 550bp in length were cut from the gel and recovered with the QIAquick gel extraction kit (Qiagen, CA). The size-selected library was quantitated with Qubit (ThermoFisher) and run on an Agilent bioanalyzer DNA high-sensitivity chip (Agilent, Santa Clara, CA) to confirm the presence of DNA fragments of the expected size range. It was further quantitated by qPCR on a BioRad CFX Connect Real-Time System (Bio-Rad Laboratories, Inc. CA) prior to sequencing for maximization of the number of clusters in the sequencing flowcell.

The PCR-free shotgun library was first sequenced on a MiSeq with v3 reagents to generate paired-reads 300nt in length. The data confirmed the DNA fragment sizes. The library was subsequently sequenced on a HiSeq2500 for 161 cycles from each end using a TruSeq Rapid SBS sequencing kit1 v1. The fastq read files were generated with the bcl2fastq v1.8.4 Conversion Software (Illumina, San Diego, CA).

PacBio: Ten micrograms of high molecular genomic DNA from Twilight was sheared with gTUBES (Covaris) in an Eppendorf® 5424 centrifuge at 4800 RPM for 2x 60 seconds. A single PacBio library was prepared from this following PacBio's protocol P/N 100-286-000-07 (20 kb Template Preparation Using BluePippin(Tm) Size-Selection System) with PacBio DNA Template Prep Kit 1.0. For the size selection, the sample was run on a 0.75% BluePippin cassette (ref: PAC20KB) using the pre-defined '0.75% DF Marker S1 high-pass 6-10kb vs3' program and a cut-off of 10-50kb. The library was sequenced on 88 SMRT cells on a PacBio RSII using

DNA/Polymerase Binding Kit P6 and DNA Sequencing Kit 4.0 (v2) sequencing reagents, magbead loading, and stage start. All SMRTcells were run through PacBio's SMRT Portal v2.3.0 pipeline RS_subreads.1 with default settings except for minimum subread and polymerase read lengths of 1kb. In addition, reads-of-insert were generated using the RS_ReadsOfInsert.1 pipeline with a minimum insert read length set to 1kb. Reads-of-insert had a mean of 4 passes and length of 11,785 bp.

Of the total initial read count, 5,934,426, we were able to create circular consensus (.ccs) reads totalling 371,943 reads. The remainder of the reads were used to generate a reads of insert file consisting of 5,562,483. These two datasets were used in the PBJelly runs described below.

CHiCago library: We generated a CHiCago library as previously described[24] using blood from Twilight.

Hi-C library: We generated a Hi-C library with primary fibroblasts from Twilight using a Hi-C protocol modified such that the chromatin immobilization took place on magnetic beads. We crosslinked the fibroblasts in formaldehyde, and lysed, washed, and resuspended as described by Lieberman-Aiden et al.[28] We then immobilized the chromatin on SPRI beads as described by Shendure et al.[29] We restriction digested the DNA with DpnII, labeled ends with biotinylated dCTP, ligated ends, and reversed crosslinks. The sample was prepared for sequencing using the NEB Ultra library preparation kit according to the manufacturer's instructions, with one exception: prior to indexing PCR, the sample was enriched by pulldown on 30 μL Invitrogen C1 Streptavidin beads, then washed to remove non-biotinylated DNA fragments.

10X Genomics library: Twilight's genomic DNA was size selected for fragments >40 Kbp on a BluePippin instrument (Sage Sciences, Beverly MA) and Illumina sequencing libraries were constructed using the 10X Genomics Chromium Controller instrument with their Genome Reagents Kit v2 chemistry (10x Genomics, Pleasanton CA) according to the manufacturer's recommendations. The resulting Illumina library was sequenced on a NextSeq500 using a High Output Kit v2 for a paired-end, 2x151 bp run (Illumina, San Diego CA). The data were analyzed and assembled using the 10x Genomics Supernova version 1.1.5 pipelines.

Assembly Generation

MaSuRCA: The 42X Illumina PE data described above was assembled into super-reads using MaSuRCA[22] version 3.1.3. The super reads produced a reduced representation of fragments with ~2X coverage (4.7Gb) with a contig N50 of 1,734 nucleotides.

Celera Assembler: The Celera Assembler[30,31], version 8.2 (downloaded from http://wgs-assembler.sourceforge.net/wiki/index.php?title=Main_Page) was used to create contigs and scaffolds using the super reads produced by MaSuRCA and the EquCab2 Sanger sequence data.

HiRise: We scaffolded the output of Celera Assembler using HiRise version 2.1.1 in serial mode with default parameters, with the CHiCago and Hi-C libraries as input libraries[24].

Identifying misassemblies: In order to identify misassemblies in the HiRise assembly relative to EquCab2, we aligned the HiRise output scaffolds to EquCab2 using nucmer with default parameters[32]. In every place where the alignment indicated a difference in order and orientation of scaffolds between the two assemblies, we used every available data type to resolve

the discrepancy and determine which was correct. Our strategies included aligning BAC-end pairs from a half-brother of Twilight[2] to the assemblies using bwa mem with default parameters[33], assessing concordance with the physical map, looking for split genes predicted by the Comparative Annotation Toolkit[27], aligning coding sequences of any genes in the region to the assemblies using gmap with default parameters[34], and examining heatmaps of long-range read pairs mapping to the assembly generated by the HiRise and longranger pipelines.

PBJelly: We filled gaps in the manually corrected HiRise scaffolds from the previous step using PacBio reads of insert and circular consensus reads as input to the PBJelly (version PBSuite_15.8.24) pipeline with the steps setup, mapping, support, extraction, assembly, and output, in that order.

Assigning scaffolds to chromosomes: We used a previously published radiation hybrid map[20] to assign scaffolds to chromosomes. We aligned each physical marker's STS primers to the assembly using bwa fastmap[35] and used only markers with both primers aligning uniquely and in the correct orientation. We then placed scaffolds on chromosomes based on the markers' mapping locations.

Quality control and assessment

Read Mapping: Short read sequence data generated in the initial phase of the equine Functional Annotation of Animal Genomes (FAANG) project has been mapped to both EquCab2 and EquCab3 for comparison of mapping fractions. Both Whole Genome Shotgun (WGS) sequence (40X), and RNA-Seq (avg 20M reads/tissue) datasets from 8 tissues types for each of two animals were trimmed using TrimGalore (a wrapper for Cutadapt[36]). For WGS data, the

program BWA[35] (version 0.6.1) aln module was used to align the reads to the reference. BWA sampe was used to produce a usable SAM file. SAMtools[37] (version 0.1.18) was used to convert from SAM to BAM format. Picard (version 1.65) FixMateInformation and MarkDuplicates modules were used, followed by GATK[38] (version 1.5) RealignerTargetCreator, and IndelRealigner (validation_strictness set to LENIENT for each). For the RNA-seq data, the mapping program STAR[39] (version 2.5.3a) was used with default parameters except for the following: --readFilesCommand zcat --outSAMtype BAM SortedByCoordinate --outBAMsortingThreadN 16 -outSAMunmapped Within.

Polishing: Since Twilight's sequence data and the EquCab3 were derived from the same animal, any homozygous differences between the PE data and the reference of which they are a component are likely errors. The differing bases were likely contributions from the sequence data generated on other platforms used for the assembly such as the Sanger or PacBio data.

The errors are either with the reference or with the miscalled/undersampled genotypes derived by the variant discovery software. To evaluate these positions, we performed variant discovery and genotyping with the UnifiedGenotyper using the Twilight PE data, the two FAANG thoroughbreds, and two additional thoroughbreds from Sarkar et al.[12] whose data was downloaded from the Sequence Read Archive (BioSample/experiment accession numbers SAMN03838869/SRX1097022, SAMN03838867/SRX1097495) and mapped as described above. The UnifiedGenotyper was used in discovery mode on the cohort. The resulting variant call format file was then parsed with custom java software looking for positions at which the Twilight data produced a homozygous genotype differing from the reference. The genotypes for

the other animals were then queried at those positions. If the reference allele was detected in one of the other horses, the reference nucleotide at that position was not changed.

Removal of Microbial Contamination: To build microbial sequence databases, all bacterial, viral, and fungal reference genomes were downloaded from RefSeq. For each of the three databases (bacteria, viruses, and fungi), the sequences were first masked with DustMasker[40]. Kraken v1.0[41] was used to generate k-mers (k=32) and to search the EquCab3 contigs for exact matches. Contigs with at least one exact 32-mer match were considered microbial contaminants and removed from the reference sequence. A total of 41 contigs were removed in this way.

Removal of Small Contigs: All scaffolds smaller than 3000 bases in length were removed from the assembly that was submitted for annotation. The contig and scaffold N50s for what was submitted were 4.73Mb and 87.2Mb, respectively.

Phasing with 10X data: The data generated for Twilight on the 10X platform described above was mapped to the reference using the longranger (version 2.1.3) wgs module. The phased variant file produced was then used to modify individual variant positions to conform to the haplotype whose allele was most common among the FAANG horses, and two other thoroughbreds described above in *Polishing*.

N50 calculation: The PBJelly (version PBSuite_15.8.24) utility summarizeAssembly.py was used to calculate N50 values. The default setting of 25 was used for the minimum gap setting. This ignored any gaps sized less than 25 Ns.

Universal ortholog analysis: For universal ortholog analysis, we used BUSCO[26] version 3.0.2 in protein mode with the lineage dataset mammalia_odb9 version 2016-02-13. For protein

set inputs, we used the official NCBI protein sets for EquCab2.0 (accession GCF_000002305.2) and EquCab3.0 (accession GCF_002863925.1).

Comparative annotation: For this analysis, a progressiveCactus[42] alignment of equCab2 and equCab3 was performed with pig (susScr3), cattle (bosTau8), white rhinoceros (cerSim1), elephant (loxAfr3) and human (hg38). The guide tree was (((Human:0.164501,((Pig:0.12,Cow:0.16908)1:0.02,(equCab3:0.0001,equCab2:0.0001):0.059397,White_rhinoceros:0.05)1:0.060727)1:0.032898)1:0.023664,Elephant:0.155646);, putting EquCab2 and EquCab3 under the same node with a branch length of 0.0001. CAT[27] was then run using the Ensembl V89 annotation of pig as the source transcript set. No RNA-seq data were provided, and so no transcript cleanup steps or comparative gene predictions were performed. Split gene analysis is performed by looking at transcripts which have multiple projections after paralog resolution and which have multiple projections whose start and stop points are within 10bp of each other in source transcript coordinates.

Read Filtering and Counting: Custom java software using the htsjdk (version 2.12.01) was written to filter the mappings that were not primary (getNotPrimaryAlignmentFlag() is false) from the mapped read (getReadUnmappedFlag() is false) count.

Ancient DNA mapping: We downloaded single-end Illumina reads produced by a previous study[17] (Supplementary Table S2, NCBI Bioproject PRJEB19970). Adapters and PCR artifacts were trimmed using AdapterRemoval v2[43]. For normalization across samples, fastq files were downsampled to 6M reads using seqtk (https://github.com/lh3/seqtk). Low complexity sequences were removed using PRINSEQ[44] following bwa mapping[35] with parameters optimized for aDNA: *aln* algorithm, “seed disable” flag, and minimum mapping phred quality of 20.

Data Availability

The sequence read datasets generated during the current study are available in the NCBI SRA repository under accession SRP126689. The final assembly generated during the current study is available in the NCBI Genbank repository under accession GCA_002863925.1. We also provide intermediate assemblies produced during the process, a de novo assembly based solely on the PacBio data, and phased variant calls from the 10X longranger pipeline as supplementary data.

Additional Information

Acknowledgements

This work was supported by Morris Animal Foundation grant D15EQ-019 and the NRPS8 Horse Genome Coordinator Fund. E.S.R. is an ARCS scholar. Support for C.J.F. was provided by the National Institutes of Health (NIH) (1K01OD015134 and L40 TR001136). Alignment and assembly work for this project were performed on the University of Louisville Cardinal Research Cluster. The authors are grateful to Mr. Harrison Simrall for his assistance in installing and running many of the applications used in this work. We also wish to thank Wim Meert for PacBio runs and library preparations, and Peter A. Schweitzer who prepared the libraries for the 10X Genomics data generated for the project. Finally, we would like to thank Drs. Tomas Bergström, Sofia Mikko, Agnes Viluma, Göran Andersson, and Petr Horin for their insightful review of the MHC locus for this assembly.

Author contributions

T.S.K., E.S.R., and B.P.W. performed the assembly. J.N.M., M.S.H., J.R.V., B.L.O., and E.S.R. contributed genomic DNA sequencing and primary data generation. T.S.K., E.S.R., M.S.D., I.T.F., and A.O.V. performed quality control on the assembly and evaluated its completeness and accuracy. J.L.P, C.J.F, and R.R.B. prepared and sequenced DNA and RNA for annotation and quality control purposes. J.N.M. designed and led the project with critical input from T.S.K., D.F.A., D.C.M., R.E.G., E.B., L.O., S.A.B., and M.E.M. T.S.K., E.S.R., M.S.D., and J.N.M. wrote the manuscript. All authors reviewed and edited the manuscript.

Competing financial interests

I.T.F. is an employee of 10x Genomics, Inc. R.E.G. is a co-founder and scientific adviser of Dovetail Genomics, LLC. No other authors have competing financial interests to disclose.

References

1. Outram, A. K. *et al.* The earliest horse harnessing and milking. *Science* **323,** 1332–1335 (2009).
2. Wade, C. M. *et al.* Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science* **326,** 865–867 (2009).
3. Coleman, S. J. *et al.* Structural annotation of equine protein-coding genes determined by mRNA sequencing. *Anim. Genet.* **41 Suppl 2,** 121–130 (2010).
4. Vanderman, K. S. *et al.* Brother of CDO (BOC) expression in equine articular cartilage. *Osteoarthritis Cartilage* **19,** 435–438 (2011).
5. Schaefer, R. J. *et al.* Developing a 670k genotyping array to tag ~2M SNPs across 24 horse breeds. *BMC Genomics* **18,** 565 (2017).
6. Petersen, J. L. *et al.* Genome-wide analysis reveals selection for important traits in domestic horse breeds. *PLoS Genet.* **9,** e1003211 (2013).
7. McCue, M. E. *et al.* A high density SNP array for the domestic horse and extant Perissodactyla: utility for association mapping, genetic diversity, and phylogeny studies. *PLoS Genet.* **8,** e1002451 (2012).
8. Bellone, R. R. Pleiotropic effects of pigmentation genes in horses. *Anim. Genet.* **41 Suppl 2,** 100–110 (2010).
9. Bellone, R. R. *et al.* Evidence for a retroviral insertion in TRPM1 as the cause of congenital stationary night blindness and leopard complex spotting in the horse. *PLoS One* **8,** e78280 (2013).
10. Brooks, S. A. *et al.* Whole-genome SNP association in the horse: identification of a deletion

in myosin Va responsible for Lavender Foal Syndrome. *PLoS Genet.* **6,** e1000909 (2010).

11. Staiger, E. A. *et al.* Host genetic influence on papillomavirus-induced tumors in the horse. *Int. J. Cancer* **139,** 784–792 (2016).
12. Sarkar, S. *et al.* Allelic Variation in CXCL16 Determines CD3+ T Lymphocyte Susceptibility to Equine Arteritis Virus Infection and Establishment of Long-Term Carrier State in the Stallion. *PLoS Genet.* **12,** e1006467 (2016).
13. Bellone, R. R. *et al.* A missense mutation in damage-specific DNA binding protein 2 is a genetic risk factor for limbal squamous cell carcinoma in horses. *Int. J. Cancer* **141,** 342–353 (2017).
14. Gaunitz, C. *et al.* Ancient genomes revisit the ancestry of domestic and Przewalski's horses. *Science* (2018). doi:10.1126/science.aao3297
15. Schubert, M. *et al.* Prehistoric genomes reveal the genetic foundation and cost of horse domestication. *Proc. Natl. Acad. Sci. U. S. A.* **111,** E5661–9 (2014).
16. Librado, P. *et al.* Tracking the origins of Yakutian horses and the genetic basis for their fast adaptation to subarctic environments. *Proc. Natl. Acad. Sci. U. S. A.* **112,** E6889–97 (2015).
17. Librado, P. *et al.* Ancient genomic changes associated with domestication of the horse. *Science* **356,** 442–445 (2017).
18. Rebolledo-Mendez, J. *et al.* Comparison of the Equine Reference Sequence with Its Sanger Source Data and New Illumina Reads. *PLoS One* **10,** e0126852 (2015).
19. Hestand, M. S. *et al.* Annotation of the Protein Coding Regions of the Equine Genome. *PLoS One* **10,** e0124375 (2015).
20. Raudsepp, T. *et al.* A 4,103 marker integrated physical and comparative map of the horse

genome. *Cytogenet. Genome Res.* **122,** 28–36 (2008).

21. Xu, X. & Arnason, U. The complete mitochondrial DNA sequence of the horse, Equus caballus: extensive heteroplasmy of the control region. *Gene* **148,** 357–362 (1994).
22. Zimin, A. V. *et al.* The MaSuRCA genome assembler. *Bioinformatics* **29,** 2669–2677 (2013).
23. Miller, J. R. *et al.* Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* **24,** 2818–2824 (2008).
24. Putnam, N. H. *et al.* Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26,** 342–350 (2016).
25. Burns, E. N. *et al.* Generation of an Equine Biobank to be Used for Functional Annotation of Animal Genomes Project. *Anim. Genet.* (in press).
26. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31,** 3210–3212 (2015).
27. Fiddes, I. T. *et al.* Comparative Annotation Toolkit (CAT) - simultaneous clade and personal genome annotation. *bioRxiv* 231118 (2017). doi:10.1101/231118
28. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–293 (2009).
29. Deng, X. *et al.* Bipartite structure of the inactive mouse X chromosome. *Genome Biol.* **16,** 152 (2015).
30. Myers, E. W. *et al.* A whole-genome assembly of Drosophila. *Science* **287,** 2196–2204 (2000).

31. Koren, S. *et al.* Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biol.* **14,** R101 (2013).

32. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5,** R12 (2004).

33. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* 1303.3997v2 (2013).

34. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21,** 1859–1875 (2005).

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

36. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10–12 (2011).

37. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

38. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).

39. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

40. Morgulis, A., Gertz, E. M., Schäffer, A. A. & Agarwala, R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J. Comput. Biol.* **13,** 1028–1040 (2006).

41. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using

exact alignments. *Genome Biol.* **15,** R46 (2014).

42. Paten, B. *et al.* Cactus: Algorithms for genome multiple sequence alignment. *Genome Res.* **21,** 1512–1528 (2011).

43. Schubert, M., Lindgreen, S. & Orlando, L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes* **9,** 88 (2016).

44. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27,** 863–864 (2011).

Figures/Tables

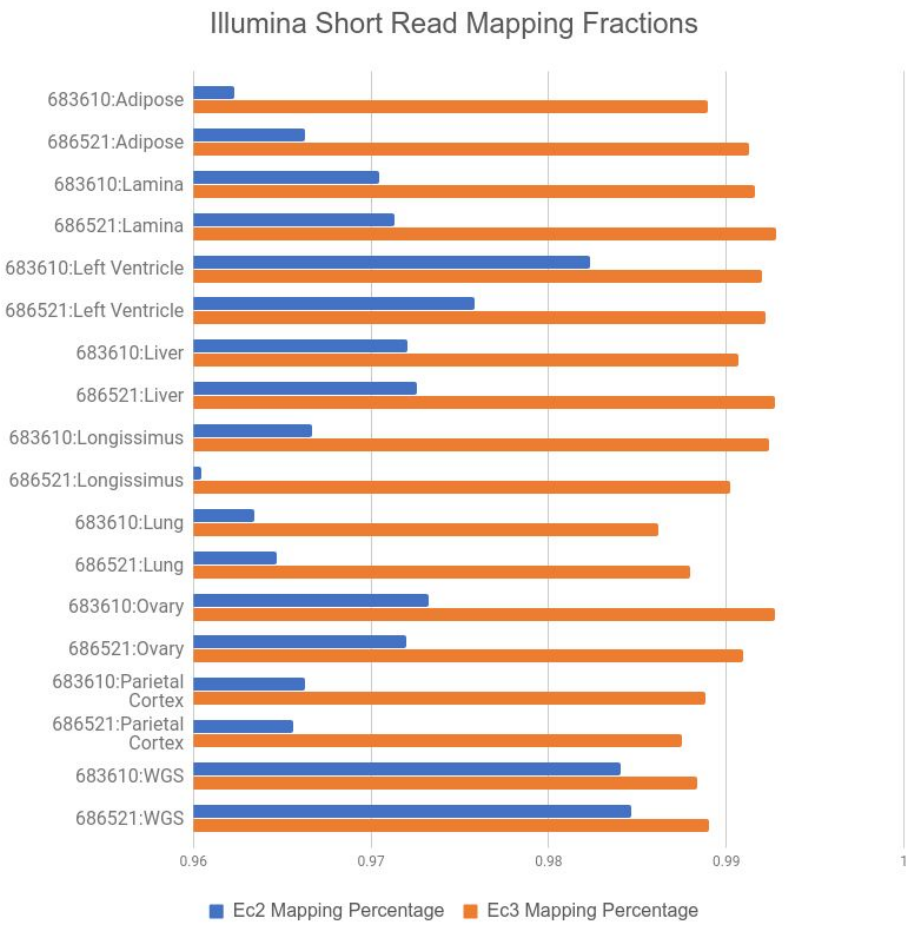

Figure 1: Percentages of RNA-seq reads from eight tissues from two horses and genomic reads mapping to EquCab2 vs. EquCab3. We used sequence data from FAANG for this mapping. More RNA-seq reads map to EquCab3 than to EquCab2 for every tissue in both horses. The percentage of genomic reads (last two rows; "WGS") mapping to EquCab3 is also larger than those mapping to EquCab2, but the difference is not as large.

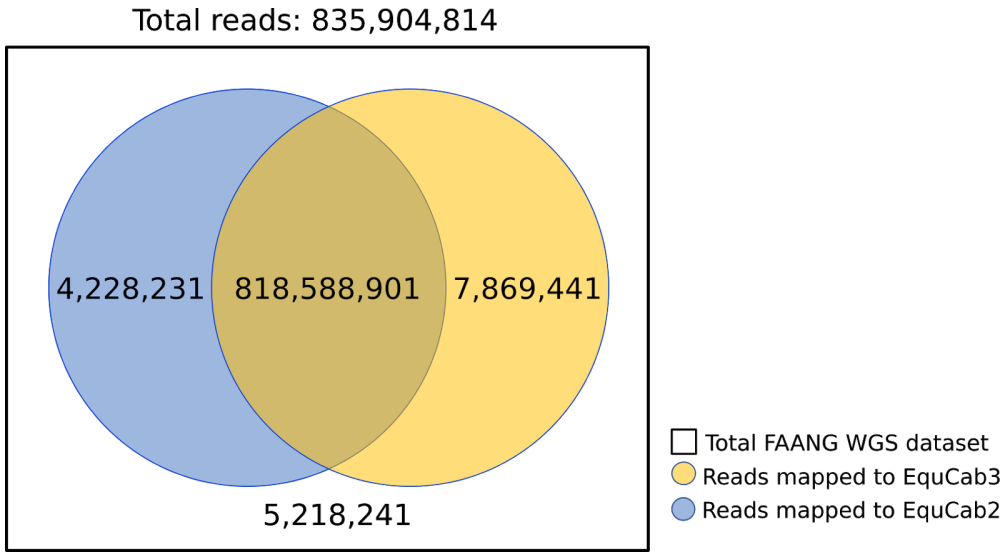

Figure 2: Number of reads from the FAANG WGS dataset mapping to EquCab2 and EquCab3. Significantly more reads map only to EquCab3 than only to EquCab2.

