Improved reference genome for the domestic horse increases assembly contiguity and composition

Theodore S. Kalbfleisch, Edward S. Rice, Michael S. DePriest Jr., Brian P. Walenz, Matthew S. Hestand, Joris R. Vermeesch, Brendan L. O’Connell, Ian T. Fiddes, Alisa O. Vershinina, Nedda F. Saremi, Jessica L. Petersen, Carrie J. Finno, Rebecca R. Bellone, Molly E. McCue, Samantha A. Brooks, Ernest Bailey, Ludovic Orlando, Richard E. Green, Donald C. Miller, Douglas F. Antczak & James N. MacLeod

Recent advances in genomic sequencing technology and computational assembly methods have allowed scientists to improve reference genome assemblies in terms of contiguity and composition. EquCab2, a reference genome for the domestic horse, was released in 2007. Although of equal or better quality compared to other first-generation Sanger assemblies, it had many of the shortcomings common to them. 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. Here, we present EquCab3. The count of non-N bases in the incorporated chromosomes is improved from 2.33 Gb in EquCab2 to 2.41 Gb in EquCab3. Contiguity has also been improved nearly 40-fold with a contig N50 of 4.5 Mb 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 ~5 kya in central Asia, 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 reference genome assembly of the domestic horse, was released in 2007. 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 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, as well as their health and evolution. However, EquCab2 contains numerous gaps in scaffolds as well as sequences unassigned to chromosomes, and genomic DNA reshuffling and gene annotation 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, as well as a radiation hybrid map and FISH data. 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, with two scaffolds, is comprised of more than one scaffold. Finally, ~16× 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 the 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 Sanger, short-read, and long-read sequencing data as well as physical chromosome maps. This new reference is derived from the same female Thoroughbred horse, Twilight, that produced EquCab2, and includes her mitochondrial genome.

We used both previously published data and newly generated data to generate this reference assembly. The previously published datasets are comprised of the data used to construct EquCab2: Sanger sequencing data, BAC-end pairs, and a physical map containing radiation hybrid and FISH markers. For this assembly, we generated shotgun Illumina short reads, Chicago and Hi-C proximity ligation libraries, PacBio long reads, and 10× 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 datasets.

First, we used the high coverage (45×) and accuracy of Illumina short reads to generate so-called super reads with MaSuRCA. We assembled these super reads together with the long and accurate but lower coverage (6.8×) Sanger reads to create an initial assembly with Celera Assembler. We scaffolded this initial assembly with the long insert-size Chicago and Hi-C proximity ligation libraries using the HiRise scaffold. 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 than Sanger reads, using PBJelly. We phased the genome using 10× 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 mapability, 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 vs. 46 Mb (Table 1). At each phase of the assembly process (described in Methods section), there was 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 6.6 Mb, less than the final value of 46 Mb reported in Wade et al. 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 46 Mb to a chromosome length-limited 86 Mb. Further, the total number of gaps in the ordered chromosomes is decreased more than 90%, from 42,304 gaps comprising 55 Mb (2.2% of the genome) in EquCab2 to 3771 gaps comprising 9 Mb (0.34% of the genome) in EquCab3.

**Reading 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-sequencing (RNA-seq) and whole-genome shotgun (WGS) 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. The first phase of the equine FAANG effort was comprised of RNA-seq data from eight tissues. As shown in Fig. 1, for RNA-seq, unique mappings of the reads are increased by an average of 2.15% over EquCab2.

| Sequence composition | Contig N50 | Scaffold N50 |
|----------------------|-----------|--------------|
| EquCab3              |           |              |
| Sanger + MaSuRCA super reads | 1.2 Mb    | 6.6 Mb       |
| Sanger + MaSuRCA super reads + Chicago + Hi-C | 1.2 Mb    | 86 Mb        |
| Sanger + MaSuRCA super reads + Chicago + Hi-C + PacBio | 4.5 Mb | 86 Mb |
| EquCab2              |           |              |
| Sanger Fosmid + BAC + Radiation Hybrid Map data | 112 kb | 46 Mb |

For comparison, the contig and scaffold N50s for the final EquCab2 product are also shown. The final EquCab3 product (Sanger + MaSuRCA super reads + Chicago + Hi-C + PacBio) improved the contig N50 40-fold, and the Scaffold N50 was improved from a chromosome arm-limited 46 Mb to a chromosome length-limited 86 Mb.
and WGS paired reads improved by 0.44%. All mapping count details are available in Supplementary Data 1. The count of reads mapping in a proper pair, that is, 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. In addition, more reads in total (i.e., not limited to proper pairs) mapped to EquCab3 than EquCab2, as shown in Fig. 2.