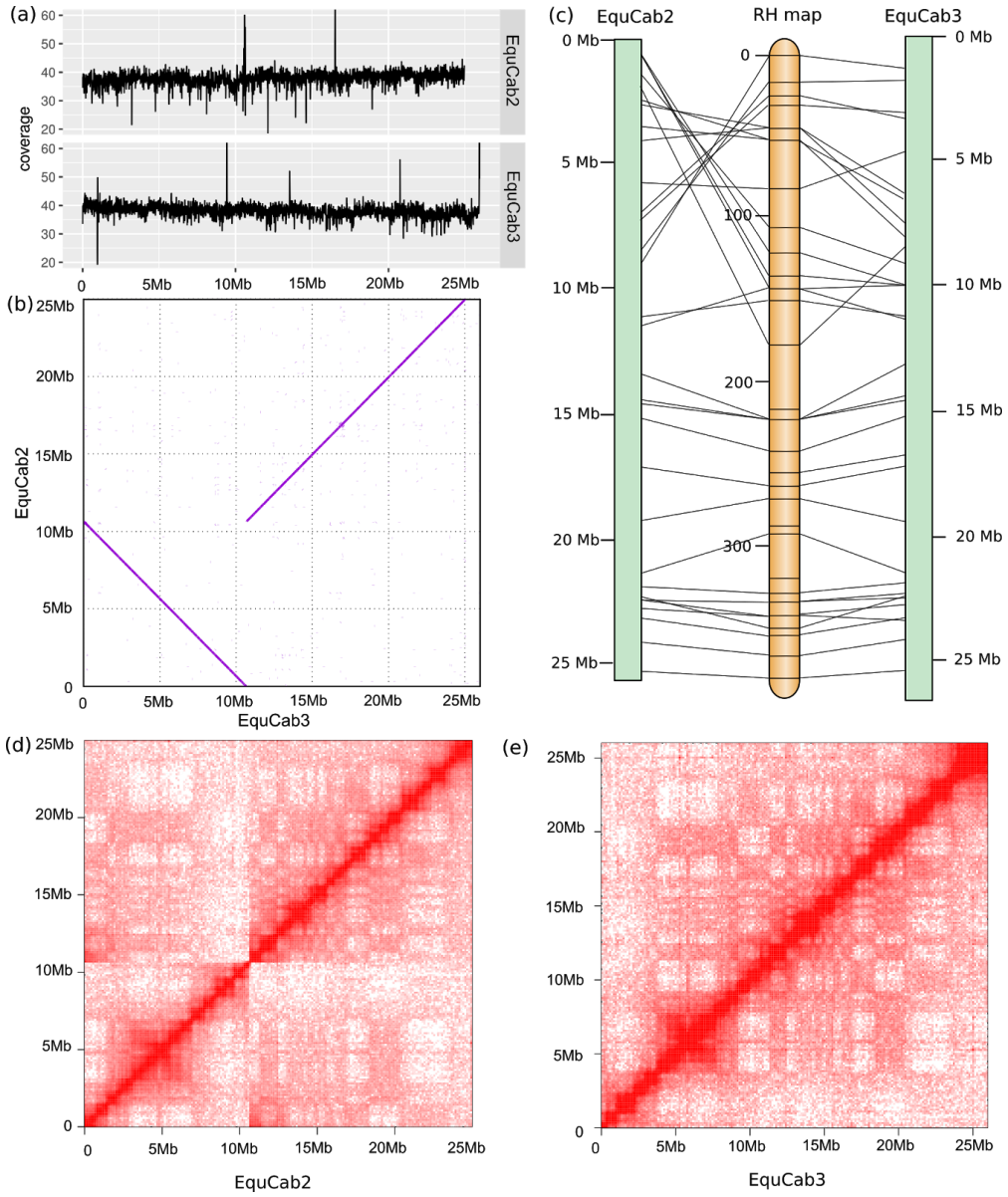

Figure 3: A comparison of equine chromosome 31 between EquCab2 and EquCab3. (a) Average coverage per 10kb window across chr31 in EquCab2 and EquCab3. EquCab3 has fewer coverage

drops and more total sequence than EquCab2. (b) An alignment of chr31 in EquCab2 and EquCab3 shows a large inversion between the two reference genomes. The RH map (c) and Hi-C contact heat maps for EquCab2 (d) and EquCab3 (e) indicate that this discrepancy is the result of a misassembly in EquCab2.

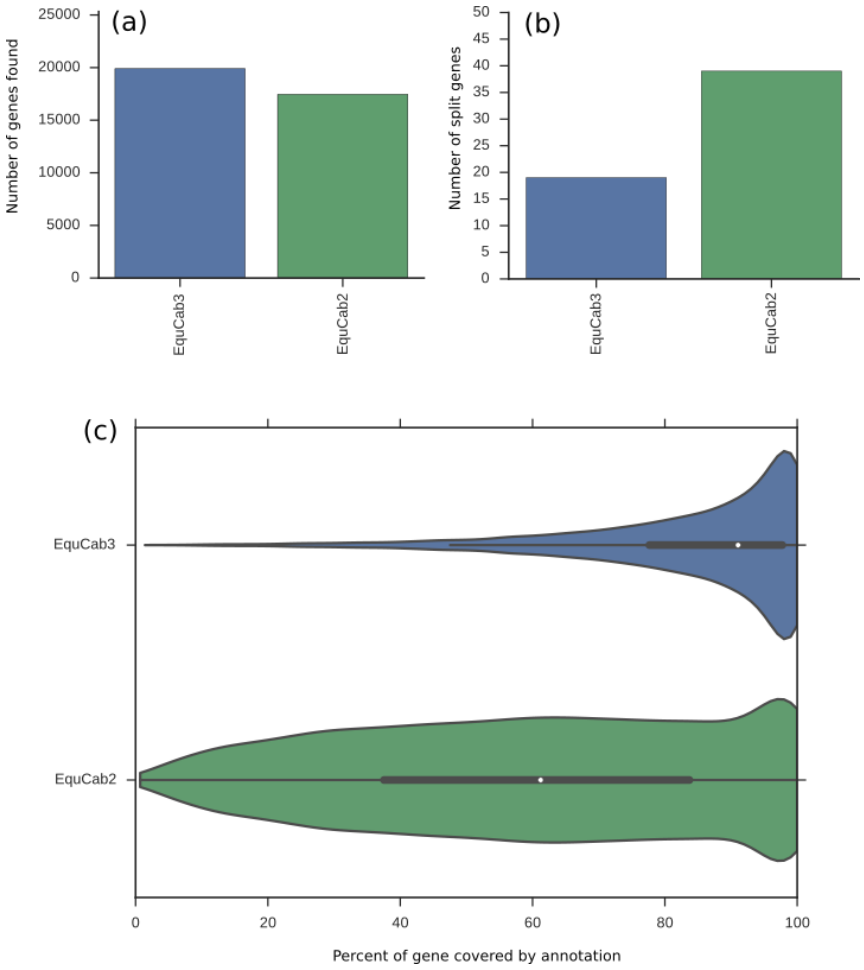

Figure 4: Annotation of EquCab2 and EquCab3 with the Comparative Annotation Toolkit shows substantial improvement in EquCab3. (a) More genes found in related species were annotated in EquCab3 than in EquCab2. (b) Fewer genes were split between contigs in EquCab3 than in EquCab2. (c) The gene coverage distribution is significantly better in EquCab3 than in EquCab2.

Table 2 attached as Excel Sheet

Chapter 3

Improved genome assembly of American alligator genome reveals conserved architecture of estrogen signaling

This chapter consists of a paper about my work on this project published in 2017 in *Genome Research*. In this project, I performed quality control and annotation on a new long-range assembly of the American alligator genome and used the new genome along with RNA-seq transcriptomic data to study the regulatory architecture of estrogen signaling during temperature-dependent sex determination. My specific contributions to this project include quality control and assessment of the genome, gene annotation of the genome, submission of the genome and annotations to NCBI, synteny analysis, PCR validation of predicted joins, all RNA-seq expression analysis, and modeling of the role of estrogen regulation in sex-biased gene expression. I wrote the manuscript with the exception of the "Comparative assembly,"

"Transposable elements," and "Small RNAs" subsections of Results; the same three subsections of Methods along with "Egg harvesting, incubation, and dissection;" and some parts of the Supplement, with significant input from Richard E. Green and Satomi Kohno and additional input and editing from other authors. I created the three figures in the main text as well as Supplemental Figures S1 and S3. S.K. and Louis J. Guillette, Jr. harvested, incubated, and dissected alligator eggs. John St. John and Jonathan Howard prepared RNA-seq libraries from embryo samples. R.E.G. designed and led the project with significant input from L.G.J., Benedict Paten, David A. Ray, Jeremy R. Sanford, and Carl Schmidt. The manuscript was reviewed and edited by three anonymous referees and the editors of *Genome Research*. I thank my coauthors for their work, without which this project would not have been possible, and for allowing me to reproduce our manuscript here.

Research

Improved genome assembly of American alligator genome reveals conserved architecture of estrogen signaling

Edward S. Rice,[1] Satomi Kohno,[2] John St. John,[3] Son Pham,[4] Jonathan Howard,[5] Liana F. Lareau,[6] Brendan L. O'Connell,[1,7] Glenn Hickey,[1] Joel Armstrong,[1] Alden Deran,[1] Ian Fiddes,[1] Roy N. Platt II,[8] Cathy Gresham,[9] Fiona McCarthy,[10] Colin Kern,[11] David Haan,[1] Tan Phan,[12] Carl Schmidt,[13] Jeremy R. Sanford,[14] David A. Ray,[8] Benedict Paten,[15] Louis J. Guillette Jr.,[16,†] and Richard E. Green[1,6,7]

[1]Department of Biomolecular Engineering, University of California, Santa Cruz, California 95064, USA; [2]Department of Biology, St. Cloud State University, St. Cloud, Minnesota 56301, USA; [3]Driver Group, LLC, San Francisco, California 94158, USA; [4]BioTuring, Incorporated, San Diego, California 92121, USA; [5]Department of Biochemistry, Stanford University, Stanford, California 94305, USA; [6]California Institute for Quantitative Biosciences, University of California, Berkeley, California 94720, USA; [7]Dovetail Genomics, LLC, Santa Cruz, California 95060, USA; [8]Department of Biological Sciences, Texas Tech University, Lubbock, Texas 79409, USA; [9]Institute for Genomics, Biocomputing & Biotechnology, Mississippi State University, Mississippi State, Mississippi 39762, USA; [10]School of Animal and Comparative Biomedical Sciences, University of Arizona, Tucson, Arizona 85721, USA; [11]Department of Animal Science, University of California, Davis, California 95616, USA; [12]HCM University of Science, Ho Chí Minh, Vietnam 748500; [13]Department of Animal and Food Sciences, University of Delaware, Newark, Delaware 19717, USA; [14]Department of Molecular, Cell, and Developmental Biology, University of California, Santa Cruz, California 95064, USA; [15]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, California 95064, USA; [16]Department of Obstetrics and Gynecology, Marine Biomedicine and Environmental Science Center, Hollings Marine Laboratory, Medical University of South Carolina, Charleston, South Carolina 29412, USA

The American alligator, *Alligator mississippiensis*, like all crocodilians, has temperature-dependent sex determination, in which the sex of an embryo is determined by the incubation temperature of the egg during a critical period of development. The lack of genetic differences between male and female alligators leaves open the question of how the genes responsible for sex determination and differentiation are regulated. Insight into this question comes from the fact that exposing an embryo incubated at male-producing temperature to estrogen causes it to develop ovaries. Because estrogen response elements are known to regulate genes over long distances, a contiguous genome assembly is crucial for predicting and understanding their impact. We present an improved assembly of the American alligator genome, scaffolded with in vitro proximity ligation (Chicago) data. We use this assembly to scaffold two other crocodilian genomes based on synteny. We perform RNA sequencing of tissues from American alligator embryos to find genes that are differentially expressed between embryos incubated at male- versus female-producing temperature. Finally, we use the improved contiguity of our assembly along with the current model of CTCF-mediated chromatin looping to predict regions of the genome likely to contain estrogen-responsive genes. We find that these regions are significantly enriched for genes with female-biased expression in developing gonads after the critical period during which sex is determined by incubation temperature. We thus conclude that estrogen signaling is a major driver of female-biased gene expression in the post-temperature sensitive period gonads.

[Supplemental material is available for this article.]

The American alligator, *Alligator mississippiensis*, like all crocodilians and many other reptiles, has temperature-dependent sex determination (TSD), in which the sex of an embryo is determined by the incubation temperature of its egg during a temperature-sensitive period (TSP) of development (Ferguson and Joanen 1982). In contrast, mammals, birds, and other animals with genetic sex determination (GSD) rely on sex chromosomes to trigger sex determination. These genetic differences induce sex differentiation during development by causing differential expression of numerous genes. Genes with sex-biased expression during development in these lineages include conserved sexual development genes such as *SOX9* and *WNT4* (De Santa Barbara et al. 1998; Hsieh et al. 2002). Such expression differences eventually cause the development of one of two sets of distinct sexual characteristics.

Corresponding authors: esrice@ucsc.edu, ed@soe.ucsc.edu
[†]Deceased August 6, 2015.
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.213595.116.

27:686–696 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/17; www.genome.org

However, in alligators and other species with TSD, males and females have identical genomes, leaving open the question of how differences in temperature lead to differential expression of genes between males and females during early development (Morrish and Sinclair 2002; Shoemaker-Daly et al. 2010; Kohno and Guillette 2013).

Insight into this question comes from the observation that exposing an alligator embryo to exogenous estrogen while incubated at a male-producing temperature (MPT) causes it to develop ovaries instead of testes. Estrogen, whose presence is detected and transduced via the transcription factor estrogen receptor alpha (Bull et al. 1988; Milnes et al. 2005; Kohno et al. 2015), is an early effector of sexual development genes in the American alligator, as it is in other vertebrates, including both species with GSD and TSD (Crews et al. 1989; Nakabayashi et al. 1998). In addition, *CYP19A1*, the gene coding for the enzyme aromatase, which converts androgen to estrogen, is expressed at significantly higher levels in embryos incubated at female-producing temperature (FPT) than those incubated at MPT (Gabriel et al. 2001). These two observations have led to the hypothesis that estrogen signaling is a master regulator of sex-biased gene expression in alligator embryos (Lance 2009). While it is clear that estrogen plays a critical role in inducing ovarian development at MPT, there is currently no direct evidence that the genes targeted by estrogen are actually involved in early TSD for embryos incubated at MPT.

Much work has been performed in alligators and other vertebrates with TSD to determine the initial switch that links temperature to sexual fate (Kohno et al. 2010; Schroeder et al. 2016) and the cause of increased expression of aromatase at FPT (Parrott et al. 2014; McCoy et al. 2016). One recent hypothesis for the gene acting as the initial switch in the American alligator is the thermosensitive TRP channel *TRPV4*, as it is activated at temperatures near MPT in vitro and targets gene expression of male development genes (Yatsu et al. 2015). However, less attention has been paid to the downstream effects of increased aromatase expression in these species, especially in terms of which genes are regulated by estrogen.

Estrogen signaling is best understood in humans, including the genes it targets and its role in sexual development. Whether these mechanisms and downstream effects are conserved in other vertebrates, including those with TSD, remains unknown. In humans, estrogen regulates gene expression through the transcription factors estrogen receptor alpha and beta, coded for by the genes *ESR1* and *ESR2*, respectively. The estrogen 17β-estradiol activates an estrogen receptor by binding to its ligand-binding domain, thus allowing the receptor's DNA-binding domain to bind to a well-defined enhancer sequence, the estrogen response element, promoting the expression of nearby genes (Nilsson et al. 2001; Dahlman-Wright et al. 2006). The motif to which human estrogen receptor alpha binds has been well characterized using chromatin immunoprecipitation (Gruber et al. 2004; Laganière et al. 2005). A majority of estrogen receptor alpha binding sites are distal enhancers—that is, they are far from the genes they regulate (Carroll et al. 2006; Lin et al. 2007; Welboren et al. 2009).

A majority of estrogen receptor binding events are associated with long-range intrachromosomal chromatin interactions, and these associated events are significantly enriched for RNA polymerase II recruitment (Fullwood et al. 2009). The zinc finger protein CTCF is responsible for many of these chromatin interactions (Zhang et al. 2010). Regions delineated by two CTCF binding sites that contain an estrogen receptor binding site are significantly more likely to contain estrogen-responsive genes in humans (Chan and Song 2008). It is currently unknown whether ESR1 and CTCF binding sites are predictive of estrogen-responsive regions in the genomes of other vertebrates or whether CTCF-mediated long-range chromatin interactions are involved in estrogen's inducement of female development in vertebrates with TSD. Because the estrogen response is a long-range phenomenon in humans, a contiguous genome assembly is necessary to fully explore the genome architecture of estrogen regulation in alligators.

Green et al. (2014) published the genomes of the American alligator and two other crocodilians: the saltwater crocodile *Crocodylus porosus* and the gharial *Gavialis gangeticus*, with scaffold N50s of 508 kb for the American alligator, 205 kb for the saltwater crocodile, and 127 kb for the gharial. The slow rate of molecular evolution within crocodilians (Green et al. 2014) makes this clade ideal for testing the ability to use a highly-contiguous genome assembly to scaffold the genome assemblies of related organisms based on synteny.

Results

Assembly and annotation

The updated American alligator genome assembly AllMis2 has a total length of 2.16 Gbp compared with 2.17 Gbp for the previously published assembly AllMis1, a difference within the range of variance between assembler runs. However, AllMis2 shows a 25-fold improvement in scaffold N50, a measure of contiguity, from 508 kbp to >13 Mbp.

To assess the quality and accuracy of AllMis2, we measured its concordance with previously published BAC-end pairs (Shedlock et al. 2007) that were not used in the assembly or scaffolding. By using BWA MEM (Li 2013) with default parameters, we aligned the forward and reverse reads of the 1309 BAC-end pairs to the new assembly and to the assembly prior to scaffolding using Chicago data. We found that while 142 BAC-end pairs had both ends aligning to the same scaffold of our assembly before scaffolding with Chicago, 1160 BAC-end pairs have both ends aligning to the Chicago-scaffolded assembly: 1143, or 98.5%, of these pairs aligning to the same scaffold are oriented correctly, and 1125, or 98.4%, of these correctly oriented pairs have an insert size between 70 and 180 kb. We thus conclude that AllMis2 is both accurate and an improvement over assembly not using the Chicago library.

We annotated AllMis2 for protein-coding genes using previously published RNA-seq reads (Green et al. 2014) and AUGUSTUS (Stanke et al. 2006), finding 32,052 transcripts and 24,713 genes. Moreover, we were able to assign names to 15,977 of these genes based on orthology with named genes in other vertebrate species. By use of both orthology and protein sequence analysis, we assigned 5960 unique Gene Ontology (GO) terms to 17,430 American alligator proteins.

Crocodilian versus mammalian genome synteny

While the previous assembly of the American alligator genome AllMis1 (Green et al. 2014) was sufficient to compare to other genomes at the sequence level, our new long-range assembly AllMis2 presents an opportunity to perform genome comparisons on a broader scale. We computed synteny between the alligator and chicken (Galgal4) genomes using SyMAP v4.2 (Soderlund et al. 2011). We used both AllMis2 and AllMis1 for comparison and found that the increased contiguity of AllMis2 vastly improved our ability to compute synteny between the chicken and alligator genomes, more than doubling the percentage of the

genome covered by synteny blocks from 35% to 90% and increasing the sizes of synteny blocks, with 57 of the 90 synteny blocks >10 Mb in length. Most scaffolds in the new alligator assembly correspond to a contiguous region of a chicken chromosome, although often with some intrachromosomal rearrangements (Fig. 1A–C). Some scaffolds in the alligator genome appear to

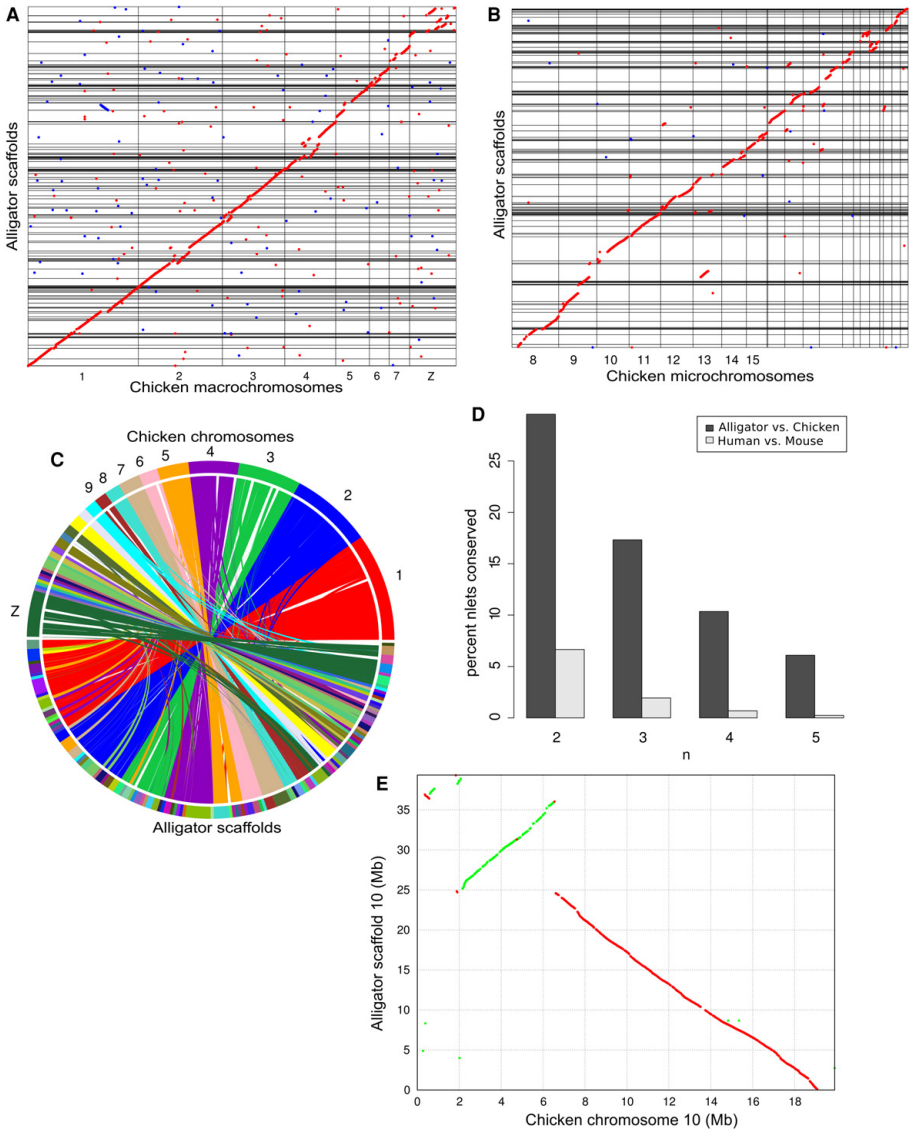

Figure 1. Our new long-range assembly of the American alligator genome allows analysis of the synteny between crocodilians and birds. (*A*,*B*) Dot plots of an anchored whole-genome alignment between the chicken and American alligator genomes show a high degree of synteny, with many long alligator scaffolds covering significant portions of chicken chromosomes, including macrochromosomes (*A*) and microchromosomes (*B*). (*C*) A circle plot of synteny between the alligator and chicken genomes made using SyMAP (Soderlund et al. 2011). (*D*) Conservation of ordered gene doublets, triplets, quadruplets, and quintuplets between alligators and chickens versus between humans and mice, showing much higher synteny between alligators and chickens than between humans and mice. (*E*) Alligator scaffold 10 covers a vast majority of the chicken microchromosome 10. However, there are several small inversions and one large inversion between the two. Green and red dots represent forward and reverse matches, respectively.

correspond to whole arms of chicken chromosomes. For example, two alligator scaffolds almost completely cover GGA7. Furthermore, the microchromosome GGA10 is almost fully covered by a single alligator scaffold, scaffold 10 (Fig. 1E), with one large inversion and numerous small local inversions.

To contrast the levels of genome rearrangement in archosaurs and mammals, we compared conservation of gene order between alligators and chickens (242 Mya TMRCA) to that between humans and mice (110 Mya TMRCA) (Crottini et al. 2012). We calculated the percentage of ordered pairs, triplets, quadruplets, and quintuplets of directly adjacent genes that occur in both alligators and chickens and both humans and mice. We found four times greater conservation of gene pair synteny between alligators and chickens than between humans and mice, nine times greater conservation of gene triplets, 15 times greater conservation of quadruplets, and 25 times greater conservation of quintuplets (Fig. 1D).

A closer look at synteny between the chicken Z Chromosome and the alligator genome reveals the expected inversion around the avian sex-determining gene *DMRT1* (Supplemental Fig. S1). This result is concordant with the *Z*-linked inversions previously predicted by examining gene synteny between the avian Z Chromosome and other reptilian outgroups such as the green anole *Anolis carolinensis*, red-tailed boa *Boa constrictor*, and Mexican musk turtle *Staurotypus triporcatus* (Kawagoshi et al. 2014; Zhou et al. 2014). While these studies show that this inversion occurred after the divergence of archosaurs from other amniotes, our result further pinpoints the time of the beginning of evolution of avian sex chromosomes by providing the first conclusive evidence that this inversion occurred in the common ancestor of birds after divergence with crocodilians.

Comparative assembly

We used the American alligator genome to scaffold the previously published genome assemblies of two other crocodilians, the saltwater crocodile *C. porosus* and the gharial *G. gangeticus*, based on synteny. These published assemblies have scaffold N50s of 205 and 127 kb, respectively. We performed comparative assembly on these genomes with Ragout (Kolmogorov et al. 2014). Through this process, we were able to increase the scaffold N50 of the saltwater crocodile genome assembly from 205 kb to 84 Mb and the gharial genome assembly from 128 kb to 96 Mb. For comparison, the mean chromosome sequence length of the saltwater crocodile and gharial genomes are 117 and 165 Mb, respectively.

To assess the accuracy of the synteny based scaffolding, we tested a random set of the scaffold joins predicted by Ragout for each species. We verified predicted scaffold joins using PCR with primers chosen such that the amplified regions would be unique in the genome assembly and would span the joins made by Ragout. We successfully amplified these gap regions for 18 out of 20 predicted joins tested in the saltwater crocodile genome and 22 out of 29 predicted joins tested in the gharial genome. Full results and primers used for join verification are in Supplemental Table S1.

Transposable elements

Repetitive sequences comprise more than one-third of the alligator genome assembly (Supplemental Table S2). Almost a quarter of the genome is derived from just three TE superfamilies: LINE CR1s (12.2%) and the DNA transposons Harbinger (7.5%) and hAT (8.2%). TEs in general appear to accumulate more slowly in crocodilians than in other vertebrate taxa (excluding Testudines), and few new TE families, or even insertions, appear in any lineage of crocodilians since their divergence (Green et al. 2014; Suh et al. 2015). Data from AllMis2 are consistent with these findings. Repeat content in general and from each of the dominant superfamilies are similar not only between alligator assemblies but also among crocodilians (Supplemental Table S2), as determined by premasked genomes (http://repeatmasker.org/genomicDatasets/RMGenomicDatasets.html; accessed March 15, 2016). Only CR1 content varies between alligator assemblies to an appreciable degree. An additional 2.6% of the AllMis2 assembly is identifiable as CR1 compared with that of AllMis1. The differences in CR1 content between assemblies may be greater than it seems when contrasted with the near uniformity in the TE annotations across existing crocodilian assemblies (Supplemental Table S2; Green et al. 2014). Highly repetitive, nearly identical sequences are difficult to assemble from short reads and are likely underrepresented in genome assemblies, so an improved assembly may be able to identify these to a greater degree. Repeats in both AllMis1 and AllMis2 are biased toward those >10% diverged from their respective consensus element (Supplemental Fig. S2). No clear "burst" of CR1 activity specific to any one divergence bin is apparent, so it is likely that the additional CR1 insertions are distributed among elements with high and low mutation loads.

Small RNAs

MicroRNAs have been identified de novo in model vertebrate species, but for nonmodel species, miRNAs are usually identified based on sequence conservation with known miRNAs in other species. We sequenced a library of small RNAs isolated from alligator testis and used the resulting reads to predict 60 putative miRNAs after filtering for quality, including one, aca-mir-425, which appears in the American alligator, saltwater crocodile, and gharial genomes, but not in the chicken genome. See Supplemental Results for more details.

Sex-biased gene expression

A crucial step toward understanding TSD in the American alligator is determining which genes are turned on or off based on temperature at various developmental stages. This necessitates the generation of a catalog of genes that show significantly different expression between eggs incubated at MPT and those incubated at FPT. To this end, we incubated a total of 168 alligator eggs at either MPT or FPT for either 0, 3, or 30 d after developmental stage 19. The TSP spans developmental stages 21 to 24 (Lang and Andrews 1994), which occur between our 3- and 30-d timepoints. We harvested the embryos after incubation, subdissected the gonad-adrenal-mesonephros (GAM) complex into its constituent parts, and performed RNA sequencing on each of these three tissues for each sample. We sequenced at least three biological replicates from different clutches for each tissue and time point combination. See Supplemental Table S3 for a list of libraries sequenced along with their NCBI accessions.

We used the resulting RNA-seq data to quantify gene expression and determine which genes are differentially expressed between developing male and female embryos at these developmental stages in these three tissues. We used Cuffdiff 2 to perform these tasks (Trapnell et al. 2013). Cuffdiff 2 generates a normalized expression value in fragments per kilobase of transcript per million mapped reads (FPKM) for each gene in each library as well as an FDR-adjusted *P*-value for determining whether gene expression is significantly different between two

sets of replicates. We considered any gene with an FDR-adjusted $P \leq 0.05$ to be differentially expressed between males and females in a given tissue at a given time point.

Due to conditions prior to egg collection, embryos can sometimes develop as a different sex than expected based on incubation temperature after collection (McCoy et al. 2015). We could not confirm sex histologically as both gonads of each embryo were used for RNA sequencing, so we confirmed the sex of each embryo by comparing gonadal expression of *CYP19A1* to *AMH* as in previous studies (Kohno et al. 2015; McCoy et al. 2015). One embryo from clutch 13 was female despite incubation at MPT (Fig. 2B), so we excluded it from differential expression analysis.

We found many genes with differential expression between males and females in each tissue at both the 3- and 30-d timepoints (Fig. 2A; Supplemental Table S4). Unsurprisingly, the gonads at the post-TSP time point displayed the most sexual dimorphism in gene expression. The genes differentially expressed between male and female embryos in these samples include many genes known to be involved in early sexual development in other vertebrates (Fig. 2B). Such male development genes include *SOX9*, which triggers testis formation, and *AMH*, which inhibits the formation of Müllerian ducts (De Santa Barbara et al. 1998). Female development genes with female-biased expression in the post-TSP gonads include *FST*, which inhibits the production of follicle-stimulating hormone (Ying et al. 1987). *CYP19A1*, which produces aromatase, the enzyme that converts androgens to estrogens (Toda and Shizuta 1993), was the gene with the largest sex-bias fold-change in either direction, with a $\log_2$ fold-change of 12.463. This is consistent with other studies of aromatase expression in embryos incubated at different temperatures (Smith et al. 1995; Gabriel et al. 2001). *ESR1*, the gene coding for estrogen receptor alpha, and *CTCF* are highly expressed in both male and female gonads at this time point, with respective average FPKM values of 24.08 and 47.02 but no significant sex bias.

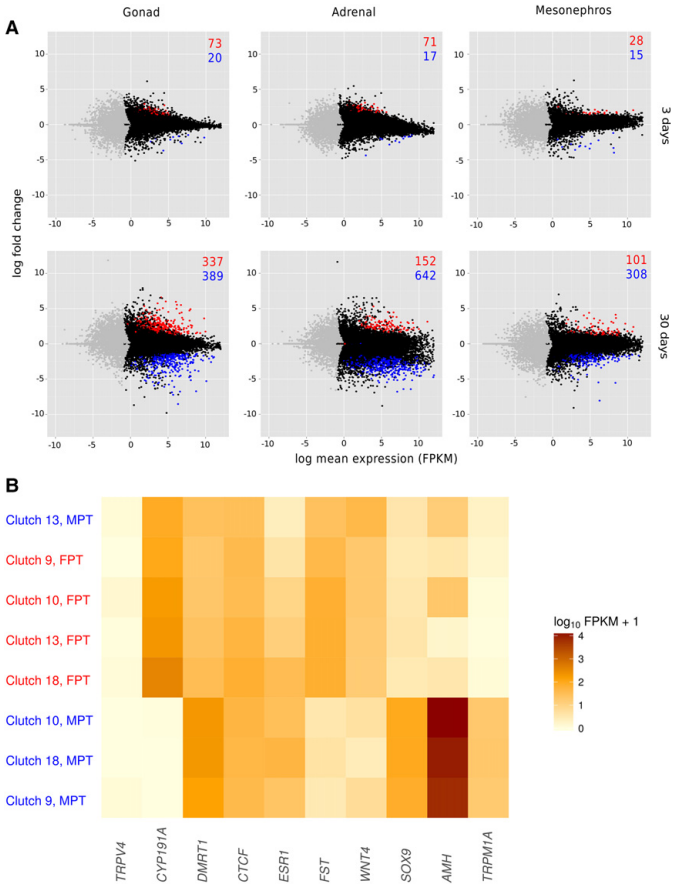

Figure 2. Sex-biased gene expression in alligator embryos. (*A*) Mean expression versus fold-change for all genes in three tissues at two developmental time points. Genes found to have female-biased expression and male-biased expression are colored in red and blue, respectively. Numbers of sex-biased genes for each tissue and time point are given in the *upper right* of each plot. (*B*) Gonadal expression of genes of interest at the 30-d time point in eight embryos. The embryo from clutch 13 incubated at MPT displays a distinctly female expression pattern despite being incubated at MPT and was thus excluded from further analyses.

We have included lists of significantly enriched GO terms among genes with male- and female-biased expression in the gonads at 30 d generated using FUNC (Prüfer et al. 2007) in Supplemental Table S5. One significantly overrepresented GO term among these male-biased genes is "detection of temperature stimulus" (GO:0016048). The only male-biased gene with this GO term is the transient receptor potential cation channel *TRPM1A*. Another transient receptor potential cation channel gene, *TRPV4*, has been suggested as one thermosensitive gene involved in TSD in the American alligator (Yatsu et al. 2015). We found no significant expression or sex-bias of *TRPV4* at any of our time points in any of the three tissues. However, Yatsu et al. (2015) found sex-biased expression of *TRPV4* only during the TSP at developmental stages 21 and 24, while we sampled only before and after the TSP.

Estrogenic regulation of gene expression

Estrogen regulation of gene expression is best understood in humans from work dissecting the molecular basis of estrogen-responsive and nonresponsive breast cancers in tissue models. That work has shown that in human estrogen-responsive tissues, estrogen promotes the expression of genes by allowing estrogen receptors to bind to enhancer DNA sequences (Dahlman-Wright et al. 2006). However, the enhancers to which estrogen receptors bind are usually distal to

the genes they regulate (Carroll et al. 2006). Due to the sex-reversing effects of estrogen exposure during crocodilian development via estrogen receptor alpha (Kohno et al. 2015) and the extreme female-biased expression of the gene coding for aromatase, we hypothesize that estrogen signaling through ESR1 binding is a major driver of female-biased gene expression during TSD in the American alligator.

We first tested this hypothesis by looking for enrichment of genes with female-biased expression in the post-TSP gonads of alligator embryos in the genomic regions surrounding computationally predicted estrogen receptor binding sites. The DNA-binding domain of ESR1 is perfectly conserved among humans, chickens, and alligators (Supplemental Fig. S3b; Supplemental Table S6), and the DNA-binding motif of ESR1 in human estrogen-responsive cells is well characterized (Gruber et al. 2004; Carroll et al. 2006; Lin et al. 2007). Therefore, we predicted ESR1 binding sites in the American alligator genome using the motif representing the human estrogen response element. We found that while 337 (2.26%) of the 14,943 genes expressed in the post-TSP gonad have female-biased expression, 62 (3.11%) of the 1991 expressed genes within 50 kb of a putative estrogen receptor binding site have female-biased expression (E = 44.9; enrichment factor = 1.38; Fisher's exact test $P = 4.79 \times 10^{-3}$). This indicates that genes are significantly more likely to have female-biased expression in the post-TSP gonad if they are near a location in the genome where ESR1 is predicted to bind.

In human tissue models of estrogen regulation of gene expression, whether a gene is likely to be estrogen responsive is based on its genomic location relative to not only estrogen receptor binding sites but also CTCF binding sites (Chan and Song 2008). Since this was established in 2008, studies of CTCF-mediated chromatin looping have shown that CTCF helps divide the genome into functional domains through a chromatin extrusion process that causes loops to form only where two adjacent CTCF binding motifs are oriented toward each other (Rao et al. 2014; Sanborn et al. 2015).

CTCF binding sites in the chicken genome have been experimentally determined (Martin et al. 2011), and the zinc finger domains of CTCF are perfectly conserved among human, chicken, and alligator orthologs (Supplemental Fig. S3a). We therefore used the CTCF binding motif in the chicken genome to predict CTCF binding sites in the American alligator genome. We used these binding site predictions and the most recent model of CTCF-mediated chromatin looping (Sanborn et al. 2015) to predict how chromatin loops form in the alligator genome (Fig. 3A). We predicted 19,482 chromatin loops based on CTCF binding sites, comparable to the 21,306 found experimentally in the human genome (Li et al. 2012); 3758 (19.3%) of these putative loops contain one or more ESR1 binding sites, and 10,074 (67.4%) of the 14,943 genes expressed in gonads after 30 d of incubation are within the boundaries of one or more predicted CTCF loops.

We found that while 337 (2.26%) of the 14,943 genes expressed in the post-TSP gonads have female-biased expression, 116 (3.09%) of the 3759 expressed genes in CTCF loops containing one or more ESR1 binding sites have female-biased expression (E = 84.8, enrichment factor = 1.37; Fisher's exact test $P = 7.76 \times 10^{-5}$). This finding shows a significant enrichment in female-biased gene expression in the regions of the genome predicted to be estrogen responsive under our model, providing support for our hypothesis that many of these genes are regulated by estrogen during sexual differentiation and development (Fig. 3B,C).

Among the female-biased genes in predicted estrogen-responsive regions is *WNT4*, a gene required for female development

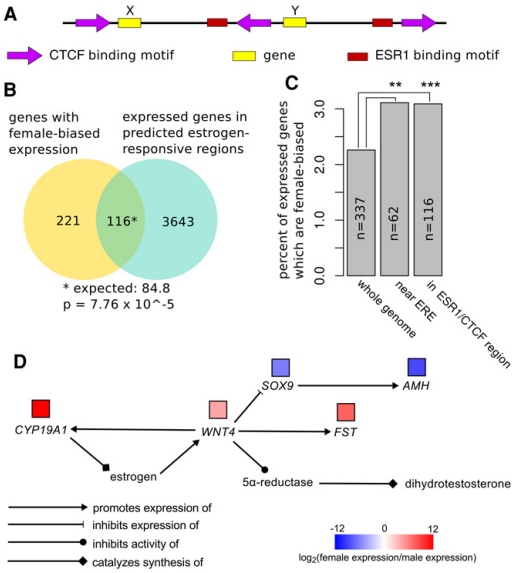

Figure 3. Genes in regions of the genome predicted to be under estrogenic regulation of gene expression are significantly more likely to be female biased in the post-TSP gonads. (*A*) Our model for predicting regions of the genome under estrogenic regulation of gene expression, based on the CTCF extrusion model (Sanborn et al. 2015) and the Chan and Song model of estrogen receptor binding site activity (Chan and Song 2008). In this example, Gene X is predicted to be estrogen responsive and Gene Y is not because Gene X is between two inward-oriented CTCF binding motifs along with an ESR1 binding site, while Gene Y is not. (*B*) Of the 14,943 genes expressed in the post-TSP gonads, 337 have female-biased expression and 3759 are in predicted estrogen-responsive genomic regions. However, 116 of these genes are both female biased and within predicted estrogen-responsive regions, a significantly higher number than the expected 84 ($P = 7.76 \times 10^{-5}$). (*C*) Percentages of expressed genes with female-biased expression in the whole genome versus near an estrogen response element and in a predicted estrogen-responsive CTCF region. Regions near an estrogen response element and predicted estrogen-responsive regions are both enriched for female-biased genes. (**) $P \leq 0.01$; (***) $P \leq 10^{-4}$. (*D*) Pathway diagram showing results of increased *CYP19A1* expression after the TSP in the gonads of embryos incubated at FPT. Sex-bias fold-changes for each gene in the pathway are shown in boxes *above* the genes.

in other vertebrates. *WNT4* suppresses *SOX9* and 5-alpha reductase activity (Fig. 3D) and promotes the formation of Müllerian ducts via frizzled receptor binding (Hsieh et al. 2002). Frizzled receptor genes *FZD2*, *FZD3*, *FZD6*, *FZD8*, and *FZD9* are all significantly expressed in the post-TSP gonads in both males and females. We therefore hypothesize that *WNT4* plays a role in sex differentiation in the American alligator similar to its role in other vertebrates, although unlike in vertebrates with GSD, its expression is determined by incubation temperature via estrogen signaling.

Discussion

We present AllMis2, an improved assembly of the American alligator (*A. mississippiensis*) genome. After demonstrating its accuracy, we used AllMis2 to examine synteny between the American

alligator and chicken (*Gallus gallus*) genomes, improve the genomes of two other crocodilian species, and predict genomic regions likely to be under estrogenic regulation of gene expression in estrogen-responsive tissues. Finally, we showed that genes in these predicted estrogen-responsive regions are significantly more likely to have female-biased expression in post-TSP gonads. We thus conclude that the genomic architecture of estrogen signaling is remarkably well conserved within vertebrates and that it is a fundamental early driver of female-biased gene expression in the post-TSP embryonic gonads of the American alligator.

Our analyses are aided by a contiguous genome, and many would not have been possible with AllMis1. Synteny blocks between the chicken genome and AllMis1 were too small and fragmented to lend significant insight to large-scale genome evolution between avians and crocodilians, while a whole-genome alignment between the chicken genome and AllMis2 shows many large synteny blocks with some inversions covering significant portions of chicken chromosomes. Synteny analysis using AllMis2 also reveals a slower rate of gene rearrangement in archosaurs than in mammals (Fig. 1C), and the first direct evidence that the initial inversion leading to the evolution of avian sex chromosomes occurred after the divergence of the crocodilian and avian lineages (Supplemental Fig. S1). Furthermore, transposable element annotation was improved by using AllMis2.

Highly repetitive, low-diversity sequences (i.e., recently active TEs) are among the most difficult to assemble, and it is likely that their presence is underestimated in genome assemblies. This could downwardly bias estimates of TE content and would particularly affect estimates of recently active TEs. It is possible that AllMis2 better represents the true TE content of the alligator genome. CR1 content increased by 2.6% between alligator assemblies (Supplemental Table S2), but sequence diversity within CR1 is similar in both assemblies (Supplemental Fig. S2). Analysis of AllMis1 suggested that TEs in general accumulate more slowly in crocodilians than in other vertebrate taxa (excluding Testudines), and few new TE families, or even insertions, have appeared in any lineage of crocodilians since their divergence (Green et al. 2014; Suh et al. 2015). Some of the variation in CR1 annotations between alligator assemblies is almost certainly due to stochasticity introduced by homology-based identification. Further, it is possible that comparable improvements to the gharial and crocodile assemblies would yield similar changes in CR1 annotation. Observations made when comparing alligator assemblies (overall increased CR1 content, few young CR1 elements) combined with our understanding of CR1 evolution in crocodilians in general (Suh et al. 2015) imply that the new alligator assembly was slightly more useful for identifying TEs.

Holleley et al. (2015) recently discovered that although the Australian bearded dragon *Pogona vitticeps* has heteromorphic sex chromosomes, it can undergo sex reversal in the wild at high temperatures. In addition, during extended hot periods, whole populations can lose their minor sex chromosomes and transition to fully TSD populations. In the context of earlier reports of thermal and hormonal overrides for GSD in several species of lizards and turtles (Barske and Capel 2008), these observations indicate that at least some components of sex determination remain sensitive to temperature even when genetic cues evolve that can override them. Here, we show that the effectors of estrogen signaling and its underlying genomic architecture are highly conserved between TSD and GSD lineages. The protein sequence of the DNA-binding domains of both ESR1 and CTCF is perfectly conserved in the alligator, human, and chicken genomes. The CTCF/ERE model for estrogen response (Chan and Song 2008) developed in estrogen-responsive human tissue culture models is predictive of female-biased gene expression in the developing alligator embryo. Aromatase and two of its downstream genes involved in sexual development in other vertebrates, *WNT4* and *SOX9*, are all differentially expressed in a temperature-dependent manner in the developing alligator embryo and in the embryos of other TSD reptiles like the red-eared slider turtle *Trachemys scripta elegans* (Ramsey and Crews 2009). We propose that some aspects of the highly conserved estrogen response may be inherently and persistently temperature sensitive. All of the 22 species of Crocodilia use TSD (Lang and Andrews 1994). Within this clade, the proposed direct link between temperature and estrogen signaling may have evolved robustness sufficient to be impervious to genetic variation. A comprehensive experimental exploration of the estrogen response in TSD versus GSD species may reveal the biochemical link between temperature and estrogen signaling.

Expression of aromatase, the enzyme that produces estrogen, has been hypothesized to be a master regulator of sex-biased gene expression in developing alligator embryos (Lance 2009) because of the ability of estrogen exposure to cause sex reversal in embryos incubated at MPT (Bull et al. 1988) and its extreme sex-biased expression in embryonic gonads after TSP (Gabriel et al. 2001). While much work is currently being performed to determine the pathway that allows aromatase expression to vary with temperature (Parrott et al. 2014; Yatsu et al. 2015; McCoy et al. 2016), less attention has been paid to the questions of which genes estrogen regulates during sexual development in American alligators or how estrogen regulates them despite its pivotal role early in embryonic sexual differentiation in alligators. Our data do not speak to the hypothesis that *TRPV4* is a component of the temperature-sensing apparatus responsible for TSD (Yatsu et al. 2015) as we find no evidence of expression of this gene in any tissue at any of our time points. Importantly, we took samples before and after the TSP. Future work measuring gene expression during the TSP may more clearly determine the roles of *TRPV4*, *TRPM1*, and perhaps other candidate thermosensitive signaling molecules.

In this article, we hypothesized that estrogen regulates gene expression in developing American alligator embryos through the same mechanism by which it is known to do so in humans and that this mechanism can explain much of the female-biased gene expression that occurs after the TSP. By using the latest model of estrogen regulation of gene expression and CTCF-mediated chromatin looping in humans, we demonstrated that the regions of the American alligator genome that are most likely to be under estrogenic regulation of gene expression are enriched for female-biased gene expression. Our results provide new evidence for Lance's hypothesis that aromatase and its production of estrogen are a major driver of sex-biased gene expression in TSD in the American alligator (Lance 2009). These results show that despite the different roles of estrogenic regulation of gene expression in sexual development between humans and alligators, much of the underlying mechanism responsible for estrogen regulation of gene expression is conserved between these two species.

Although our study does not fully elucidate the downstream effects of female-biased gene expression caused by estrogen signaling in the post-TSP gonads, *WNT4*'s female-biased expression and presence in a predicted estrogen-sensitive region provide a possible explanation for some of these effects. In mammals, *WNT4* expression prevents the formation of male-specific vasculature by preventing migration of endothelial and steroidogenic cells from mesonephros tissues to gonads (Jeays-Ward et al. 2003). It

performs this action through up-regulation of follistatin (Yao et al. 2004). *FST*, the gene coding for follistatin, is among the genes we find to have female-biased expression in the post-TSP alligator gonad, suggesting that *FST* may be among the genes indirectly regulated by estrogen signaling after the TSP. Furthermore, *WNT4* promotes expression of aromatase in mammals (Boyer et al. 2010). If the same is true in post-TSP embryonic alligator gonads, *WNT4* and aromatase may cooperate through a feed-forward mechanism in which estrogen promotes the expression of *WNT4* and *WNT4* promotes the expression of aromatase, which then creates more estrogen.

Methods

Sequencing and assembly

DNA was extracted with Qiagen blood and cell midi kits according to the manufacturer's instructions. Briefly, cells were lysed and centrifuged to isolate the nuclei. The nuclei were further digested with a combination of Proteinase K and RNase A. The DNA was bound to a Qiagen genomic column, washed, eluted and precipitated in isopropanol, and pelleted by centrifugation. After drying, the pellet was resuspended in 200 μL TE (Qiagen). We generated the Chicago library as previously described by Putnam et al. (2016). Briefly, high-molecular-weight DNA was assembled into chromatin in vitro and then chemically cross-linked before being restriction digested. The overhangs were filled in with a biotinylated nucleotide, and the chromatin was incubated in a proximity-ligation reaction. The cross-links were then reversed, and the DNA purified from the chromatin. The library was then sonicated and finished using the NEB ultra library preparation kit (NEB catalog no. E7370), according to the manufacturer's instructions, with the exception of a streptavidin bead capture step prior to indexing PCR. We sequenced the Chicago library on a single lane on the Illumina HiSeq 2500, resulting in 210 million read pairs.

The contig assembly was made with MERACULOUS (Chapman et al. 2011) and scaffolded using the Chicago library with Dovetail Genomics' HiRise scaffolder as previously described by Putnam et al. (2016).

Annotation

We made gene predictions using AUGUSTUS version 3.0.3 (Stanke et al. 2006). We provided as extrinsic evidence to AUGUSTUS RNA-seq alignments made using TopHat2 version 2.0.14 (Kim et al. 2013), repetitive region predictions made using RepeatScout (Price et al. 2005) and RepeatMasker Open-4.0 (Smit et al. 2015), and alignments of published chicken protein sequences made using Exonerate version 2.2.0 (Slater and Birney 2005). We assigned names to these predicted proteins and genes using reciprocal best hits BLAST searches between the set of predicted protein sequences and published protein sequences from related organisms. We also assigned GO terms to our predicted proteins using InterProScan (Jones et al. 2014).