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. EquCab3 also contains fewer gaps than EquCab2 due to the long-read gap-filling step, resulting in fewer dips in alignment coverage, shown in Fig. 3a. In addition, EquCab3 has more sequence assigned to chromosomes, giving reads more total sequence to map to, also demonstrated by Fig. 3a from the length increase in chr31 from EquCab2 to EquCab3. Finally, EquCab3 improves the characterization of GC-rich regions.

The total GC content of EquCab3 is roughly equivalent to that of EquCab2 (both near 41.6%). However, the guanine-cytosine (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. We compared mapping statistics between EquCab3 and EquCab2 for 13 previously sequenced ancient horses (Supplementary Data 2). A paired Wilcoxon’s 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 radiation hybrid map.** We used a radiation hybrid map of the horse genome to assign scaffolds to chromosomes. EquCab3 agrees with the radiation hybrid map more often than EquCab2. Of the 4103 markers on the physical map, 2982 map to EquCab2, while 3039 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 Fig. 3b–e.

Of the 395 misoriented marker pairs on EquCab3, 352 are oriented the same way on both EquCab2 and EquCab3, but 94 are misoriented.

**Fig. 1** Percentages of RNA-seq reads from eight tissues from two horses (designated 683610 and 686521) and genomic reads mapping to EquCab2 vs. EquCab3. We used sequence data from the Functional Annotation of Animal Genomes (FAANG) project 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.

**Fig. 2** Number of reads from the Functional Annotation of Animal Genomes (FAANG) project WGS dataset mapping to EquCab2 and EquCab3. Significantly more reads map only to EquCab3 than only to EquCab2 (one-tailed two-sample binomial test $p < 2.2 \times 10^{-16}$).
differently on the map. Twenty-one of the remaining 39 marker pairs that disagree with EquCab2 agree with EquCab3, while 18 do not have both pairs mapping to EquCab3. 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 radiation hybrid 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 seven marker pairs agreeing with EquCab2 but not EquCab3. Given that the radiation hybrid 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\textsuperscript{26} and the mammalian
universal ortholog set. Out of 4104 mammalian universal orthologs, BUSCO found 4092 (99.7%) as complete orthologs in EquCab3 with 5 fragmented and 7 missing, compared to 4064 (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\(^2\). 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 disjointed 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 more orthologs of genes in the other genomes in EquCab3 (Fig. 4a), fewer predicted genes split between contigs in EquCab3 (Fig. 4b), and a better distribution of gene coverage in EquCab3 (Fig. 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 10× 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 10× reads and the longranger\(^2\) 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 four Thoroughbreds, the two FAANG horses, and data from two other Thoroughbreds.

---

**Fig. 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 better in EquCab3 than in EquCab2.
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 radiation hybrid map and FISH data for the horse than was EquCab2. Going forward, the lenses 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 also long reads and proximity ligation data. All equine data produced by any of these technologies should be well served going forward. Illumina short reads, currently the most common data types for genetic and genomic studies, for two Thoroughbreds not related to Twilight or each other have been demonstrated to map to the new reference at an average rate of 99.06% for RNA-seq and 98.87% for WGS libraries, an increase of 2.15% and 0.44%, respectively. 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. Sample collection and DNA extraction: A single reference horse was used for this study. Her care and collection of the blood sample for genomic DNA isolation are covered under the Cornell University Institutional Animal Care and Use Committee (IACUC) Protocol 1986–0216, which is approved from 21 January 2016 through 21 January 2019. Venous jugular whole-blood samples were collected from Twilight into evacuated tubes containing heparin. High molecular weight DNA was extracted from whole blood using the Qiagen Mag-Attract HMW DNA Kit (Qiagen, Valencia, CA, USA) following the manufacturer’s protocol. After isolation, DNA size and quality were determined using a Femto Pulse Automated Pulsed-Field Instrument (Advanced Analytical, Ankeny, IA, USA) at the Cornell University’s DNA Core Lab.

Sanger data: The sequence data set comprised of 30,111,484 Sanger reads for the Thoroughbred mare Twilight, produced for and used to build EquCab2, were downloaded from the NCBI Trace Archive using the NCBI Trace Archive program to trace the data from the NCBI website with the query string query page size 30,000 page_number 223 binary SPECIES_CODE = EQUUS CABELLUS and CENTER_NAME = BI and CENTER_PROJECT = G386. In compliance with Trace Archive rules, 1004 individual shell scripts were executed to download the maximum 30,000 records per search request. This is described in more detail in ‘Methods’.