To annotate the genome for microRNAs, we extracted and purified small RNAs from testis tissue of a reproductively-mature alligator caught in the Rockefeller Wildlife Refuge (Grand Chenier, LA) using TRIzol reagent followed by an ethanol precipitation. We sequenced the resulting library on a MiSeq and then, after filtering, used the miRDeep2 pipeline (Friedländer et al. 2012) and MapMi (Guerra-Assunção and Enright 2010) to align these sequences to and predict miRNAs in the alligator genome.

For more detail on our annotation process, see the Supplemental Methods.

Synteny

We created synteny maps and calculated synteny statistics using SyMAP 4.2 (Soderlund et al. 2011), considering only scaffolds of at least 100 kb and ordering the alligator scaffolds based on the chicken genome. We determined synteny for Galgal4 against both the previous version of the alligator genome (Green et al. 2014) and the updated alligator genome for comparison.

To calculate conservation of ordered gene *n*-lets between the alligator and chicken genomes, as well as the human and mouse genomes, we first found homologs in the second genome for genes in the first genome by performing a blastp search of the protein sequence of the primary isoform of each gene in the first genome against a database of all protein sequences in the second genome. We consider *n*-lets only of directly adjacent genes on the same scaffold. We then counted the number of ordered gene *n*-lets in the first genome whose homologs also appear contiguously in the same order in the second genome.

Comparative assembly

We used synteny blocks to separate large structural variants from small polymorphisms, taking a hierarchical approach, with multiple sets of synteny blocks, each defined at a different resolution, from the coarsest, karyotype level all the way down to the fine-grained base level. To create the hierarchy, we used the principles developed by Sibelia tool (Minkin et al. 2013), which can create such a hierarchy for bacterial genomes, but adapted to use a multi-size A-Bruijn graph algorithm for constructing synteny blocks from a multiple genome alignment file in HAL format (Hickey et al. 2013), produced by Progressive Cactus (Paten et al. 2011). At each level of resolution, we used Ragout (Kolmogorov et al. 2014) to decompose the input genomes into synteny blocks and join scaffolds based on this synteny.

We assessed the accuracy of joins by designing primer pairs bracketing the gaps using Primer3 (Untergasser et al. 2012). We PCR amplified saltwater crocodile or gharial DNA with these primers at annealing temperatures ranging from 58°C–62°C for 20 cycles. The joins, primers, and full results are in Supplemental Table S1.

Transposable elements

We identified transposable elements and low complexity repetitive sequences in the alligator (*A. mississippiensis*) genome using RepeatMasker Open-4.0 (Smit et al. 2015) and homology based searches with all known alligator repeats (RepBase Update v21.02). We created a repeat accumulation profile by calculating the Kimura 2-parameter (Kimura 1980) genetic distance between individual insertions and the homologous repeat in the *A. mississippiensis* library.

Egg harvesting, incubation, and dissection

All field and laboratory work were conducted under permits from the Florida Fish and Wildlife Conservation Commission and US Fish and Wildlife Service (permit no. SPGS-1 0-44). Five clutches of alligator eggs were collected from the Lake Woodruff National Wildlife Refuge, where relatively low chemical contamination of persistent organic pollutants allow American alligators to exhibit healthy reproductive activity. One egg from each clutch was dissected to identify the developmental stage of the embryo based on criteria described by Ferguson (1985). Eggs were incubated at 30°C (FPT) until they reached stage 19 based on an equation predicting their development (Kohno and Guillette 2013). At the predicted stage 19, which was before the TSP (stage 21–24) for alligator

TSD (Lang and Andrews 1994), the incubation temperature was either kept constant at FPT or increased to 33°C (MPT). The alligator embryos were dissected and the GAM complex was isolated and preserved in ice-cold RNA*later* (Ambion/Thermo Fisher Scientific) at 3 or 30 d after the stage 19. Gonadal tissues were carefully isolated from GAM under a dissection microscope after RNA stabilization in RNA*later* and stored at −80°C until RNA isolation.

RNA sequencing, expression quantification, and differential expression analysis

Total RNAs were then extracted from the GAM samples using TRIreagent LS (Sigma). Poly(A)$^+$ RNA sequencing libraries were made from each sample using the TruSeq RNA library preparation kit v1 (Illumina). A total of 60 libraries were created by PCR amplification with Illumina barcoding primers at 17 reaction cycles and quantified using a Bioanalyzer DNA 1000 kit (Agilent). Libraries were then pooled and sequenced on a HiSeq 2000 Sequencing system (Illumina).

We removed adapters from the reads using SeqPrep (https://github.com/jstjohn/SeqPrep) with default parameters and aligned them to the alligator genome using TopHat2 (Kim et al. 2013) with default parameters. We used Cuffdiff 2 to calculate normalized expression values, fold-changes, and FDR-adjusted *P*-values for each gene in each tissue at each time point (Trapnell et al. 2013). Cuffdiff 2 reports expression values normalized by transcript length and library size in FPKM for reporting the expression of individual genes in each library in values that are comparable between different genes. Expression values in FPKM are useful for generating heatmaps and reporting average expression values for a gene, but Cuffdiff 2 uses raw counts rather than FPKM for differential expression analysis. For each gonad sample at the 30 d, we compared FPKMs of two genes, *CYP19A1* and *AMH*, to verify the sex of the embryo as in previous studies (Kohno et al. 2015; McCoy et al. 2015), resulting in one sample, the embryo from clutch 13 incubated at MPT, being removed from further analysis. We used an FDR-adjusted *P*-value reported by Cuffdiff 2 for each gene for the null hypothesis that expression levels of that gene in tissues incubated at MPT and FPT are drawn from the same distribution. We considered a gene to be sex-biased if its FDR-adjusted *P*-value was ≤ 0.05.

We used FUNC to perform GO enrichment analysis (Prüfer et al. 2007). We ran the hypergeometric variant of FUNC with default options and the October 2016 release of GO tables.

Predicting estrogen-responsive regions of the alligator genome

The DNA-binding domain of estrogen receptor alpha (ESR1) and the zinc fingers of CTCF are identically conserved in protein sequence among human, chicken, and alligator (Supplemental Fig. S3), suggesting that the DNA-binding motifs of these proteins are also conserved among these species. We predicted binding locations for these proteins by searching the alligator genome for sequences matching the human ESR1-binding motif (Lin et al. 2007) and the chicken CTCF-binding motif (Martin et al. 2011) using PoSSuM-search (Beckstette et al. 2006) with *P*-value cutoffs of 4.388×10^{-6} for ESR1 and 1.214×10^{-6} for CTCF. We considered any genomic region between two inward-facing CTCF motifs within 700 kb to be possibly estrogen responsive if it contained one or more ER-binding motifs.

Data access

The sequence data from this study have been submitted to the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra/) under accession number SRP057608 and to the BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession numbers PRJNA322197, PRJNA163131, PRJNA172383, and PRJNA285470. The genome assembly AllMis2 from this study has been submitted to the NCBI Assembly database (https://www.ncbi.nlm.nih.gov/assembly/) under accession number GCA_000281125.4.

Competing interest statement

R.E.G. is a cofounder and paid consultant of Dovetail Genomics LLC. B.L.O. is a paid consultant.

Acknowledgments

We thank the Florida Fish and Wildlife Conservation Commission and the US Fish and Wildlife Service for their assistance in obtaining collection permits; Steven Weber, Darrin Schultz, Stefany Rubio (UC Santa Cruz), and Jenny Korstian (Texas Tech University) for technical assistance; and Beth Shapiro and Angela Brooks (UC Santa Cruz) for discussion regarding the project. This work was supported by the National Institutes of Health under award numbers 5U54HG007990 (National Human Genome Research Institute), GM085121, and GM109146 and by grants from the W.M. Keck Foundation and the Simons Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Barske LA, Capel B. 2008. Blurring the edges in vertebrate sex determination. *Curr Opin Genet Dev* **18:** 499–505.
Beckstette M, Homann R, Giegerich R, Kurtz S. 2006. Fast index based algorithms and software for matching position specific scoring matrices. *BMC Bioinformatics* **7:** 389.
Boyer A, Lapointe E, Zheng X, Cowan RG, Li H, Quirk SM, DeMayo FJ, Richards JS, Boerboom D. 2010. WNT4 is required for normal ovarian follicle development and female fertility. *FASEB J* **24:** 3010–3025.
Bull JJ, Gutzke WH, Crews D. 1988. Sex reversal by estradiol in three reptilian orders. *Gen Comp Endocrinol* **70:** 425–428.
Carroll JS, Meyer CA, Song J, Li W, Geistlinger TR, Eeckhoute J, Brodsky AS, Keeton EK, Fertuck KC, Hall GF, et al. 2006. Genome-wide analysis of estrogen receptor binding sites. *Nat Genet* **38:** 1289–1297.
Chan CS, Song JS. 2008. CCCTC-binding factor confines the distal action of estrogen receptor. *Cancer Res* **68:** 9041–9049.
Chapman JA, Ho I, Sunkara S, Luo S, Schroth GP, Rokhsar DS. 2011. Meraculous: de novo genome assembly with short paired-end reads. *PLoS One* **6:** e23501.
Crews D, Wibbels T, Gutzke WH. 1989. Action of sex steroid hormones on temperature-induced sex determination in the snapping turtle *(Chelydra serpentina)*. *Gen Comp Endocrinol* **76:** 159–166.
Crottini A, Madsen O, Poux C, Strauss A, Vieites DR, Vences M. 2012. Vertebrate time-tree elucidates the biogeographic pattern of a major biotic change around the K-T boundary in Madagascar. *Proc Natl Acad Sci* **109:** 5358–5363.
Dahlman-Wright K, Cavailles V, Fuqua SA, Jordan VC, Katzenellenbogen JA, Korach KS, Maggi A, Muramatsu M, Parker MG, Gustafsson JA. 2006. International Union of Pharmacology. LXIV. Estrogen receptors. *Pharmacol Rev* **58:** 773–781.
De Santa Barbara P, Bonneaud N, Boizet B, Desclozeaux M, Moniot B, Sudbeck P, Scherer G, Poulat F, Berta P. 1998. Direct interaction of SRY-related protein SOX9 and steroidogenic factor 1 regulates transcription of the human anti-Müllerian hormone gene. *Mol Cell Biol* **18:** 6653–6665.
Ferguson MWJ. 1985. Development. In *Biology of the reptilia* (ed. Gans C, et al.), Vol. 14, pp. 329–491. John Wiley and Sons, New York.
Ferguson MW, Joanen T. 1982. Temperature of egg incubation determines sex in Alligator mississippiensis. *Nature* **296:** 850–853.
Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N. 2012. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res* **40:** 37–52.

Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH, et al. 2009. An oestrogen-receptor-α-bound human chromatin interactome. *Nature* **462:** 58–64.
Gabriel WN, Blumberg B, Sutton S, Place AR, Lance VA. 2001. Alligator aromatase cDNA sequence and its expression in embryos at male and female incubation temperatures. *J Exp Zool* **290:** 439–448.
Green RE, Braun EL, Armstrong J, Earl D, Nguyen N, Hickey G, Vandewege MW, St John JA, Capella-Gutiérrez S, Castoe TA, et al. 2014. Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. *Science* **346:** 1254449.
Gruber CJ, Gruber DM, Gruber IM, Wieser F, Huber JC. 2004. Anatomy of the estrogen response element. *Trends Endocrinol Metab* **15:** 73–78.
Guerra-Assunção JA, Enright AJ. 2010. MapMi: automated mapping of microRNA loci. *BMC Bioinformatics* **11:** 133.
Hickey G, Paten B, Earl D, Zerbino D, Haussler D. 2013. HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* **29:** 1341–1342.
Holleley CE, O'Meally D, Sarre SD, Marshall Graves JA, Ezaz T, Matsubara K, Azad B, Zhang X, Georges A. 2015. Sex reversal triggers the rapid transition from genetic to temperature-dependent sex. *Nature* **523:** 79–82.
Hsieh M, Johnson MA, Greenberg NM, Richards JS. 2002. Regulated expression of Wnts and Frizzleds at specific stages of follicular development in the rodent ovary. *Endocrinology* **143:** 898–908.
Jeays-Ward K, Hoyle C, Brennan J, Dandonneau M, Alldus G, Capel B, Swain A. 2003. Endothelial and steroidogenic cell migration are regulated by WNT4 in the developing mammalian gonad. *Development* **130:** 3663–3670.
Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30:** 1236–1240.
Kawagoshi T, Uno Y, Nishida C, Matsuda Y. 2014. The *Staurotypus* turtles and aves share the same origin of sex chromosomes but evolved different types of heterogametic sex determination. *PLoS One* **9:** e105315.
Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14:** R36.
Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* **16:** 111–120.
Kohno S, Guillette L Jr. 2013. Endocrine disruption and reptiles: using the unique attributes of temperature-dependent sex determination to assess impacts. In *Endocrine disrupters: hazard testing and assessment methods, (ed*. Mattheissen P), pp. 245–271. John Wiley and Sons, New York.
Kohno S, Katsu Y, Urushitani H, Ohta Y, Iguchi T, Guillette LJ Jr. 2010. Potential contributions of heat shock proteins to temperature-dependent sex determination in the American alligator. *Sex Dev* **4:** 73–87.
Kohno S, Bernhard MC, Katsu Y, Zhu J, Bryan TA, Doheny BM, Iguchi T, Guillette LJ. 2015. Estrogen receptor 1 (ESR1; ERα), not ESR2 (ERβ), modulates estrogen-induced sex reversal in the American alligator, a species with temperature-dependent sex determination. *Endocrinology* **156:** 1887–1899.
Kolmogorov M, Raney B, Paten B, Pham S. 2014. Ragout: a reference-assisted assembly tool for bacterial genomes. *Bioinformatics* **30:** i302–i309.
Laganière J, Deblois G, Lefebvre C, Bataille AR, Robert F, Giguère V. 2005. From the Cover: location analysis of estrogen receptor α target promoters reveals that *FOXA1* defines a domain of the estrogen response. *Proc Natl Acad Sci* **102:** 11651–11656.
Lance VA. 2009. Is regulation of aromatase expression in reptiles the key to understanding temperature-dependent sex determination? *J Exp Zool A Ecol Genet Physiol* **311:** 314–322.
Lang JW, Andrews HV. 1994. Temperature-dependent sex determination in crocodilians. *J Exp Zool A Ecol Genet Physiol* **270:** 28–44.
Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v1 [q-bio.GN].
Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, Zhang J, et al. 2012. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148:** 84–98.
Lin CY, Vega VB, Thomsen JS, Zhang T, Kong SL, Xie M, Chiu KP, Lipovich L, Barnett DH, Stossi F, et al. 2007. Whole-genome cartography of estrogen receptor α binding sites. *PLoS Genet* **3:** e87.
Martin D, Pantoja C, Fernández Miñán A, Valdes-Quezada C, Moltó E, Matesanz F, Bogdanović O, de la Calle-Mustienes E, Domínguez O, Taher L, et al. 2011. Genome-wide CTCF distribution in vertebrates defines equivalent sites that aid the identification of disease-associated genes. *Nat Struct Mol Biol* **18:** 708–714.
McCoy JA, Parrott BB, Rainwater TR, Wilkinson PM, Guillette LJ Jr. 2015. Incubation history prior to the canonical thermosensitive period determines sex in the American alligator. *Reproduction* **150:** 279–287.
McCoy JA, Hamlin HJ, Thayer L, Guillette LJ, Parrott BB. 2016. The influence of thermal signals during embryonic development on intrasexual and sexually dimorphic gene expression and circulating steroid hormones in American alligator hatchlings (*Alligator mississippiensis*). *Gen Comp Endocrinol* **238:** 47–54.
Milnes MR, Bryan TA, Medina JG, Gunderson MP, Guillette LJ. 2005. Developmental alterations as a result of in ovo exposure to the pesticide metabolite p,p′-DDE in Alligator mississippiensis. *Gen Comp Endocrinol* **144:** 257–263.
Minkin I, Patel A, Kolmogorov M, Vyahhi N, Pham S. 2013. Sibelia: a scalable and comprehensive synteny block generation tool for closely related microbial genomes. In *Algorithms in bioinformatics* (ed. Darling A, Stoye J), pp. 215–229. Springer-Verlag, Berlin.
Morrish BC, Sinclair AH. 2002. Vertebrate sex determination: many means to an end. *Reproduction* **124:** 447–457.
Nakabayashi O, Kikuchi H, Kikuchi T, Mizuno S. 1998. Differential expression of genes for aromatase and estrogen receptor during the gonadal development in chicken embryos. *J Mol Endocrinol* **20:** 193–202.
Nilsson S, Mäkelä S, Treuter E, Tujague M, Thomsen J, Andersson G, Enmark E, Pettersson K, Warner M, Gustafsson JA. 2001. Mechanisms of estrogen action. *Physiol Rev* **81:** 1535–1565.
Parrott BB, Kohno S, Cloy-McCoy JA, Guillette LJ. 2014. Differential incubation temperatures result in dimorphic DNA methylation patterning of the SOX9 and aromatase promoters in gonads of alligator (*Alligator mississippiensis*) embryos. *Biol Reprod* **90:** 2.
Paten B, Earl D, Nguyen N, Diekhans M, Zerbino D, Haussler D. 2011. Cactus: algorithms for genome multiple sequence alignment. *Genome Res* **21:** 1512–1528.
Price AL, Jones NC, Pevzner PA. 2005. De novo identification of repeat families in large genomes. *Bioinformatics* **21**(Suppl 1): i351–i358.
Prüfer K, Muetzel B, Do H, Weiss G, Khaitovich P, Rahm E, Pääbo S, Lachmann M, Enard W. 2007. FUNC: a package for detecting significant associations between gene sets and ontological associations. *BMC Bioinformatics* **8:** 41.
Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet CW, et al. 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res* **26:** 342–350.
Ramsey M, Crews D. 2009. Steroid signaling and temperature-dependent sex determination: reviewing the evidence for early action of estrogen during ovarian determination in turtles. *Semin Cell Dev Biol* **20:** 283–292.
Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, et al. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159:** 1665–1680.
Sanborn AL, Rao SS, Huang SC, Durand NC, Huntley MH, Jewett AI, Bochkov ID, Chinnappan D, Cutkosky A, Li J, et al. 2015. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci* **112:** E6456–E6465.
Schroeder AL, Metzger KJ, Miller A, Rhen T. 2016. A novel candidate gene for temperature-dependent sex determination in the common snapping turtle. *Genetics* **203:** 557–571.
Shedlock AM, Botka CW, Zhao S, Shetty J, Zhang T, Liu JS, Deschavanne PJ, Edwards SV. 2007. Phylogenomics of nonavian reptiles and the structure of the ancestral amniote genome. *Proc Natl Acad Sci* **104:** 2767–2772.
Shoemaker-Daly CM, Jackson K, Yatsu R, Matsumoto Y, Crews D. 2010. Genetic network underlying temperature-dependent sex determination is endogenously regulated by temperature in isolated cultured *Trachemys scripta* gonads. *Dev Dyn* **239:** 1061–1075.
Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6:** 31.
Smit A, Hubley R, Green P. 2015. *RepeatMasker Open-4.0*. http://www.repeatmasker.org/.
Smith CA, Elf PK, Lang JW, Joss JMP. 1995. Aromatase enzyme activity during gonadal sex differentiation in alligator embryos. *Differentiation* **58:** 281–290.
Soderlund C, Bomhoff M, Nelson WM. 2011. SyMAP v3.4: a turnkey synteny system with application to plant genomes. *Nucleic Acids Res* **39:** e68.
Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* **34:** W435–W439.
Suh A, Churakov G, Ramakodi MP, Platt RN, Jurka J, Kojima KK, Caballero J, Smit AF, Vliet KA, Hoffmann FG, et al. 2015. Multiple lineages of ancient CR1 retroposons shaped the early genome evolution of amniotes. *Genome Biol Evol* **7:** 205–217.
Toda K, Shizuta Y. 1993. Molecular cloning of a cDNA showing alternative splicing of the 5′-untranslated sequence of mRNA for human aromatase P-450. *Eur J Biochem* **213:** 383–389.
Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. 2013. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* **31:** 46–53.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. 2012. Primer3: new capabilities and interfaces. *Nucleic Acids Res* **40:** e115.

Welboren WJ, van Driel MA, Janssen-Megens EM, van Heeringen SJ, Sweep FC, Span PN, Stunnenberg HG. 2009. ChIP-Seq of ERα and RNA polymerase II defines genes differentially responding to ligands. *EMBO J* **28:** 1418–1428.

Yao HH, Matzuk MM, Jorgez CJ, Menke DB, Page DC, Swain A, Capel B. 2004. Follistatin operates downstream of Wnt4 in mammalian ovary organogenesis. *Dev Dyn* **230:** 210–215.

Yatsu R, Miyagawa S, Kohno S, Saito S, Lowers RH, Ogino Y, Fukuta N, Katsu Y, Ohta Y, Tominaga M, et al. 2015. TRPV4 associates environmental temperature and sex determination in the American alligator. *Sci Rep* **5:** 18581.

Ying SY, Becker A, Swanson G, Tan P, Ling N, Esch F, Ueno N, Shimasaki S, Guillemin R. 1987. Follistatin specifically inhibits pituitary follicle stimulating hormone release in vitro. *Biochem Biophys Res Commun* **149:** 133–139.

Zhang Y, Liang J, Li Y, Xuan C, Wang F, Wang D, Shi L, Zhang D, Shang Y. 2010. CCCTC-binding factor acts upstream of FOXA1 and demarcates the genomic response to estrogen. *J Biol Chem* **285:** 28604–28613.

Zhou Q, Zhang J, Bachtrog D, An N, Huang Q, Jarvis ED, Gilbert MT, Zhang G. 2014. Complex evolutionary trajectories of sex chromosomes across bird taxa. *Science* **346:** 1246338.

Received August 1, 2016; accepted in revised form December 13, 2016.

Improved genome assembly of American alligator genome reveals conserved architecture of estrogen signaling

Edward S. Rice, Satomi Kohno, John St. John, et al.

Genome Res. 2017 27: 686-696 originally published online January 30, 2017
Access the most recent version at doi:10.1101/gr.213595.116

Supplemental Results

Small RNAs

In model vertebrate species (ex. human, mouse, chicken, leopard frog, zebrafish) a few hundred to more than a thousand miRNAs have been identified (Kozomara and Griffiths-Jones 2014). In non-model taxa, which include the crocodilians, miRNAs are frequently identified based on sequence conservation to known miRNAs. Using this technique some conserved miRNAs in Alligator mississippiensis have been annotated by mapping small RNA reads to miRNAs from the chicken and green anole (Lyson et al. 2012) but no lineage-specific miRNAs are identifiable. Results presented here represent a first step in understanding the lineage-specific evolution of miRNAs in the crocodilians.

A total of 15 million reads from the testis library was reduced to 1.12 million unique, quality- and size-filtered reads used for miRNA prediction with miRDeep2. miRDeep2 mapped reads to 114 chicken miRNAs, confirming their presence and expression in alligator testis. Initial predictions of novel miRNAs (n = 145) were filtered using various criteria. Putative miRNAs with less than 10 reads mapping to the predicted mature miRNA (n = 15), a miRDeep score < 1 (n = 13), non-significant randFold scores (n = 11), more reads mapping to the hairpin loop than the miRNA* strand (n = 7), homology to ribosomal or transfer RNAs (n = 2), or overlapping loci (n = 2) were removed from downstream analyses. The remaining putative miRNAs were re-predicted in the alligator genome and compared to the crocodile, gharial, and chicken genomes to identify homologous miRNAs using MapMi. MapMi removed 31 putative miRNAs with homology to TEs and one putative miRNA with a low complexity sequence. Three miRDeep miRNAs failed re-prediction in MapMi, though two were identified in either the crocodile or gharial. In all, 60 putative miRNAs passed all quality filters and were predicted by both the miRDeep2 and MapMi algorithms, 25 were present in all crocodilians, 17 were alligator specific, and 11 were in the crocodilians and the chicken. Seven were present either the alligator and the gharial or the alligator and the crocodile, but not all three crocodilians. Blast results against NCBI's non-redundant nucleotide

database identified four putative miRNAs with homologs in *Anolis carolinensis* and one with *Danio rerio*. Four of the 5 miRNAs with NCBI homologs were found in all four taxa examined with MapMi (aca-mir146-a, aca-mir-34c, dar-mir-144-5, aca-mir-1388). The fifth (aca-mir-425) was in all three crocodilians, but not in the chicken. Due to the deep divergences of these taxa and strong selection on many miRNAs (Quach et al. 2009), it is likely that these putative miRNAs are functional in crocodilians. In addition, the ability to identify these conserved-functional miRNAs demonstrates the ability of the methods employed herein to identify true miRNAs that are lineage-specific. Additional work is necessary to verify and ascribe function to the putative miRNAs. Putative miRNAs were deposited in miRBase and all sequence data used for miRNA prediction was deposited in the NCBI Short read archive (PRJNA285470).

Supplemental Methods

Gene prediction

We made gene predictions using the AUGUSTUS gene prediction software version 3.0.3 (Stanke et al. 2006). AUGUSTUS predicts genes based on a hidden Markov model trained on gene structures from a related species as well as extrinsic evidence provided by the user. We provided RNA-seq alignments, repetitive element predictions, and chicken protein alignments to AUGUSTUS as extrinsic evidence. We aligned previously-published RNA-seq reads from various tissues of *Alligator mississippiensis* (Green et al. 2014) to the genome (SRA: SRP057608) using TopHat 2.0.14 (Kim et al. 2013) with default parameters. We found repetitive elements in the genome using RepeatScout (Price et al. 2005) and RepeatMasker Open-4.0 (Smit et al. 2015) with default parameters. We aligned all *Gallus gallus* (chicken) proteins from UniProt to the genome using Exonerate version 2.2.0 (Slater and Birney 2005) with the protein2genome model. Finally, we ran AUGUSTUS using these sources of extrinsic evidence and parameters trained on gene structures from *G. gallus*.

Functional annotation

We assigned protein names, gene nomenclature, and Gene Ontology (GO) terms to the predicted genes. We chose protein names based on reciprocal best hits BLAST from orthologous proteins from vertebrate species with a gene nomenclature project, specifically *G. gallus* (chicken), *A. carolinensis* (Green anole), *D. rerio* (Zebrafish), and *H. sapiens* (Human). We define orthologous proteins as those with a reciprocal best hit using default blastp parameters and an E-value cutoff of 0.00001. We assigned gene names using the same strategy, resulting in the assignment of 15,977 protein and gene names. We assigned GO terms to predicted proteins based upon a combinatorial approach. We mapped predicted proteins to InterPro identifiers and GO (assigned the GO evidence code of "IEA" or Inferred from Electronic Annotation) based on InterProScan (Jones et al. 2014). We also transferred GO using reciprocal blast from orthologous vertebrate genes experimental evidence codes (assigned the GO evidence code "ISA" or Inferred from Sequence Alignment). We merged GO annotations from these two sources, removed duplicates, and manually reviewed GO terms to eliminate those that are not species-appropriate, such as "sex chromosome" and "fin development." Following this strategy, 17,430 American alligator proteins were assigned 5,960 unique GO terms.

Small RNAs

Testis tissue was harvested from a wild-caught, reproductively mature, male alligator from Rockefeller National Wildlife (Grand Chenier, LA) and a horizontal cross section was homogenized for small RNA isolation. Small RNAs were purified using TRIzol reagent followed by an ethanol precipitation. RNA quantity and quality was measured using a Bioanalyzer, to assure that RNA Integrity Number (RIN) was greater than 7.5. The small RNA pools was prepped for Illumina sequencing using a NEBNext Small RNA Library Prep Set with converted RNA fragments ranging from 15 to 35 nt

(excluding sequencing adapters) selected via PippinHT. The resulting library was sequenced on a single MiSeq lane 1x50 nt.

Adapters and low quality base calls were removed from small RNA sequences using the FASTX-Toolkit (v0.0.13; http://hannonlab.cshl.edu/fastx_toolkit/index.html). Specifically, reads with scores below Q20 across 50% or more of the read, after adapter trimming, were discarded. Once filtered, reads falling outside of an 18-24 nt range were culled. miRNAs were predicted from the remaining reads using the miRDeep2 pipeline (Friedländer et al. 2012). All high quality small RNA reads were mapped to known chicken (*Gallus gallus*) miRNAs (mature and hairpin) and the new alligator genome using miRDeep2's mapper.pl. Additional parameters included collapsing unique reads (-m) and limiting the maximum mapping locations to five or fewer (-r 5). Once mapped, miRNAs were predicted from reads without homology to known chicken miRNAs using the miRDeep2.pl script.

Several filters were applied to novel miRNAs predicted by miRDeep2. Any novel miRNAs that were similar to ribosomal or transfer RNAs, had fewer than 10 reads from the mature miRNA, had a miRDeep score less than 1, did not have a significant randFold score, overlapped with other predicted or known miRNAs, or contained more reads mapping to the miRNA hairpin loop than the miRNA* were removed from further analyses. Known chicken miRNAs were accepted regardless of these constraints. MapMi (Guerra-Assunção and Enright 2010) was used to identify homologous loci to the putative miRNAs predicted by miRDeep2 in the crocodile (*Crocodylus porosus*; JRXG00000000.1), gharial (*Gavialis gangeticus*; JRWT00000000.1), and chicken (CM000000.4) genomes. Initial steps in the MapMi uses Dust3 to remove low complexity sequences and then culls sequences with homology to TEs. MapMi predictions scoring less than 35 were considered low quality and removed. In addition, miRDeep2 putative miRNAs not re-predicted by MapMi in the alligator genome were removed as well.

Supplemental Figures

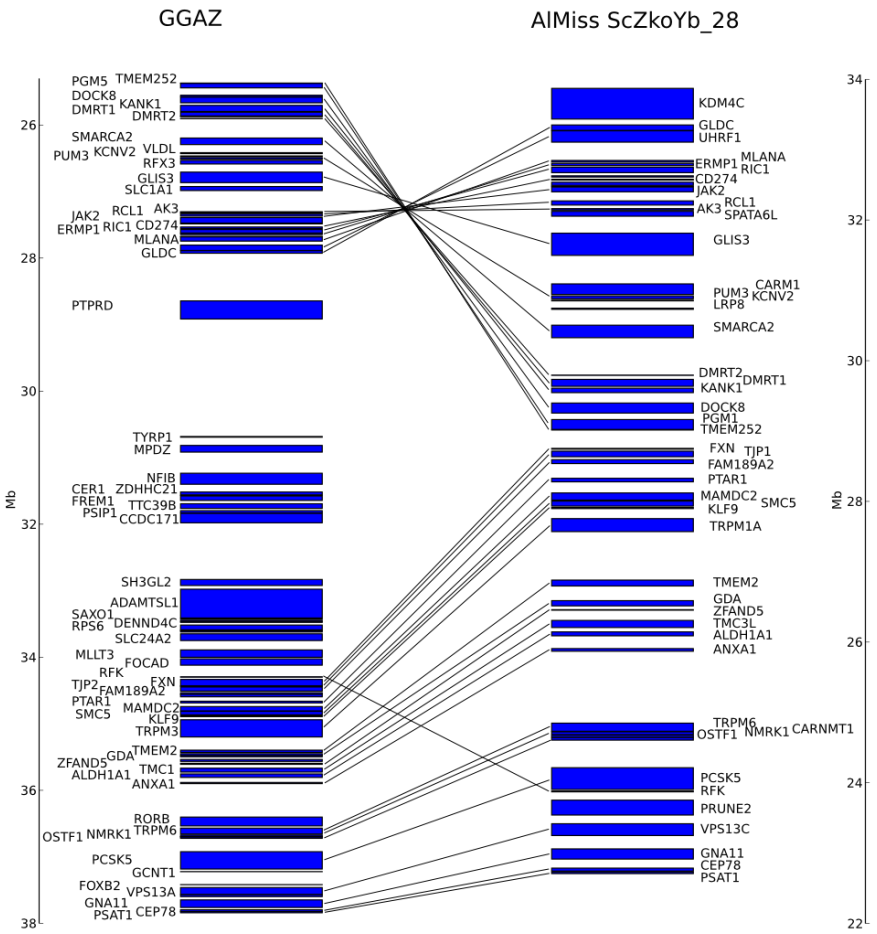

Figure S1. Synteny between the chicken Z chromosome and scaffold 28 of the alligator assembly, around the avian sex-determination gene DMRT1. Orthologous genes are connected with lines.

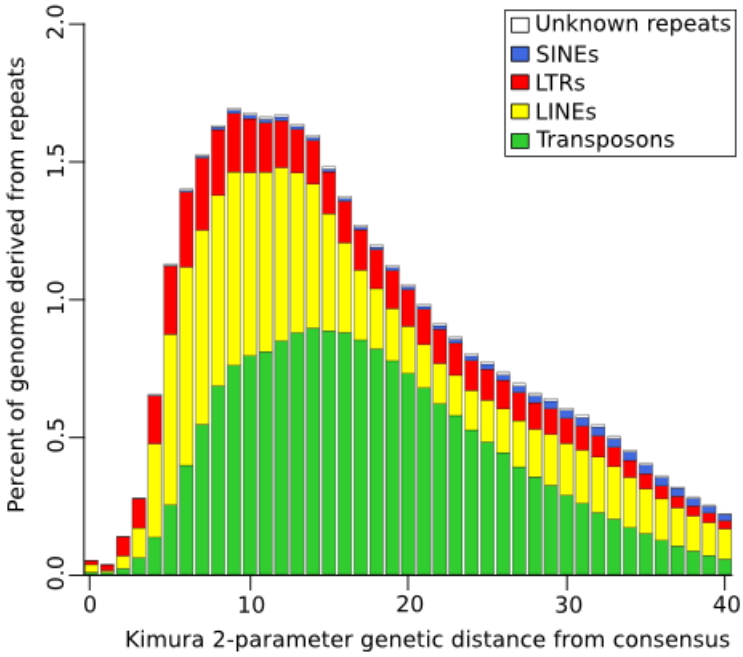

Figure S2. The Kimura 2-parameter (Kimura 1980) between individual transposable element insertions and their respective consensus sequences as a percentage of the genome. Genetic distance increases with element insertion age.

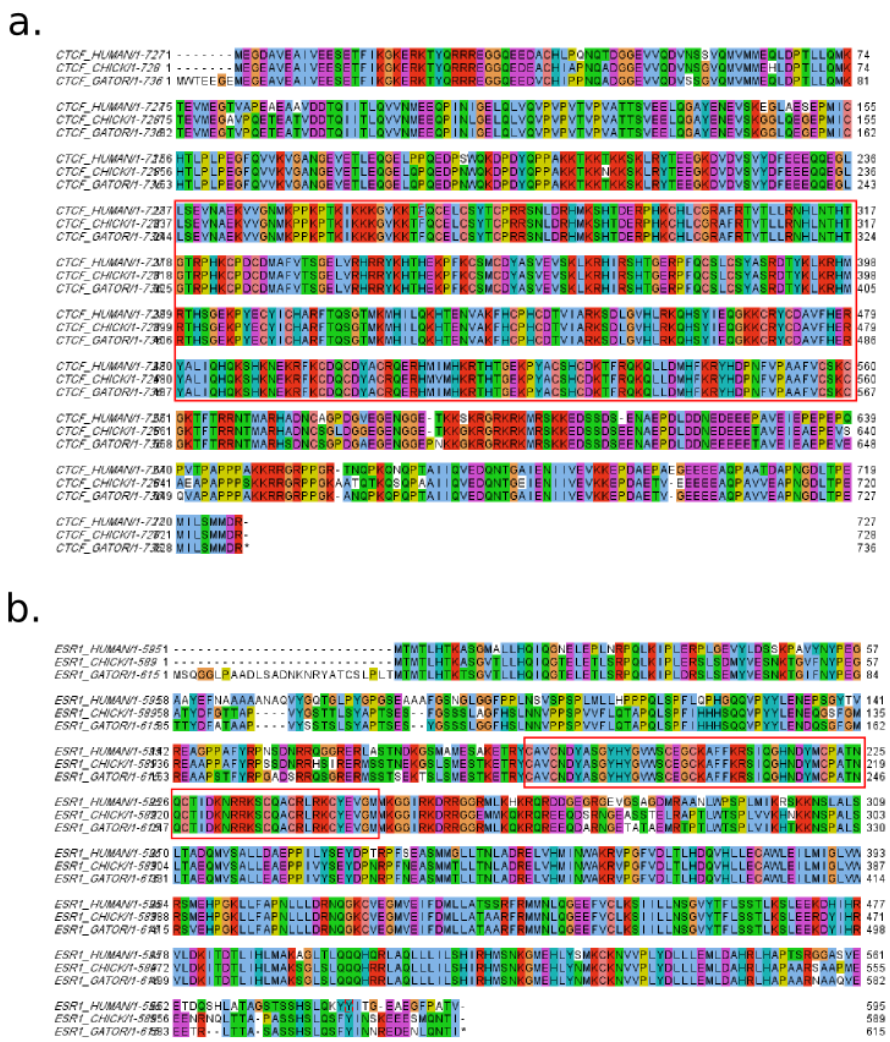

Figure S3. Alignments of the protein sequences of human, chicken, and alligator orthologs of CTCF (a) and ESR1 (b). The DNA-binding domains of each are highlighted in a red box, showing perfect conservation.

Supplemental Tables

Table S1. Scaffold joins in the saltwater crocodile and gharial genomes verified by PCR, including the primers used and results.

Table S2. Total repetitive content in new alligator assembly and percent of genome derived from all repeats as well as the three dominant TE superfamilies in crocodilians. Repeats were identified using RepeatMasker (Smit et al. 2015) and known alligator repeats present in RepBase (v21.02).

Table S3. Embryonic alligator GAM complex libraries for RNA-sequencing, along with their NCBI accessions.

Table S4. Genes determined to have sex-biased expression in alligator embryos, including expression values in FPKM, fold changes, and FDR-adjusted p-values.

Table S5. Enriched gene ontology terms for genes with male- and female-biased expression in the gonads at the 30-day time point.

Table S6. ESR1 DNA-binding domain conservation, showing perfect protein sequence conservation of the binding domain in human, mouse, chicken, alligator, and turtle orthologs of this protein.

References

Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N. 2012. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res* **40**: 37-52.

Green RE, Braun EL, Armstrong J, Earl D, Nguyen N, Hickey G, Vandewege MW, St John JA, Capella-Gutiérrez S, Castoe TA et al. 2014. Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. *Science* **346**: 1254449.

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**: 1236-1240.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**: R36.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* **16**: 111-120.

Kozomara A, Griffiths-Jones S. 2014. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res* **42**: D68-73.

Lyson TR, Sperling EA, Heimberg AM, Gauthier JA, King BL, Peterson KJ. 2012. MicroRNAs support a turtle + lizard clade. *Biol Lett* **8**: 104-107.

Price AL, Jones NC, Pevzner PA. 2005. De novo identification of repeat families in large genomes. *Bioinformatics* **21 Suppl 1**: i351-358.

Quach H, Barreiro LB, Laval G, Zidane N, Patin E, Kidd KK, Kidd JR, Bouchier C, Veuille M, Antoniewski C et al. 2009. Signatures of purifying and local positive selection in human miRNAs. *Am J Hum Genet* **84**: 316-327.

Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**: 31.

Smit A, Hubley R, Green P. 2015. RepeatMasker Open-4.0.

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* **34**: W435-439.

Chapter 4

NOVA1 and RNA splicing in the Neanderthal brain

In this chapter, I discuss my work on a project involving the splicing factor NOVA1 that is different in Neanderthals and modern humans. NOVA1 is a trans-regulatory factor that controls the splicing of many genes during brain development. A high-quality reference genome for humans already exists, and has been used as a reference for the alignment of ancient DNA reads from Neanderthal remains, which are too degraded to assemble. This has in turn been used to find places in the genome where Neanderthals were different from modern humans, such as an RNA-binding domain of NOVA1. Using RNA-seq transcriptomic data from a cell line edited to have the Neanderthal version of this gene, as well as a wild-type modern human control cell line, I was able to find genes that are spliced differently by these two versions of NOVA1.

The writing and figures herein are my own unless otherwise noted below, and focus

on my contributions to the project, but the extensive work of several collaborators was necessary for me to perform the analyses I describe. Alysson R. Muotri and Richard E. Green conceived and led this project. Cleber A. Trujillo performed the CRISPR editing, tissue culture, and related assays, and created **Figure 4.2**. Ashley Byrne performed the RNA-seq library preparation and wrote the methods for this step, Maximilian Marin assisted with splicing quantification, Lars Fehren-Schmidt performed Sanger sequencing, and Angela N. Brooks provided valuable guidance and input into the project. I thank all of these collaborators and look forward to coauthoring a manuscript with them.

4.1 Introduction

Neanderthals were a group of archaic humans who coexisted with modern humans and went extinct around 40 thousand years ago (kya) (Higham et al., 2014). Genetic evidence suggests that they interbred with modern humans, leaving most humans today outside of sub-Saharan Africa with some amount of Neanderthal ancestry, ranging from about 1% to 4% (Green et al., 2010).

Questions about the mental capacity of Neanderthals compared to that of modern humans are the subject of much speculation, but few data currently exist that could help answer these questions. While they are known to have hunted (Gardeisen, 1999; Richards et al., 2000) and used technologies such as fire (Roebroeks and Villa, 2011), it is unclear whether Neanderthals engaged in other behaviors that are human-specific among currently extant primates, such as art, speech, and ceremonial burial.

Many regions of the genome are depleted for Neanderthal ancestry among current human populations (Sankararaman et al., 2014; Vernot and Akey, 2014). In some locations where Neanderthals and humans differed, there are no longer any modern humans with the Neanderthal variant. These Neanderthal-specific variants, especially the ones that cause coding differences, may be important to understanding phenotypic differences between Neanderthals and modern humans. Some may have been purged from modern human genomes due to purifying selection because they were incompatible with other variants present in modern humans. At genomic loci where the Neanderthal allele persists in modern human populations, the Neanderthal variant can be regulated differently than the modern human variant. In modern humans heterozygous for Neanderthal variants, McCoy et al. (2017) found evidence that about a quarter of genes show differential expression between the human and Neanderthal alleles. No work has been done to date on whether Neanderthal variants are spliced differently, however.

One gene containing a nonsynonymous Neanderthal-specific variant in its coding sequence is the splicing factor NOVA1. NOVA1 is a good target for studying the effects of Neanderthal variants in the human genome because its action in humans, namely, regulating alternative splicing during brain development, is well-understood. NOVA1 is a sequence-specific RNA-binding protein (Buckanovich et al., 1996; Buckanovich and Darnell, 1997) that regulates alternative splicing in neurons (Jensen et al., 2000) and is a master regulator of splicing in genes responsible for synapse formation (Ule et al., 2005). It can promote inclusion or exclusion of cassette exons in a mature mRNA depending on where it binds to the pre-mRNA in relation to splice junctions (Ule et al., 2006). NOVA1 has three K homology (KH) domains that bind to RNA. Structural work on NOVA1 has shown that RNA can bind to the KH1 and KH2 domains

simultaneously, and that NOVA1 can dimerize with contact between the two molecules' KH2 domains, allowing RNA to bind to both KH1 domains (Teplova et al., 2011).

The CRISPR/Cas9 system for genome editing presents an unprecedented opportunity to experimentally test the effects of Neanderthal-specific variants in human genomes. In the remainder of this chapter, I describe how we used RNA from neural organoids containing either the human version of NOVA1, the Neanderthal version, or a non-functional knockout version to determine how splicing during neural development would have been different in humans with the Neanderthal version of NOVA1.

4.2 Results

4.2.1 All modern humans have a private coding variant in the KH2 domain of NOVA1

The KH2 domain of NOVA1 has a different amino acid sequence in humans and Neanderthals, caused by a single base pair change in the coding sequence. These sequences are different in position 200 of the full protein sequence; humans have a valine in this position while Neanderthals had an isoleucine (**Figure 4.1**(a)). Based on the 1000 Genomes Project panel, there is no evidence that the Neanderthal version of NOVA1 is present today in modern human populations (1000 Genomes Project Consortium et al., 2015). Other vertebrates, such as mice and chickens, also have an isoleucine in this position, suggesting that Neanderthals had the basal version of NOVA1 and modern humans have a derived allele that has reached fixation (**Figure 4.1**(b-c)).

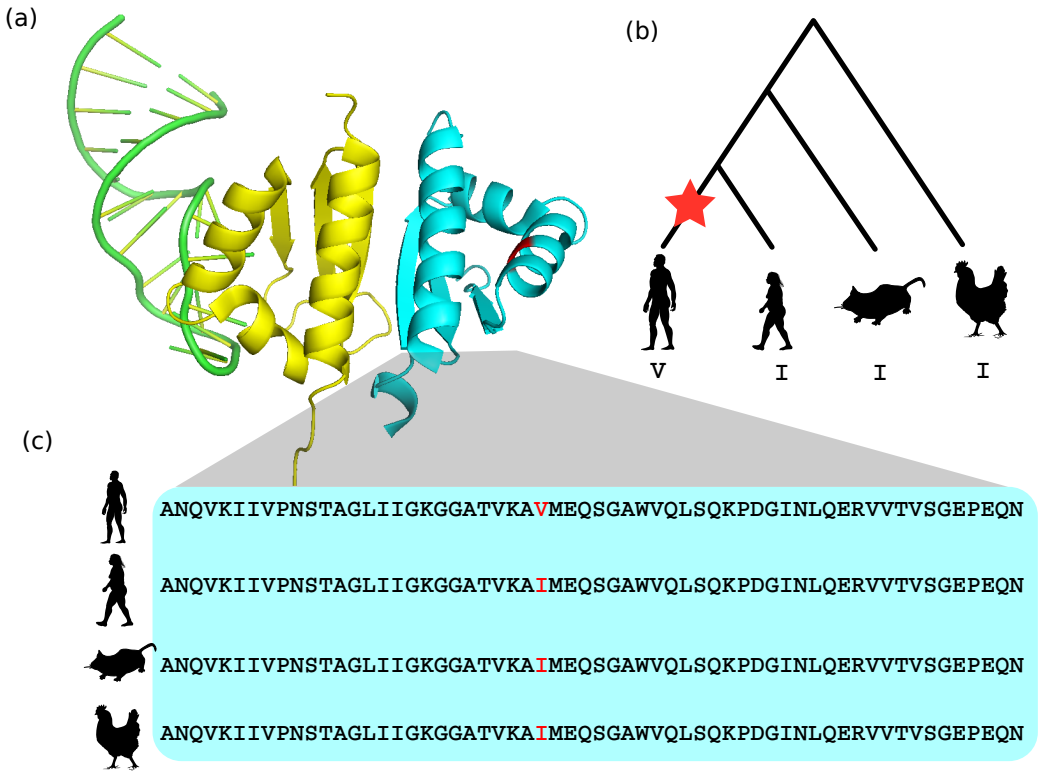

Figure 4.1: Structure and phylogeny of NOVA1. (a) Partial structure of NOVA1, showing the KH1 (yellow) and KH2 (blue) domains bound to RNA (green). The structure of the KH3 domain has not been studied. RNA can simultaneously bind to both the KH1 and KH2 domains of NOVA1. The location in KH2 of the Ile200Val difference between humans and other amniotes is highlighted in red. (b) Phylogeny of modern humans, Neanderthals, mice, and chickens, showing the amino acid at position 200. (c) An alignment of the protein sequences of KH2 for modern human, Neanderthal, mouse, and chicken. The Ile200Val change is highlighted in red.

4.2.2 Neural organoids grown from stem cells with Neanderthal version of NOVA1 are phenotypically distinct

We used CRISPR/Cas9 to edit the genomes of induced pluripotent stem cells (iPSCs) to have either the Neanderthal version of NOVA1 or a nonfunctional knockout version with a premature stop codon. We grew these cells, along with unedited human iPSCs, into neural progenitor cells and then neuronal three-dimensional organoids. **Figure 4.2** shows the develop-

ment of organoids grown from these three cell lines and the qualitative phenotypic differences between them. Organoids from the three different cell lines were also quantitatively different in size and proliferation distance (**Figure 4.2**(b-c)).