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 500 bp (range 300–650 bp) was constructed from 2 µg of Twilight’s genomic DNA after sonication with a Covaris M220 (Covaris, MA, USA) with the Hyper Library Preparation Kit from Kapa Biosystems (Roche) with no PCR amplification. The DNA library with adapters was loaded onto a 2% agarose gel and fragments 450 to 550 bp in length were cut from the gel and recovered with the QiaQuick Gel Extraction Kit (Qiagen, CA, USA). The size-selected library was quantified with Qubit (ThermoFisher) and run on an Agilent bioanalyzer DNA high-sensitivity chip (Agilent, Santa Clara, CA, USA) to confirm the presence of DNA fragments of the expected size range. It was further quantitated by qPCR on a Bio-Rad CFX, Connect Real-Time System (Bio-Rad Laboratories, Inc., Hercules, CA, USA) 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 300 nt in length. The fragment size distribution was measured by running the MiSeq dataset. The mode of the distribution was 504 bp with an average insert size of 544.18 bp. The library was subsequently sequenced on two lanes of the HiSeq2500 for 151 cycles from each end using a TrueSeq Rapid SBS Sequencing Kit v1. The fastq read files were generated with the bcl2fastq v1.8.4 Conversion Software (Illumina, San Diego, CA, USA) to the FASTQ format.

PacBio: Ten micrograms of high molecular weight genomic DNA from Twilight was sheared with gTUBEs (Covaris) in an Eppendorf® 5424 centrifuge at 4800 rpm for 2 × 60 s. A single PacBio library was prepared from this following PacBio’s protocol P/1N 100–286-000-07 (20 kb Template Preparation Using BluePippin™ Size-Selection System) with PacBio DNA Template Prep Kit. For the size selection, the sample was run on a 0.75% BluePippin cassette (ref: PAC20KB) using the predefined “0.75% DF Marker S1 high-pass 6–10 kb vs3” program and a cutoff of 10–50 kb. The library was sequenced on 88 SMRT cells on a PacBio RSII using DNA/Polynucleotide Binding Kit P6 and DNA Sequencing Kit 4.0 (v2) sequencing reagents, magbead loading, and stage start. All SMRT cells were run through PacBio’s SMRT Portal v2.3.0 pipeline _rs.1 with default settings except for minimum subread and polymerase read lengths of 1 kb. In addition, reads of insert were generated using the RS_ReadsOfInsert pipeline with a minimum insert read length set to 1 kb. Reads of insert had a mean of four passes and a depth of 11,785.

From the total initial read count of 5,934,426, we were able to create circular consensus (ccs) reads totaling 371,943 reads. The remainder of the reads were used to generate an error corrected subreads file using canu33 (version 1.0) consisting of 5,270,838 reads. These two datasets were used in the PBjelly reads described below.

CHICago library: We generated a CHICago library as previously described34 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, was resuspended and described by Lieberman et al.35 We then immobilized the chromatin on SPRI beads as described by Deng et al.36 We restriction digested the DNA with DpnII labeled ends with biotinylated dCTP, ligated ends, and reversed crosses. The sample was prepared for sequencing using the NEN 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 Inviotrogen C1 streptavidin beads, and then washed to remove non-biotinylated DNA fragments.

10x Genomics library: Twilight’s genomic DNA was size selected for fragments >40 kb on a BluePippin instrument (Sage Sciences, Beverly, MA, USA) and Illumina sequencing libraries were constructed using the 10x Genomics Chromium Controller instrument with their Genome Reagents Kit v2 chemistry (10x Genomics, Pleasanton, CA, USA) 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, 2 × 151 bp run (Illumina, San Diego, CA, USA). The data were analyzed and assembled using the 10x Genomics Supernova version 1.15.1 pipelines.

Assembly generation. MaSuRCA: We assembled all short reads from the Illumina genomic library into super reads using MaSuRCA37 version 3.1.3. The total length of sequences assembled into super reads was 4.7 Gb (approximately 2x coverage of the genome). These super read sequences had a contig N50 of 1734 nucleotides.

CelerA Assembler: The CelerA Assembler38 was downloaded from http://wgs-assembler.sourceforge.net/wiki/index.php?title=CelerA (version 8.2, downloaded on 2015-07-18). From the total number of reads, we 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 the small fragment mapping with default parameters, with the CHICago and Hi-C libraries as input libraries39.

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 parameters40. 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 Twilight41 to the assemblies using wma with default parameters41, assessing concordance with the physical map, looking for split genes predicted by the CAT42,43,44, aligning coding sequences of any genes in the region to the assemblies using gmap with default parameters45, and examining heatmaps of long-range read pairs mapping to the assembly generated by the HiRise and lonagranger pipelines46.

PBjelly: We filled gaps in the manually corrected HiRise scaffolds from the paired-end step using PacBio error corrected subreads and circular consensus sequences as input to the PBjelly (version PBjelly2.15.0/2.46b) 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 to assign scaffolds to chromosomes. We aligned each physical marker’s short tandem sequence primers to the assembly using bwa fastmap 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 positions.