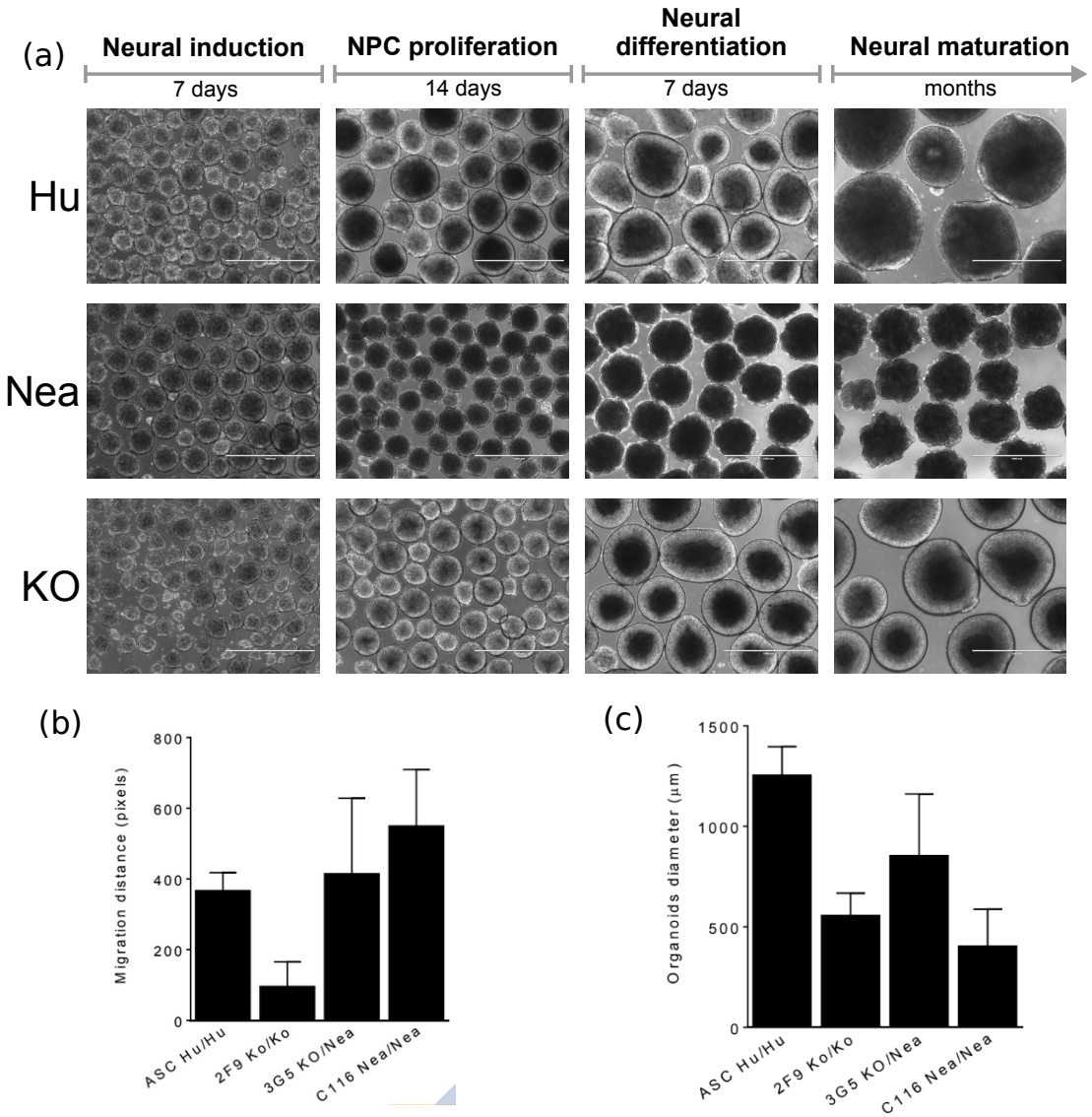

Figure 4.2: (a) Organoids grown from iPSCs homozygous for human (Hu), Neanderthal (Nea), and knockout (KO) versions of NOVA1. We captured images of the organoids at four different developmental stages: neural induction, neural progenitor cell proliferation, neural differentiation, and neural maturation. We extracted RNA after one month, at the end of the differentiation stage, and after two months, during the maturation stage. (b) Organoids containing the human, Neanderthal, and knockout versions of NOVA1 had significantly different migration distances. (c) The organoids from the different cell lines had significantly different sizes.

4.2.3 Changing NOVA1 to the Neanderthal version causes global changes in splicing and gene expression

In order to study how splicing is different between cells with the human versus Neanderthal versions of NOVA1 during neural development, we extracted RNA from organoids at two different developmental stages: after one month of growth, during neural differentiation, and after two months of growth, during neural maturation, with at least two replicates for each cell line and time point combination. We performed RNA-seq on these samples. After verifying that the expected versions of NOVA1 were being expressed, we used the RNA-seq data to quantify gene expression and alternative splicing across the samples.

To quantify splicing, we used juncBASE (Brooks et al., 2011) to calculate a PSI (percent spliced in) value for each alternative splicing event (**Figure 4.3**). To visualize the differences in splicing between the different cell lines and time points, we performed principal components analysis (PCA) on the PSI values for each sample and cassette exon splicing event (**Figure 4.4**a). The replicates from each sample clustered together in the first two principal components. We also compared each of the principal components to NOVA1 expression and found that the second principal component is negatively correlated with NOVA1 expression with $R^2 = 0.65$ (**Figure 4.4**b).

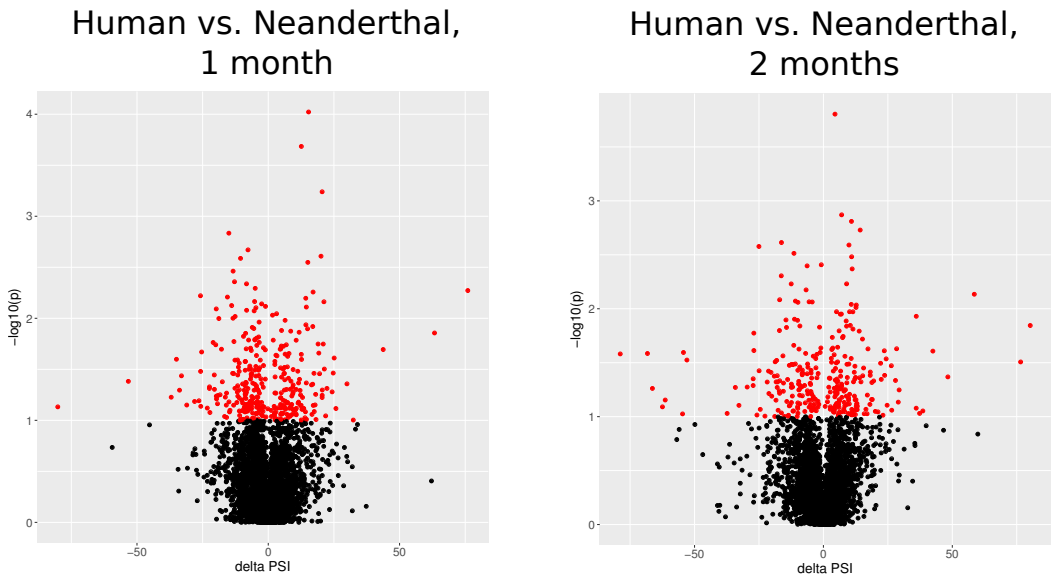

Figure 4.3: Volcano plots showing the difference in percent spliced in between human and Neanderthal versus the p-value that an event is differentially spliced between these two cell lines for each splicing event, at one month (a) and two months (b). Significant events are colored in red.

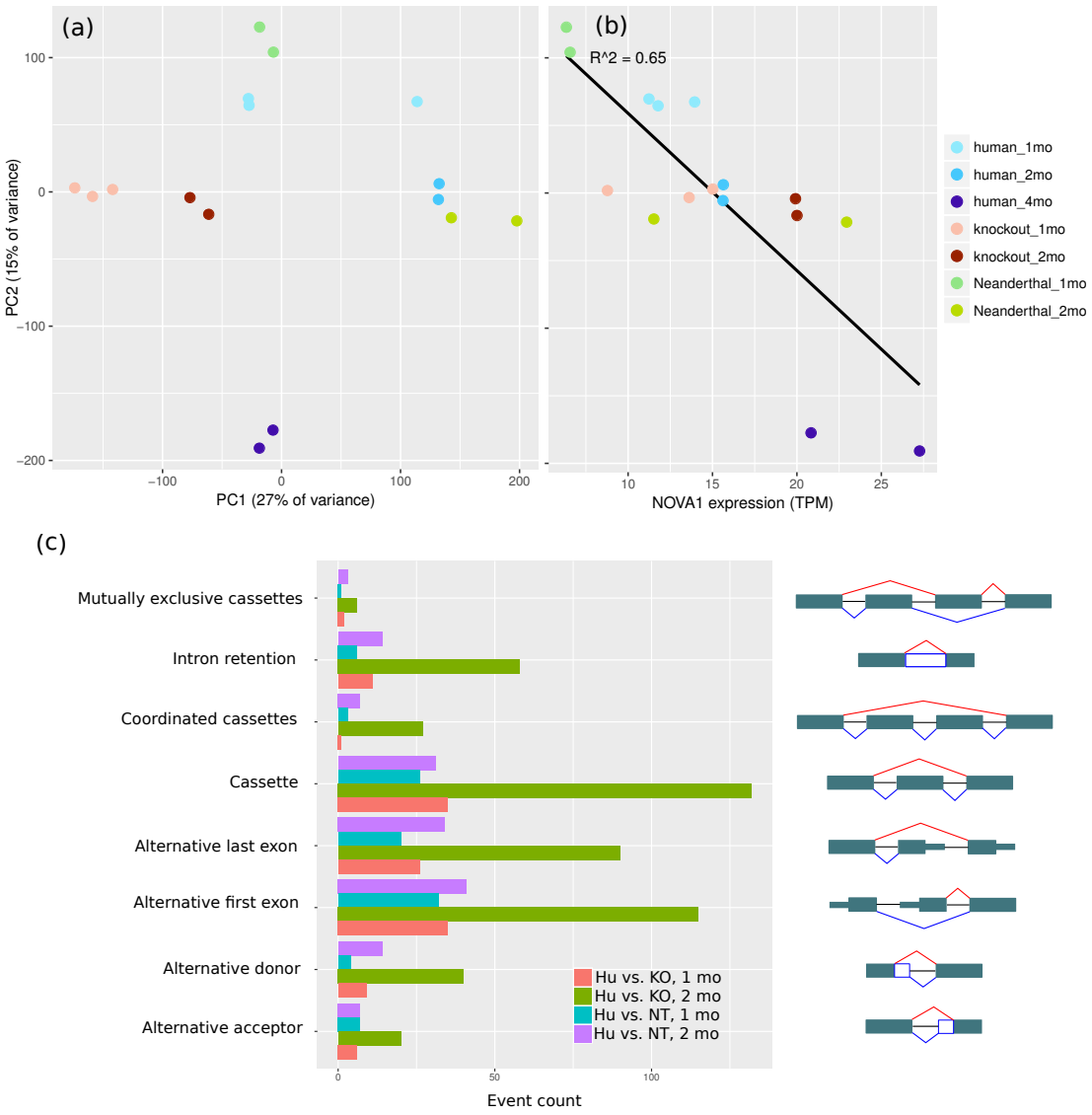

Figure 4.4: Global analysis of splicing among different samples. (a) A plot of the first two principal components from a principal components analysis of cassette inclusion frequency shows that replicates from different cell lines cluster together. (b) The second principal component negatively correlates with NOVA1 expression. (c) Numbers of differential splicing events of different types based on comparisons between human and Neanderthal at one month, human and knockout at one month, human and Neanderthal at two months, and human and knockout at two months. More differential splicing is found between human and knockout than between human and Neanderthal, and at two months than at one month.

We then examined splicing events that occurred at significantly different rates between human and knockout or Neanderthal cell lines at the one or two month time points. At one month, we found 113 alternative splicing (AS) events that occur at significantly different rates between human and Neanderthal cell lines, affecting 122 genes. At two months, we found 166 significantly different splicing events affecting 156 genes. Cassette inclusion and alternative first and last exons were the most common differential AS events for all comparisons, and all classes of AS events were more commonly differentially spliced between human and knockout than human and Neanderthal, and at two months than at one month (**Figure 4.4**c).

Although NOVA1 regulates splicing rather than overall expression intensity of its targets, changes in splicing can have downstream effects on gene expression. To find these events, we quantified gene expression and found genes that are differentially expressed between human and Neanderthal or human and knockout at one or two months. We found 277 differentially expressed genes between human and Neanderthal at one month and 757 at two months (**Figure 4.5**). Similarly to differential splicing, there were more differentially expressed genes between human and knockout than between human and Neanderthal, and more at two months than at one month. This is unsurprising as we expect knocking out a gene to have a larger effect on downstream expression than modifying one amino acid in its protein sequence, and we also expect the downstream effects of a modification to a gene involved in neural development to be amplified as development progresses.

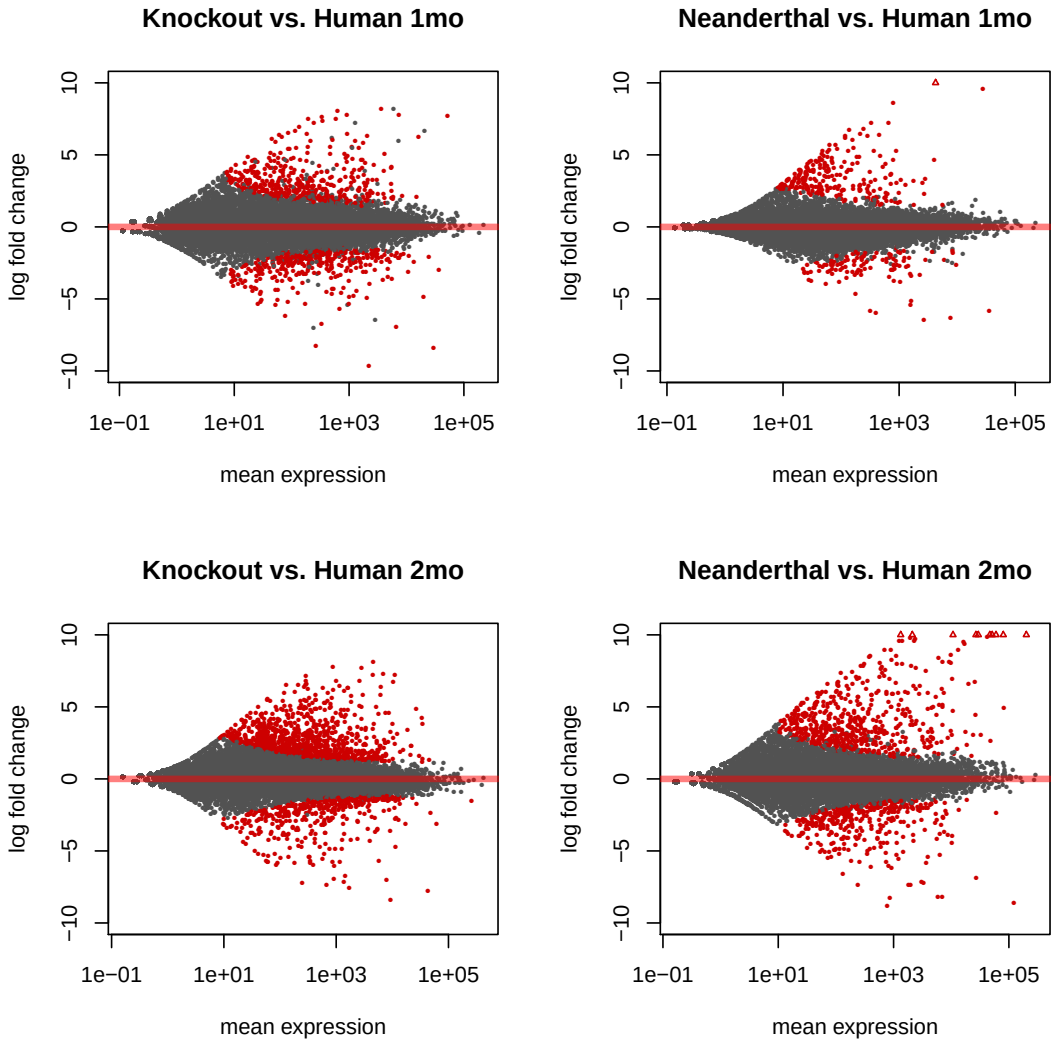

Figure 4.5: MA plots showing overall expression on the X-axis and log-2 fold change on the Y-axis for every gene. Points colored red represent genes that were significantly differentially expressed with an FDR $\alpha = 0.01$.

The genes with the most different expression are listed in **Table 4.1**. Genes with Neanderthal-biased expression at one month are NNAT, which codes for a protein involved in transforming stem cells to neurons through calcium signaling during neural development (Lin

et al., 2010); TDGF1, a membrane signaling protein involved in cell proliferation and migration during development; and EPCAM, a cell adhesion protein (Litvinov et al., 1994). The genes with the most human-biased expression at one month are FEZF1, a protein involved in axon guidance and neural migration; PAX6, a transcription factor that regulates gene expression during embryonic brain development (Davis et al., 2008); and LHX5, a transcription factor controlling cell differentiation during brain development (Heide et al., 2015). The genes with the most Neanderthal-biased expression at two months are MYL1, which forms part of the myosin muscle complex; ACTC1, an actin expressed in cardiac muscle; and MYLPF, another part of the myosin complex. The genes with the most human-biased expression at two months are PMCH, which is involved in melanin production; TTR, a thyroid hormone transport protein; and TBR1, a DNA-binding protein involved in neural migration.

Gene	Log-2 fold change	Time point
NNAT	9.60	1 month
TDGF1	8.60	1 month
EPCAM	7.21	1 month
FEZF1	-6.47	1 month
PAX6	-6.28	1 month
LHX5	-5.94	1 month
MYL1	14.49	2 months
ACTC1	13.14	2 months
MYLPF	13.00	2 months
PMCH	-8.78	2 months
TTR	-8.58	2 months
TBR1	-8.27	2 months

Table 4.1: Most differentially expressed genes between human and Neanderthal cell lines at one and two months, with log-2 fold changes (LFCs). Positive LFCs indicate higher expression in human than Neanderthal cells while negative LFCs indicate higher expression in Neanderthal than human cells.

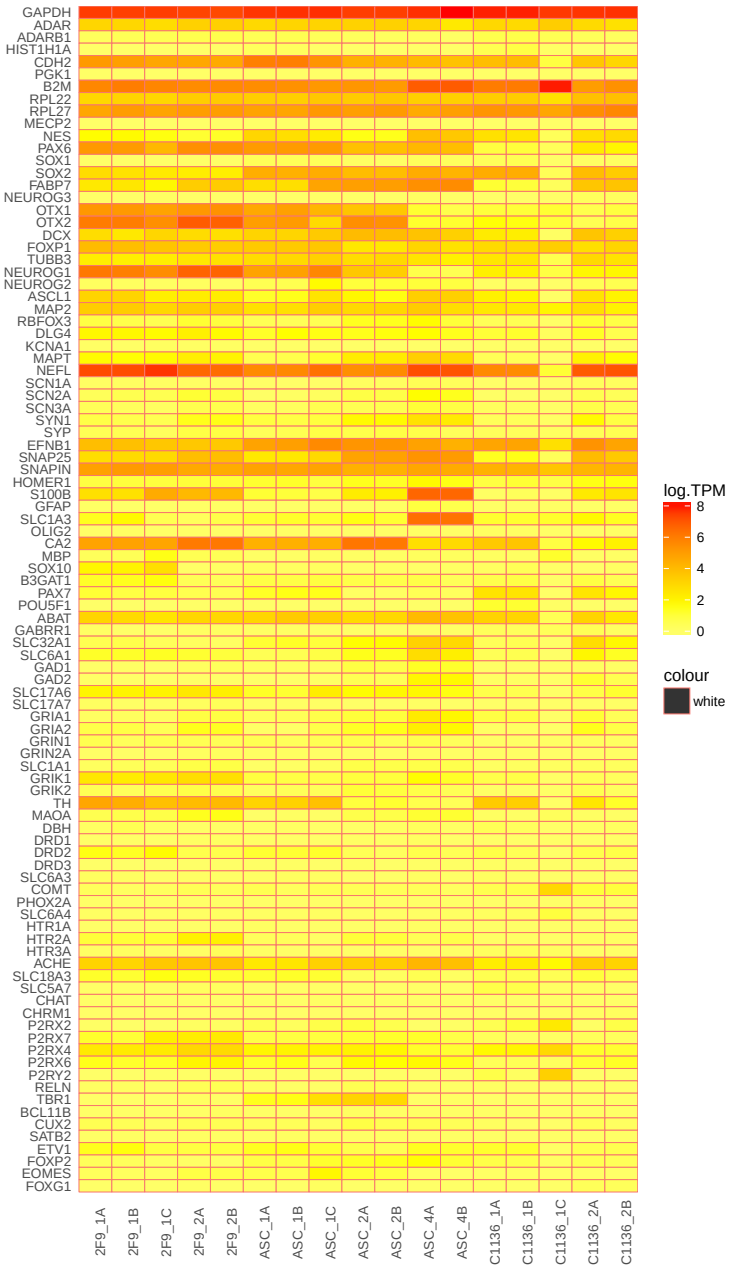

Figure 4.6: Gene expression for 96 key genes involved in neural development across all time points and cell lines.

4.2.4 The Neanderthal version of NOVA1 causes changes in the splicing of genes involved in brain development

We examined the genes with the largest significant differences in alternative splicing between human and Neanderthal cell lines at one and two months. Most of these genes are involved in neural developmental processes, and some may be explanatory of the different phenotypes found in the human versus Neanderthal organoids.

HOMER3 is a member of the HOMER family of scaffold proteins, which are localized in the postsynaptic density. HOMER proteins are involved in calcium signaling and all have multiple isoforms known to fulfill different roles (Shiraishi-Yamaguchi and Furuichi, 2007). *HOMER3* is differentially spliced between humans and Neanderthals at both one and two months, as shown in **Figure 4.7**. A different last exon is used between the two isoforms differentially expressed. In human cells at one month, only the earlier last exon is used, while in Neanderthal cells, both isoforms are expressed. However, at two months, the Neanderthal cells use only the earlier last exon while the human cells express both isoforms.

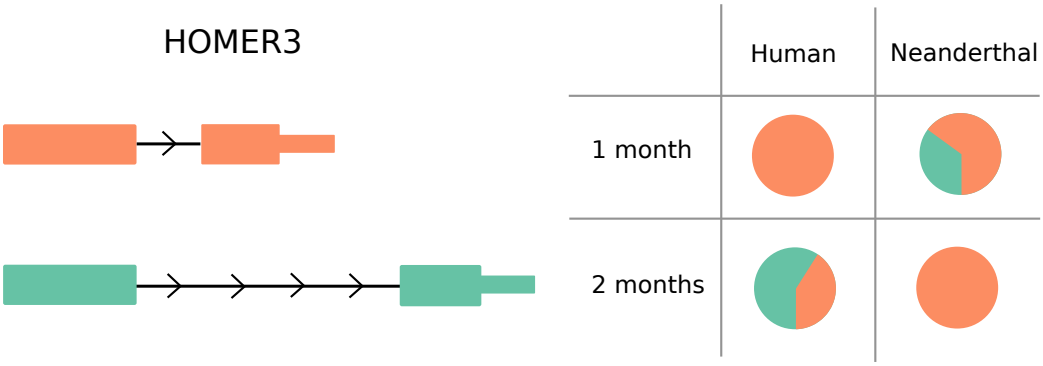

Figure 4.7: Differential last exon usage in *HOMER3* between human and Neanderthal cells at one and two months.

Two other genes of note that are differentially expressed at one month are *COMMD5* and *ANP32E*. *COMMD5* is a calcium-regulated gene (Matsuda et al., 2014) that inhibits cell proliferation (Maine and Burstein, 2007). Two alternate isoforms use different 5' splice sites. In Neanderthal cells, the downstream splice site is always used, while in human cells, both isoforms are expressed (**Figure 4.8**a). Usage of the downstream splice site causes inclusion of an upstream open reading frame (uORF) in the mature mRNA. *ANP32E* is involved in cerebellar synaptogenesis (Costanzo et al., 2006). Two different isoforms of this gene use different 3' splice sites. Neanderthal cells use only the upstream splice site, while human cells express both isoforms (**Figure 4.8**b).

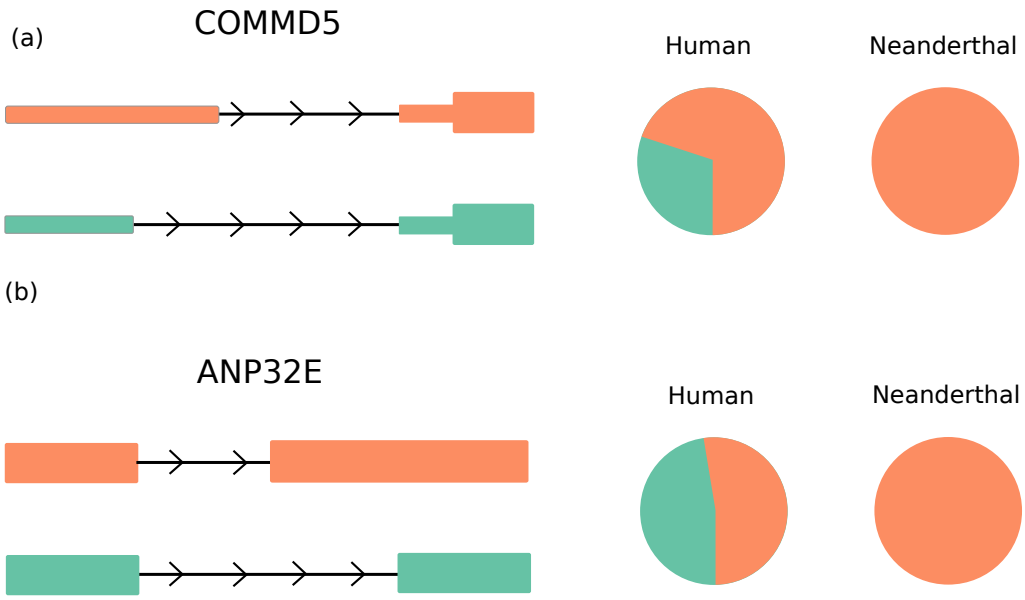

Figure 4.8: Alternative splicing of *COMMD5* (a) and *ANP32E* (b) between human and Neanderthal cells at one month.

At 2 months, *SEPT5* and *TPM3* both have differential first exon usage between human and Neanderthal cells, as shown in **Figure 4.9**. *SEPT5* is a gene involved in dopamine-

dependent neurotoxicity (Son et al., 2005), and the Neanderthal cells exclusively use a first exon containing a uORF while the human cells almost exclusively use a first exon without a uORF. *TPM3* is a known NOVA1 target (Irimia et al., 2011) with different first exon usage between humans and Neanderthals at two months. The human cells only use the downstream first exon while the Neanderthal cells express both isoforms.

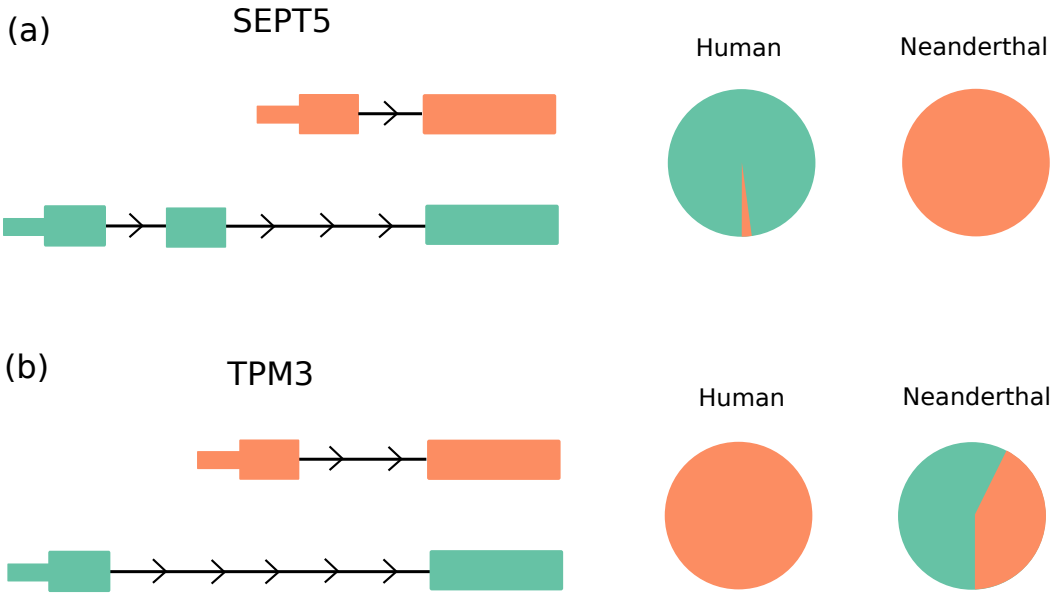

Figure 4.9: Alternative splicing of *SEPT5* (a) and *TPM3* (b) between human and Neanderthal cells at two months.

GNAS and *BIN1* are two other known NOVA1 targets differentially spliced at two months (**Figure 4.10**). *GNAS* has three first exons used by either human or Neanderthal at two months. Human cells primarily use the furthest upstream exon while Neanderthal cells primarily use the furthest downstream exon. *BIN1* has three different cassette splicing events with differential usage between human and Neanderthal cells. One of these is a coordinated cassette event, in which two adjacent cassette exons are either both used or neither used.

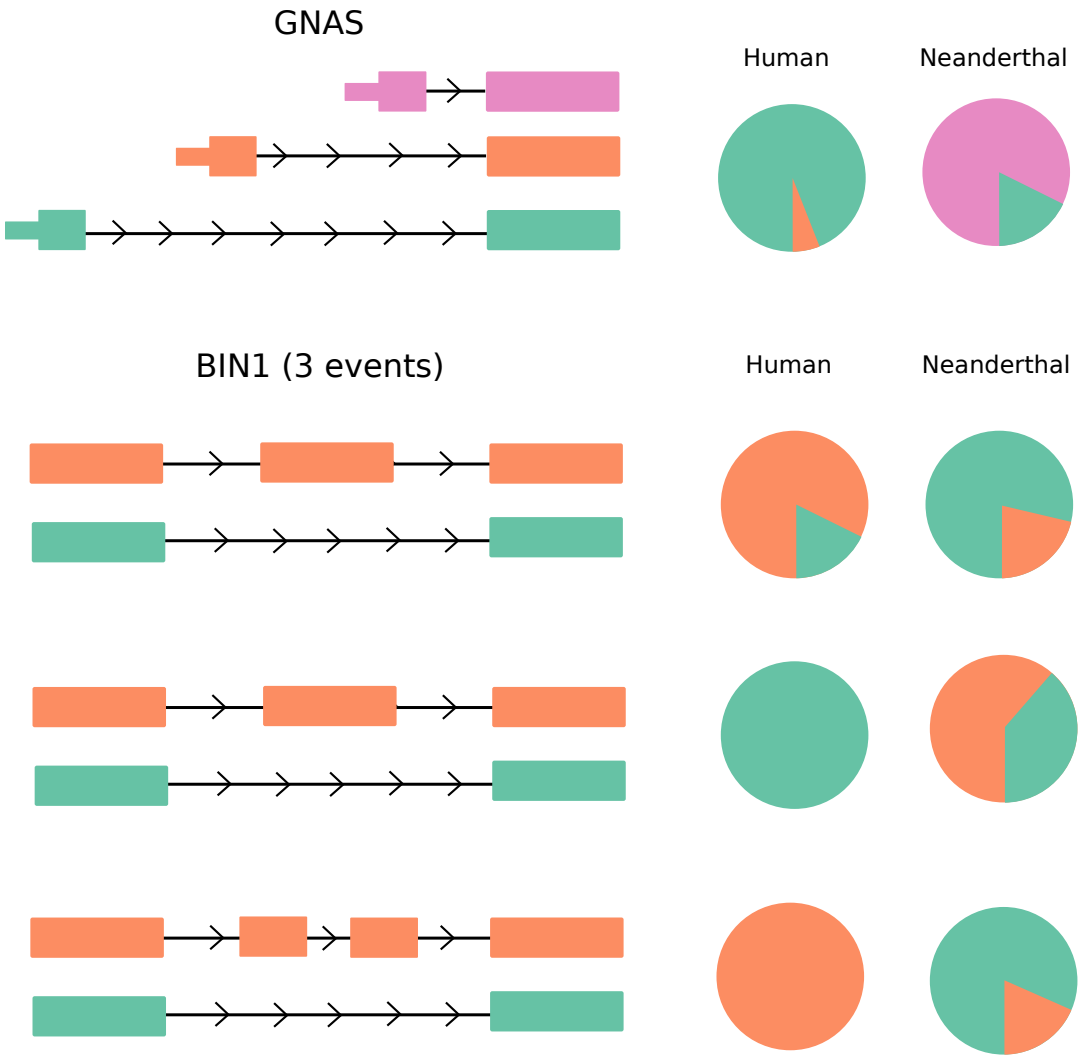

Figure 4.10: Alternative splicing of *GNAS*(a) and *TPM3* (b) at two months. *GNAS* has three different possible first exons, while *TPM3* has three separate cassette splicing events which are expressed at different rates in human and Neanderthal cells at two months.

We also performed a Gene Ontology (GO) enrichment analysis on the sets of genes differentially spliced to find overrepresented functions in these genes. These results are in **Table 4.2** (one month) and **Table 4.3** (two months).

Term ID	Term description	FDR α
GO:0005887	Integral component of plasma membrane	9.4×10^{-17}
GO:0005509	Calcium ion binding	2.3×10^{-25}
GO:0007156	Homophilic cell adhesion via plasma membrane adhesion molecules	5.5×10^{-38}
GO:0007399	Nervous system development	5.2×10^{-19}
GO:0007267	Cell-cell signaling	2.8×10^{-14}
GO:0016339	Calcium-dependent cell-cell adhesion via plasma membrane cell adhesion molecules	5.3×10^{-3}
GO:0051085	Chaperone mediated protein folding requiring cofactor	5.3×10^{-3}

Table 4.2: Enriched GO terms in set of genes differentially spliced in human versus Neanderthal NOVA1 cell lines at one month.

Term ID	Term description	FDR α
GO:0005887	Integral component of plasma membrane	7.9×10^{-14}
GO:1990023	Mitotic spindle midzone	3.4×10^{-3}
GO:0005509	Calcium ion binding	2.3×10^{-22}
GO:0097718	Disordered domain specific binding	2.1×10^{-3}
GO:0007156	Homophilic cell adhesion via plasma membrane adhesion molecules	3.5×10^{-20}
GO:0007267	Cell-cell signaling	4.9×10^{-11}
GO:0060789	Hair follicle placode formation	3.8×10^{-4}
GO:0070527	Platelet aggregation	1.8×10^{-3}
GO:0061077	Chaperone-mediated protein folding	8.5×10^{-3}
GO:0016339	Calcium-dependent cell-cell adhesion via plasma membrane cell adhesion molecules	9.8×10^{-3}
GO:0031116	Positive regulation of microtubule polymerization	9.8×10^{-3}

Table 4.3: Enriched GO terms in set of genes differentially spliced in human versus Neanderthal NOVA1 cell lines at two months.

4.2.5 Motif search

NOVA1 promotes or inhibits cassette exon inclusion in mature mRNAs based on where it binds to pre-mRNAs relative to splice sites (Ule et al., 2006). Therefore, we searched for the NOVA1 binding motif near splice sites involved in differential cassette usage between human and Neanderthal cell lines using the YCAY cluster score as described by Ule et al. (2006). We found that splice sites involved in differential cassette usage at one month were enriched for NOVA1's YCAY binding motif, with 21 out of 106 (19.8%) differentially spliced cassette splice sites having a net YCAY score > 1 versus 5307 out of 58294 (9.1%) across all cassettes (Fisher's Exact Test $p = 5.4 \times 10^{-4}$).

In our Neanderthal NOVA1 cell lines, the Neanderthal version of NOVA1 is present in a genome that is otherwise that of a modern human. Therefore, we tested whether the sequences around splice sites involved in differential cassette usage between human and Neanderthal versions of NOVA1 were different between Neanderthals and modern humans in a way that affects the strength of the binding motif. We performed this test by calculating the YCAY cluster scores around these splice sites using both the human reference genome and the Vindija Neanderthal genome. We found no major differences in YCAY cluster scores between the human and Neanderthal genomes near splice sites involved in differential cassette usage between the human and Neanderthal cell lines (**Figure 4.11**). The only place YCAY scores differ at all between the human and Neanderthal sequences around affected splice sites is in the gene *CD74*, where a single base pair change in the intronic sequence 151nt downstream of the cassette gives the human sequence a higher YCAY score.

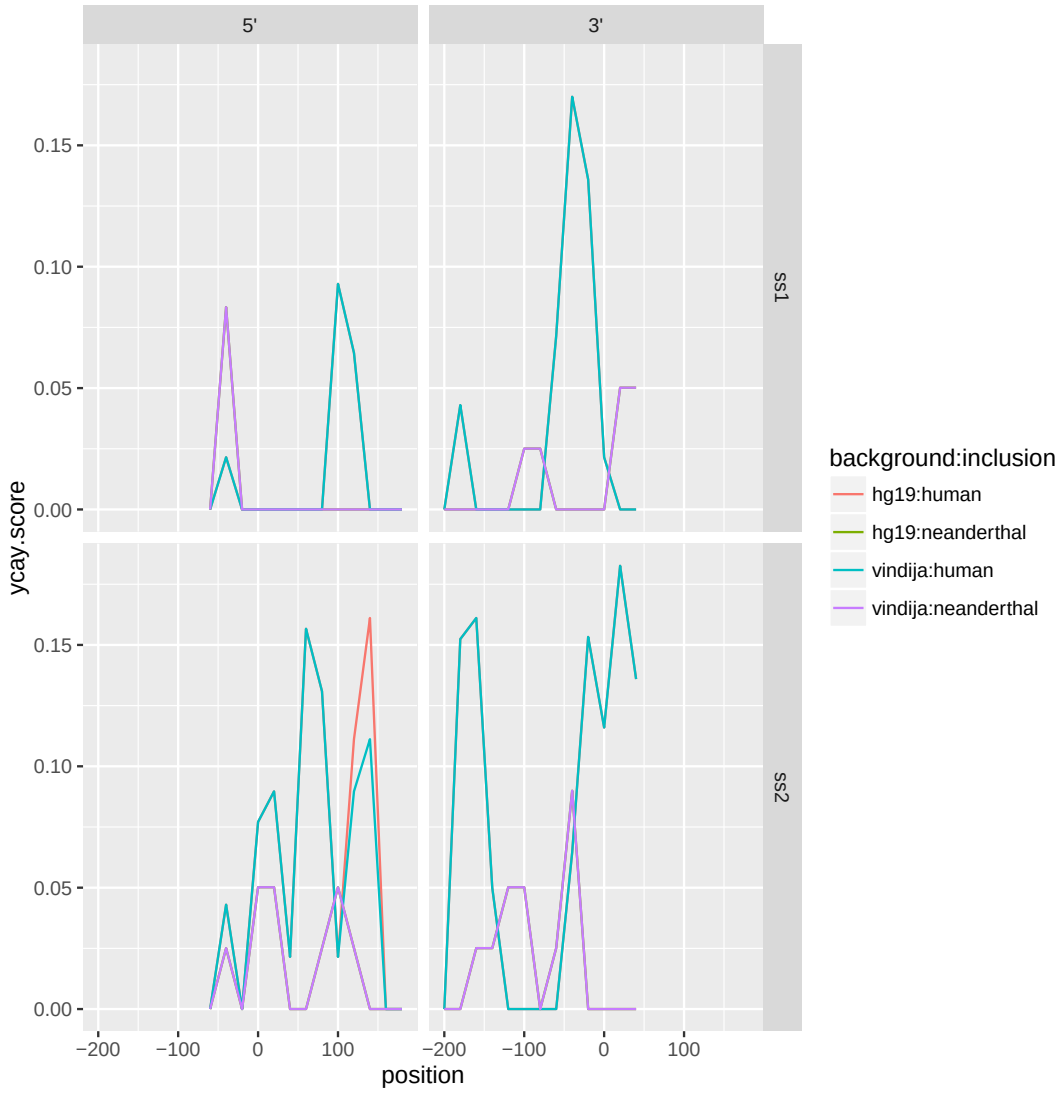

Figure 4.11: Mean YCAY cluster scores for sliding windows around splice sites involved in differential cassette inclusion between human and Neanderthal cell lines. We divided differential cassette inclusion events between those in which the cassette was included more often in human than in Neanderthal and those in which the cassette was included more often in Neanderthal than human. We calculated mean YCAY cluster scores in sliding windows around the four intron-exon boundaries involved in a cassette inclusion event using both the reference human and Vindija Neanderthal genomes, finding only one window (5'ss2, position +140) where the YCAY cluster score differed between these two genomes.

The position-dependent effect of YCAY clusters on splicing has only been determined

with respect to cassette exon inclusion. Given that we found cassette exon inclusion events to be differentially spliced less often than other AS event types, such as alternate first and last exon usage, we performed a de novo search for overrepresented motifs in the sequence around all splice sites differentially used between the human and Neanderthal cell lines. In splice sites differentially used at one month, the most overrepresented motif is the pyrimidine-rich YTK-BYHYYKBYYYHYYYYY, with log score 3738.2. At two months, the most overrepresented motif is KBYYWBYKYYYKNBWBYBYYY, also pyrimidine-rich, with log score 5148.9. Intronic polypyrimidine tracts are known to be involved in spliceosome assembly (Reed, 1989), so finding these motifs enriched near splice sites is unsurprising.

4.2.6 No evidence of depletion for Neanderthal ancestry in targets of NOVA1

The RNA-binding domains of NOVA1 are highly conserved, causing NOVA1 to have similar binding-position-dependent effects on splicing in taxa as distantly related as *Drosophila* and mammals (Brooks et al., 2011). However, the binding targets of NOVA1 are much more divergent, leading NOVA1 to bind to different pre-mRNAs in the genomes of different animals (Jelen et al., 2007). This means that our Neanderthal cell lines do not show how NOVA1 would have behaved in developing Neanderthal brains, but rather how the Neanderthal version of NOVA1 interacts with a modern human genetic background. Therefore, it is possible that the differential splicing, and thus the different phenotypes, we observe between cell lines with the human versus Neanderthal version of NOVA1 is a result of incompatibility between the Neanderthal version of NOVA1 and the modern human genetic background (**Figure 4.12**). So, we sought to determine whether the splice sites involved in differential alternative splicing between

human and Neanderthal versions of NOVA1 are depleted for Neanderthal ancestry in modern humans.

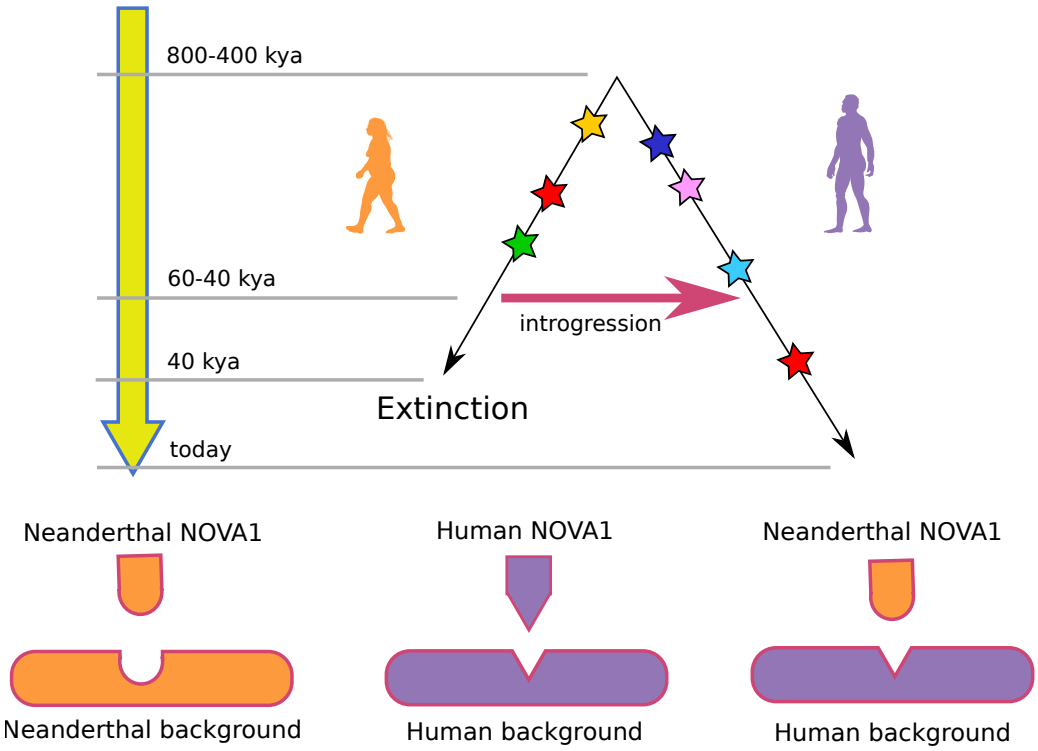

Figure 4.12: Because our "Neanderthal" cell lines contain the Neanderthal version of NOVA1 but an otherwise modern human genetic background, it is possible that the phenotypic and splicing differences we observe between these cell lines is a result of an incompatibility between the Neanderthal version of NOVA1 and the human genetic background.

We first tested for depletion of Neanderthal ancestry among modern humans in the genes differentially spliced between the human and Neanderthal Nova1 organoids using published coordinates of Neanderthal "deserts" where modern humans are depleted for Neanderthal ancestry from Vernot and Akey (2014) and Sankararaman et al. (2014). For genes differentially spliced at one month, 28.6% of their summed length intersects with Vernot and Akey deserts

and 47.7% intersects with Sankararaman et al. deserts. At two months, 42.4% intersects with Vernot and Akey deserts and 45.4% intersects with Sankararaman et al. deserts. Based on permutation tests, none of these values were significantly higher or lower than the desert intersection percentage over the summed length of all protein-coding genes in the genome: 43.2% for Vernot and Akey deserts and 38.4% for Sankararaman et al. deserts.

We next tested for depletion of Neanderthal ancestry at splice sites involved in differential alternative splicing between human and Neanderthal versions of NOVA1 by counting the number of locations near these splice sites where the Vindija (Prüfer et al., 2017) or Altai (Prüfer et al., 2014) Neanderthal genome differs from all human genomes in the 1000 genomes panel (1000 Genomes Project Consortium et al., 2015). We call these Neanderthal-specific variants or NSVs. For genes differentially spliced at one month, we found 486 NSVs per Mb near splice sites using the Vindija genome and 660 NSVs per Mb using the Altai genome. At two months, we found 229 NSVs per Mb near splice sites using the Vindija genome and 679 NSVs per Mb using the Altai genome. All of these values were larger than the average rate of NSVs per Mb near splice sites across the genome: 221 NSVs per Mb for the Vindija genome and 601 NSVs per Mb for the Altai genome. Based on permutation tests, the most significant enrichment of NSVs per Mb was around splice sites involved in differential alternative splicing at one month using the Vindija genome, with $p = 0.0253$ (**Figure 4.13**).

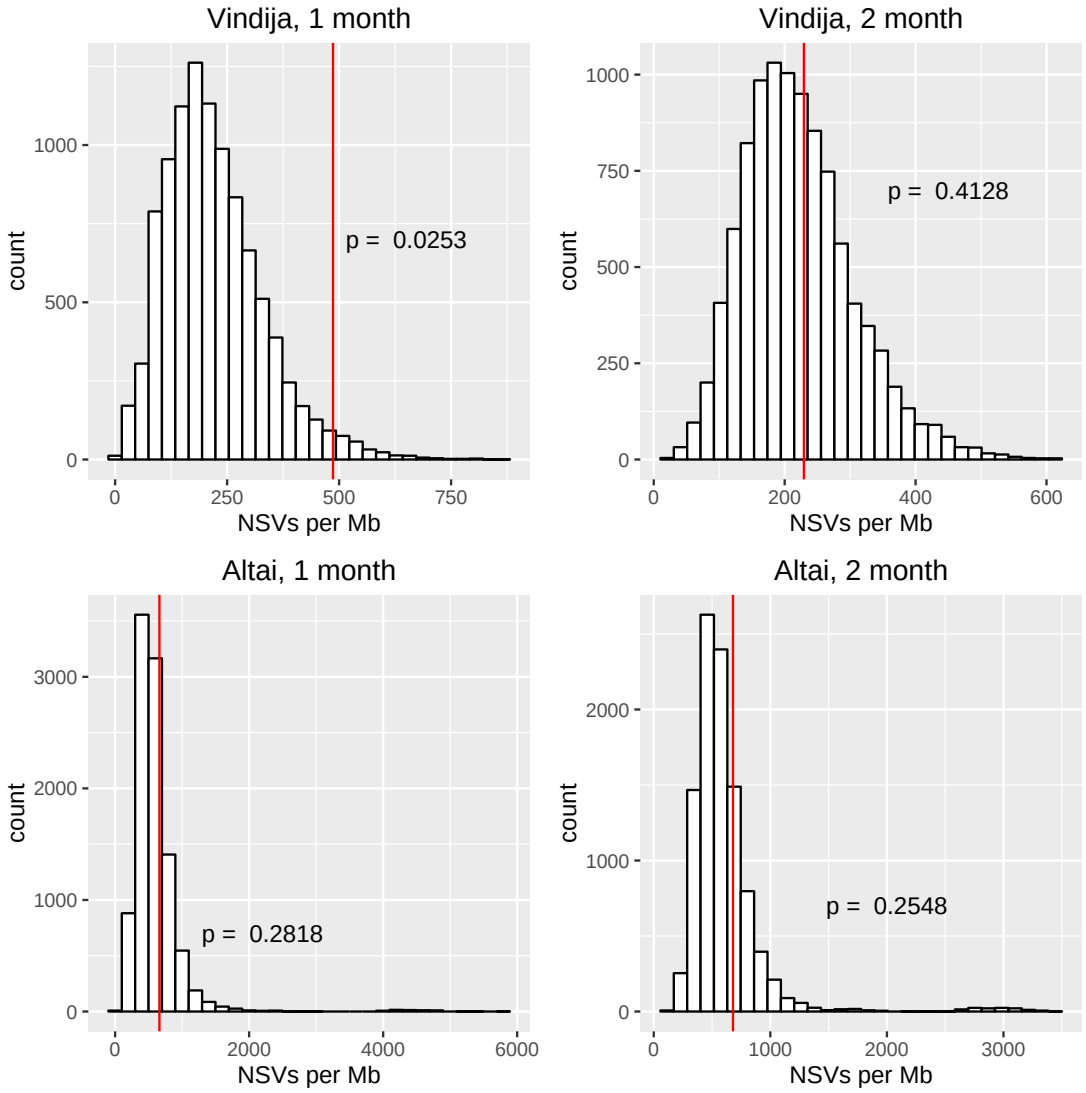

Figure 4.13: Results of permutation tests to determine the significance of the number of Neanderthal-specific variants per Mb found near splice sites involved in differential alternative splicing at one month and two months using the Vindija and Altai Neanderthal genomes. Permutation values are shown as histograms and the actual number of Neanderthal-specific variants per Mb is shown as a red vertical line in each histogram.

We also tested whether allele frequencies of Neanderthal-introgressed variants are different near differentially regulated splice sites compared to all splice sites and the whole

genome. We consider a variant to be Neanderthal-introgressed if it is present in the Neanderthal genome, not present in present-day African populations to account for incomplete lineage sorting, and present in present-day European populations. For the 94,653 sites that meet this definition, the mean allele frequency in present-day European populations is 8.46×10^{-3}. For the 5,597 of these sites that are near splice sites, the mean allele frequency in present-day European populations is 7.41×10^{-3}. For the 8 of these variants near 1-month differentially regulated splice sites, the mean allele frequency in Europeans is 1.46×10^{-2}; and for the seven 2-month differentially regulated splice sites, 9.94×10^{-3}. Neither of these values is significantly less than the mean allele frequency around all splice sites in the genome.

Given the results of these three analyses, we do not find sufficient evidence to conclude that the genes differentially spliced by the human and Neanderthal versions of NOVA1 are depleted for Neanderthal ancestry.

4.3 Discussion

NOVA1 is an important regulator of splicing in genes involved in neural development. Modern humans have a different version of NOVA1 than Neanderthals did, with a single amino acid substitution in the KH2 RNA binding domain of the protein. The Neanderthal version of the KH2 domain of NOVA1 is the ancestral allele as other amniotes share this sequence with Neanderthals, meaning that the modern human version is private to modern humans. Despite the Neanderthal ancestry present in various parts of the genomes of modern humans due to incomplete lineage sorting and admixture, no modern humans today have the Neanderthal version

of NOVA1.

We grew stem cells containing the human or Neanderthal version of NOVA1 into neural progenitor cells and then mature neurons. These cell lines display significant phenotypic differences during development, with neural organoids with the Neanderthal version of NOVA1 smaller and less spherical than the human neural organoids. We measured relative splicing event frequency between the human and Neanderthal cell lines during both neural differentiation and maturation, and found that the two different cell lines display significant differences in the splicing of many genes, a significant number of which are involved in communication and adhesion between cells and therefore may potentially explain the phenotypic differences between the cell lines.

The effects on splicing and phenotype of changing the NOVA1 to the Neanderthal version are clear, but these could be the result of two different causes. It is possible that the splicing and phenotype differences we observe are indicative of a difference that existed in vivo between Neanderthal and modern human brains, but it is also possible that they are the result of an incompatibility between the Neanderthal version of NOVA1 and the human genetic background. The cells we call "Neanderthal" are not purely genetically Neanderthal, but rather contain human genomes edited such that they have the Neanderthal version of a single gene, NOVA1, which is a trans-regulatory RNA-binding protein that regulates splicing of other genes through binding to their transcripts.

To test whether the changes we observe are the result of an incompatibility between the Neanderthal version of NOVA1 and the modern human genetic background, we looked for depletion of Neanderthal ancestry around the splice sites differentially used in cell lines with

the human and Neanderthal versions of NOVA1. We found no significant evidence that there is a depletion of Neanderthal ancestry around these splice sites, or that there are differences in the strength of NOVA1 binding motifs around these splice sites. However, we did not find significant evidence to the contrary, either. Testing this more conclusively will require further modifications to these cells to make the genetic background closer to that of Neanderthals.

4.4 Methods

4.4.1 Construction of RNA-seq Libraries

All RNA libraries were prepared using a modified SmartSeq2 method (Byrne et al., 2017). A total of 2 µL of RNA (50 ng) of each sample was reverse transcribed using Smartscribe Reverse Transcriptase (Clontech) in a 10 µL reaction containing a Smart-seq2 TSO according to manufacturer's instructions for 60 min at 42°C (Picelli et al., 2014b). The resulting cDNA was treated with 1 µL of 1:10 dilutions of RNAse A (Thermo) and Lambda Exonuclease (NEB) for 30 min at 37°C. The cDNA was then amplified using KAPA Hifi Readymix 2x (KAPA) and incubated at 95°C for 3 min, followed by 15 cycles of (98°C for 20 s, 67°C for 15 s, 72°C for 4 min), with a final extension of 72°C for 5 min. The resulting polymerase chain reaction (PCR) product was then treated with our Tn5 enzyme (Picelli et al., 2014a) custom loaded with Tn5ME-A/R and Tn5ME-B/R. The Tn5 reaction was performed using 5 µL of the amplified product, 1 µL of the loaded Tn5 enzyme, 10 µL of H2O and 4 µL of 5x TAPS-PEG buffer. The sample was then incubated at 55°C for 5 min. The Tn5 reaction was then inactivated using 5 µL of 0.2% sodium dodecyl sulphate (SDS). 5 µL of the Tn5 product was then nick-

translated at 72°C for 6 min and further amplified using KAPA Hifi Polymerase (KAPA) using both Nextera_Primer_B and Nextera_Primer_A primers. The sample was incubated at 98°C for 30 s, followed by 10 cycles of (98°C for 10 s, 63°C for 30 s, 72°C for 2 min) with a final extension at 72°C for 5 min. The Tn5 treated PCR product was then size selected using a 2% EX E-gel (Thermo) to a size range of 300-850 bp. All libraries were quantified using qPCR and Qubit prior to sequencing and pooled equally based on concentration. Furthermore, all libraries were prepared in parallel on the same day and pooled once prior to sequencing so they should be at similar ratios on each sequencing lane, which should mitigate any batch effects. The libraries were sequenced on an Illumina HiSeq 2x100 run on 2 lanes.

4.4.2 Edit verification

To ensure that the expected version of NOVA1 was being expressed in each sample, we amplified a 204 bp sequence around the variant base in the cDNA library with the primers 5' GGTAAAGATTATAGTTCCCAACAGC 3' and 5' CTTCTGGATGATAAGTTCAACAGC 3' using KAPA Taq polymerase with the provided kit protocol, an annealing temperature of 61°C, 40 cycles, and 20 µL reaction volumes. We then ran the product on a gel and purified the band at 204bp with the Zymoclean Gel DNA Recovery kit. Finally, we Sanger-sequenced the purified DNA with the same primers on an Applied Biosystems 310 Genetic Analyzer and compared it with the reference cDNA.

4.4.3 Expression quantification

To measure gene expression across all samples, we first aligned reads to hg19 using TopHat2 v2.0.8 (Kim et al., 2013), a spliced aligner, with default parameters. We then calculated raw counts of fragments mapping to gene features in each library using featureCounts v1.5.1 (Liao et al., 2013) with the parameters `-t exon -p -g gene_id` and GENCODE v19 (Harrow et al., 2012) as the annotation set. For comparisons of expression levels across both genes and samples, we normalized the raw counts using the transcripts per million (TPM) normalization method (Wagner et al., 2012). For differential expression analysis, we used DESeq2 with unnormalized raw counts and parameters `lfcThreshold = 1, altHypothesis = "greaterAbs", alpha = 0.01` (Love et al., 2014).

4.4.4 Splicing quantification

We quantified and compared splicing between samples using juncBASE (Brooks et al., 2011). We ran juncBASE on the read alignments, which we created as described above, using GENCODE v19 (Harrow et al., 2012) as the annotation set and the parameters `--by_chr` in steps 1B, 2, 4, and 5; `--majority_rules` in step 1B; and `--jcn_seq_len 188` in steps 5 and 6. To call differentially spliced events, we used the pairwise Fisher's test script with parameters `--jcn_seq_len 188 --method BH`. We considered a splicing event to be differentially spliced if the replicates of the human control were not significantly different from each other but the replicates of the sample being compared to the control were all significantly different from the control.

4.4.5 Motif search

To calculate YCAY cluster scores, we implemented the scoring method described by Ule et al. (2006), who searched for YCAY clusters in sliding 25bp windows in the 205bp of intronic and 65bp of exonic sequence adjacent to splice sites. For the general motif search, we used GLAM2 version 4.12.0, part of the MEME suite (Bailey et al., 2015), with default parameters.

4.4.6 Gene Ontology analysis

We found Gene Ontology (GO) terms that were significantly enriched among genes with differential splicing using func v0.4.8 (Prüfer et al., 2007). For a given time point, we used as a background set all genes with expression level higher than 5 TPM in at least one replicate at that time point. We then used the `func_hyper` script and the refinement script it creates to calculate FDR-corrected p-values for overrepresentation of terms.

4.4.7 Testing for depletion of Neanderthal ancestry

For the published desert analysis, we used bed files containing the genomic coordinates of Neanderthal deserts from Vernot and Akey (2014) or Sankararaman et al. (2014). We created bed files containing the genomic coordinates of the full lengths of genes differentially spliced between the human and Neanderthal Nova1 organoids. We used bedtools intersect v2.25.0 (Quinlan and Hall, 2010) with default parameters to calculate the intersections between differentially spliced genes and Neanderthal deserts. We calculated the percentage of the lengths of differentially spliced genes intersecting with deserts by dividing the summed length of all in-

tersections in the bedtools output by the summed length of all differentially spliced genes. To assign a p-value to this statistic, we performed a permutation test with 10,000 permutations in which for each permutation we randomly selected N genes from the full set of protein-coding genes where N is the number of differentially spliced genes and calculated the same statistic.

We found NSVs by comparing the published genotypes for either the Altai Neanderthal (Prüfer et al., 2014) or the Vindija Neanderthal (Prüfer et al., 2017) to the 1000 Genomes Project panel of modern human genotypes (1000 Genomes Project Consortium et al., 2015). We consider a locus to be the site of an NSV if at least one of the two haplotypes in the given Neanderthal genome differs from the reference (hg19) but none of the haplotypes in the 1000 Genomes panel differ from the reference. To count the number of NSVs near splice sites involved in differential alternative splicing between human and Neanderthal version of NOVA1, we used bedtools intersect with default parameters to find all the NSVs in the 205bp of intronic sequence and 65bp of exonic sequence directly adjacent to an intron-exon boundary determined by juncBASE to be involved in a differential splicing event. We chose these distances from intron-exon boundaries because these were the windows within which Ule et al. (2006) found NOVA1 binding to alter splicing. We then calculated the number of NSVs per megabase by dividing the number of NSVs by the summed length of all regions around splice sites involved in differential alternative splicing events. To assign a p-value to this statistic, we performed a permutation test with 10,000 permutations in which for each permutation we randomly selected N genes from the full set of protein-coding genes where N is the number of differentially spliced genes and calculated the same statistic.

For the allele-frequency analysis, we used the Vindija genome as the source of in-

trogression, 1000 genomes population EUR as the reference population, and 1000 genomes population AFR as the outgroup. We considered a variant to be Neanderthal-introgressed if it is present in the Neanderthal genome, has allele frequency $AF = 0$ in the AFR population, and has allele frequency $AF > 0$ in EUR.

Bibliography

1000 Genomes Project Consortium et al. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68.

Bailey, T. L., Johnson, J., Grant, C. E., and Noble, W. S. (2015). The meme suite. *Nucleic acids research*, 43(W1):W39–W49.

Bellone, R. R., Holl, H., Setaluri, V., Devi, S., Maddodi, N., Archer, S., Sandmeyer, L., Ludwig, A., Foerster, D., Pruvost, M., et al. (2013). Evidence for a retroviral insertion in TRPM1 as the cause of congenital stationary night blindness and leopard complex spotting in the horse. *PLoS One*, 8(10):e78280.

Brooks, A. N., Yang, L., Duff, M. O., Hansen, K. D., Park, J. W., Dudoit, S., Brenner, S. E., and Graveley, B. R. (2011). Conservation of an RNA regulatory map between Drosophila and mammals. *Genome Research*, 21(2):193–202.

Buckanovich, R. J. and Darnell, R. B. (1997). The neuronal RNA binding protein Nova-1 recognizes specific RNA targets in vitro and in vivo. *Molecular and cellular biology*, 17(6):3194–3201.

Buckanovich, R. J., Yang, Y., and Darnell, R. B. (1996). The onconeural antigen Nova-1 is a neuron-specific rna-binding protein, the activity of which is inhibited by paraneoplastic antibodies. *Journal of Neuroscience*, 16(3):1114–1122.

Bull, J. J., Gutzke, W. H., and Crews, D. (1988). Sex reversal by estradiol in three reptilian orders. *General and Comparative Endocrinology*, 70(3):425–428.

Byrne, A., Beaudin, A. E., Olsen, H. E., Jain, M., Cole, C., Palmer, T., DuBois, R. M., Forsberg, E. C., Akeson, M., and Vollmers, C. (2017). Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nature Communications*, 8:16027.

Chan, C. S. and Song, J. S. (2008). CCCTC-binding factor confines the distal action of estrogen receptor. *Cancer Research*, 68(21):9041–9049.

Coleman, S., Zeng, Z., Wang, K., Luo, S., Khrebtukova, I., Mienaltowski, M., Schroth, G., Liu, J., and MacLeod, J. (2010). Structural annotation of equine protein-coding genes determined by mRNA sequencing. *Animal genetics*, 41(s2):121–130.

Costanzo, R. V., Vilá-Ortíz, G. J., Perandones, C., Carminatti, H., Matilla, A., and Radrizzani, M. (2006). Anp32e/cpd1 regulates protein phosphatase 2a activity at synapses during synaptogenesis. *European Journal of Neuroscience*, 23(2):309–324.

Davis, L., Meyer, K., Rudd, D., Librant, A., Epping, E., Sheffield, V., and Wassink, T. (2008). Pax6 3 deletion results in aniridia, autism and mental retardation. *Human genetics*, 123(4):371–378.

Ferguson, M. W. and Joanen, T. (1983). Temperature-dependent sex determination in *Alligator mississippiensis*. *J. Zool.*, 200(2):143–177.

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., Orlov, Y. L., Velkov, S., Ho, A., Mei, P. H., et al. (2009). An oestrogen-receptor-α-bound human chromatin interactome. *Nature*, 462(7269):58.

Gardeisen, A. (1999). Middle palaeolithic subsistence in the west cave of "le portel" (pyrénées, france). *Journal of Archaeological Science*, 26(9):1145–1158.

Gaunitz, C., Fages, A., Hanghøj, K., Albrechtsen, A., Khan, N., Schubert, M., Seguin-Orlando, A., Owens, I. J., Felkel, S., Bignon-Lau, O., et al. (2018). Ancient genomes revisit the ancestry of domestic and Przewalskis horses. *Science*, page eaao3297.

Green, R. E., Braun, E. L., Armstrong, J., Earl, D., Nguyen, N., Hickey, G., Vandewege, M. W., John, J. A. S., Capella-Gutiérrez, S., Castoe, T. A., et al. (2014). Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. *Science*, 346(6215):1254449.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., et al. (2010). A draft sequence of the Neandertal genome. *Science*, 328(5979):710–722.

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: the reference human genome annotation for the ENCODE project. *Genome research*, 22(9):1760–1774.

Heide, M., Zhang, Y., Zhou, X., Zhao, T., Miquelajáuregui, A., Varela-Echavarría, A., and Alvarez-Bolado, G. (2015). Lhx5 controls mamillary differentiation in the developing hypothalamus of the mouse. *Frontiers in neuroanatomy*, 9:113.

Higham, T., Douka, K., Wood, R., Ramsey, C. B., Brock, F., Basell, L., Camps, M., Arrizabalaga, A., Baena, J., Barroso-Ruíz, C., et al. (2014). The timing and spatiotemporal patterning of Neanderthal disappearance. *Nature*, 512(7514):306.

Irimia, M., Denuc, A., Burguera, D., Somorjai, I., Martín-Durán, J. M., Genikhovich, G., Jimenez-Delgado, S., Technau, U., Roy, S. W., Marfany, G., et al. (2011). Stepwise assembly of the nova-regulated alternative splicing network in the vertebrate brain. *Proceedings of the National Academy of Sciences*, 108(13):5319–5324.

Jelen, N., Ule, J., Živin, M., and Darnell, R. B. (2007). Evolution of Nova-dependent splicing regulation in the brain. *PLoS genetics*, 3(10):e173.

Jensen, K. B., Dredge, B. K., Stefani, G., Zhong, R., Buckanovich, R. J., Okano, H. J., Yang, Y. Y., and Darnell, R. B. (2000). Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. *Neuron*, 25(2):359–371.

Kalbfleisch, T. S., Rice, E., DePriest, M. S., Walenz, B. P., Hestand, M. S., Vermeesch, J. R., O'Connell, B. L., Fiddes, I. T., Vershinina, A. O., Petersen, J. L., Finno, C. J., Bellone, R. R., McCue, M. E., Brooks, S. A., Bailey, E., Orlando, L., Green, R. E., Miller, D. C., Antczak, D. F., and MacLeod, J. N. (2018). Equcab3, an updated reference genome for the domestic horse. *bioRxiv*.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, 14(4):R36.

Lance, V. A. (2009). Is regulation of aromatase expression in reptiles the key to understanding temperature-dependent sex determination? *Journal of Experimental Zoology Part A: Ecological and Integrative Physiology*, 311(5):314–322.

Liao, Y., Smyth, G. K., and Shi, W. (2013). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930.

Lin, H.-H., Bell, E., Uwanogho, D., Perfect, L. W., Noristani, H., Bates, T. J., Snetkov, V., Price, J., and Sun, Y.-M. (2010). Neuronatin promotes neural lineage in escs via ca2+ signaling. *Stem Cells*, 28(11):1950–1960.

Litvinov, S. V., Velders, M. P., Bakker, H., Fleuren, G. J., and Warnaar, S. O. (1994). Ep-cam: a human epithelial antigen is a homophilic cell-cell adhesion molecule. *The Journal of cell biology*, 125(2):437–446.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12):550.

Maine, G. N. and Burstein, E. (2007). Commd proteins: Comming to the scene. *Cellular and molecular life sciences*, 64(15):1997–2005.

Matsuda, H., Hamet, P., and Tremblay, J. (2014). Hypertension-related, calcium-regulated gene (hcarg/commd5) and kidney diseases: Hcarg accelerates tubular repair. *Journal of nephrology*, 27(4):351–360.

McCoy, R. C., Wakefield, J., and Akey, J. M. (2017). Impacts of neanderthal-introgressed sequences on the landscape of human gene expression. *Cell*, 168(5):916–927.

Picelli, S., Björklund, Å. K., Reinius, B., Sagasser, S., Winberg, G., and Sandberg, R. (2014a). Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome research*, 24(12):2033–2040.

Picelli, S., Faridani, O. R., Björklund, Å. K., Winberg, G., Sagasser, S., and Sandberg, R. (2014b). Full-length RNA-seq from single cells using Smart-seq2. *Nature protocols*, 9(1):171.

Prüfer, K., de Filippo, C., Grote, S., Mafessoni, F., Korlević, P., Hajdinjak, M., Vernot, B., Skov, L., Hsieh, P., Peyrégne, S., et al. (2017). A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science*, page eaao1887.

Prüfer, K., Muetzel, B., Do, H.-H., Weiss, G., Khaitovich, P., Rahm, E., Pääbo, S., Lachmann, M., and Enard, W. (2007). Func: a package for detecting significant associations between gene sets and ontological annotations. *BMC bioinformatics*, 8(1):41.

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., De Filippo, C., et al. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505(7481):43.

Putnam, N. H., O'Connell, B. L., Stites, J. C., Rice, B. J., Blanchette, M., Calef, R., Troll, C. J., Fields, A., Hartley, P. D., Sugnet, C. W., et al. (2016). Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome research*, 26(3):342–350.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

Reed, R. (1989). The organization of 3'splice-site sequences in mammalian introns. *Genes & development*, 3(12b):2113–2123.

Rice, E. S., Kohno, S., John, J. S., Pham, S., Howard, J., Lareau, L. F., O'Connell, B. L., Hickey, G., Armstrong, J., Deran, A., et al. (2017). Improved genome assembly of american alligator genome reveals conserved architecture of estrogen signaling. *Genome research*, 27(5):686–696.

Richards, M. P., Pettitt, P. B., Trinkaus, E., Smith, F. H., Paunović, M., and Karavanić, I. (2000). Neanderthal diet at Vindija and Neanderthal predation: the evidence from stable isotopes. *Proceedings of the National Academy of Sciences*, 97(13):7663–7666.

Roebroeks, W. and Villa, P. (2011). On the earliest evidence for habitual use of fire in Europe. *Proceedings of the National Academy of Sciences*, 108(13):5209–5214.

Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N., and Reich, D. (2014). The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*, 507(7492):354.

Shiraishi-Yamaguchi, Y. and Furuichi, T. (2007). The homer family proteins. *Genome biology*, 8(2):206.

Son, J. H., Kawamata, H., Yoo, M. S., Kim, D. J., Lee, Y. K., Kim, S., Dawson, T. M., Zhang, H., Sulzer, D., Yang, L., et al. (2005). Neurotoxicity and behavioral deficits associated with septin 5 accumulation in dopaminergic neurons. *Journal of neurochemistry*, 94(4):1040–1053.

St. John, J. A., Braun, E. L., Isberg, S. R., Miles, L. G., Chong, A. Y., Gongora, J., Dalzell, P., Moran, C., Bed'Hom, B., Abzhanov, A., et al. (2012). Sequencing three crocodilian genomes to illuminate the evolution of archosaurs and amniotes. *Genome Biol.*, 13(1):415.

Teplova, M., Malinina, L., Darnell, J. C., Song, J., Lu, M., Abagyan, R., Musunuru, K., Teplov, A., Burley, S. K., Darnell, R. B., et al. (2011). Protein-rna and protein-protein recognition by dual kh1/2 domains of the neuronal splicing factor nova-1. *Structure*, 19(7):930–944.

Ule, J., Stefani, G., Mele, A., Ruggiu, M., Wang, X., Taneri, B., Gaasterland, T., Blencowe, B. J., and Darnell, R. B. (2006). An RNA map predicting Nova-dependent splicing regulation. *Nature*, 444(7119):580.

Ule, J., Ule, A., Spencer, J., Williams, A., Hu, J.-S., Cline, M., Wang, H., Clark, T., Fraser, C., Ruggiu, M., et al. (2005). Nova regulates brain-specific splicing to shape the synapse. *Nature genetics*, 37(8):844.

Vernot, B. and Akey, J. M. (2014). Resurrecting surviving Neandertal lineages from modern human genomes. *Science*, page 1245938.

Wade, C., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., Lear, T., Adelson, D., Bailey, E., Bellone, R., et al. (2009). Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science*, 326(5954):865–867.

Wagner, G. P., Kin, K., and Lynch, V. J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in biosciences*, 131(4):281–285.

Zhang, Y., Liang, J., Li, Y., Xuan, C., Wang, F., Wang, D., Shi, L., Zhang, D., and Shang, Y. (2010). CCCTC-binding factor acts upstream of FOXA1 and demarcates the genomic response to estrogen. *Journal of Biological Chemistry*, 285(37):28604–28613.