Mitochondrial assembly: Illumina data were adapter trimmed using SeqPrep27. A subset of 24 million randomly selected Illumina reads were created using seqtk sample. The subsets Illumina reads were used as input into an iterative assembler (mia version 1.0.2)28, using the horse mitochondrial sequence (NC_001640.1) as the reference with a slope of 1.75 and intercept of 100. Sanger data were used to guide the correct mapping. We repeated the alignment and copy number steps of the control region80. Sanger reads were trimmed using Figaro (version 1.05.11) and aligned into the mitochondrial assembly using bwa (version 0.7.12)42. Alignments were manually inspected by eye using IGV (version 2.4)43. Sanger reads that aligned to the control region were extracted, visualized by eye, and compared to the initial mitochondrial assembly. Only Sanger reads unaligned both sides of the control region 8-mer repeats and was used to update the number of 8-mer repeats in the mitochondrial assembly. The updated mitochondrial assembly sequence was used as the reference sequence for an assembly with mia, using 40 million randomly selected Illumina reads, with a slope of 180 and intercept of 100. The assembly was filtered such that sites not having at least 10-fold coverage and 90% consensus among bases were changed to N. The final assembly had an average coverage of 40x, and 6Ns. The new Twilight mitochondrial sequence has been deposited into Genbank (accession number MH386816).

Quality control and assessment. Read mapping: Short-read sequence data generated in the initial phase of the equine FAANG project was mapped to both EquCab2 and EquCab3 for data comparison and mapping fractions. The sequence data can be found at the Sequence Read Archive. Both WGS sequence (40x) and RNA-seq (average coverage 20x) were aligned to the control region to assign scaffolds to chromosomes. We aligned each physical

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_002869295.1. The mitochondrial sequence has been deposited into Genbank (accession number MH386816). 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 longrange pipeline in a CyVerse Data Commons repository at https://doi.org/10.7946/P2034B5.

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. Osteoarthr. Cartil. 19, 435–438 (2011).
5. Schafer, R. J. et al. Developing a 670k genotyping array to tag ~2M SNPs across 24 horse breeds. BMC Genom. 18, 365 (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. Pleistocene 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 a deletion in myosin Va responsible for Lavender Foal Syndrome. PLoS Genet. 6, e1001999 (2010).
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, e1001999 (2010).
11. Stanger, E. A. et al. Host genetic influence on papillomavirus-induced tumors in the horse. Int. J. Cancer 139, 794–796 (2012).
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**, 342–352 (2017).
14. Gaunitz, C. et al. Ancient genomes reveal the ancestry of domestic and Przewalski’s horses. *Science* [https://doi.org/10.1126/science.aao3297](https://doi.org/10.1126/science.aao3297) (2018).
15. Schubert, M. et al. Prehistoric genomes reveal the genetic foundation and cost of horse domestication. *Proc. Natl Acad. Sci. USA* **111**, E5661–E5669 (2014).
16. Librado, P. Tracking the origins of Yakutian horses and the genetic basis for their fast adaptation to arctic environments. *Proc. Natl Acad. Sci. USA* **112**, E6889–E6897 (2015).
17. Librado, P. et al. Ancient genomic changes associated with domestication of the horse. *Science* **356**, 442–445 (2017).
18. Rebollo-Mendez, J. et al. Comparison of the equine reference sequence with its same-source derived 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. *Cytogetgen. Genome Res.* **122**, 28–36 (2008).
21. Zimin, A. V. *et al.* The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669–2677 (2013).
22. Miller, J. R. et al. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* **24**, 2818–2824 (2008).
23. Putnam, N. H. *et al.* Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).
24. Miler, J. R. et al. Revisiting the Full Spectrum of Human Genome Variation using Linked Reads. *BioRxiv*, 230946. [https://doi.org/10.1101/230946](https://doi.org/10.1101/230946) (2018).
25. Burns, E. N. et al. Generation of an Equine Biobank to be used for functional annotation of animal genomes project. *Anim. Genet.* [https://doi.org/10.1111/age.12717](https://doi.org/10.1111/age.12717) (2018).
26. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. J. *et al.* Tracking origins of Yakutian horses and the genetic basis for their fast adaptation to arctic environments. *Cytogenet. Genome Res.* **122**, 28–36 (2008).
27. Fiddes, I. T. et al. Comparative Annotation Toolkit (CAT)—simultaneous clade and personal genome annotation. *Genome Res.* [https://doi.org/10.1101/gr.233460.117](https://doi.org/10.1101/gr.233460.117) (2018).
28. Berlin, K. *et al.* Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.* **33**, 623–630 (2015).
29. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2012).
30. Deng, X. et al. Bipartite structure of the inactive mouse X chromosome. *Genome Biol.* **16**, 152 (2015).
31. Myers, E. W. *et al.* A whole-genome assembly of Drosophila. *Science* **287**, 2196–2204 (2000).
32. Koren, S. et al. Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biol.* **14**, R101 (2013).
33. Kurtz, S. et al. Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
34. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, arXiv [Q-bio. GN] 1303, 3997v2 (2013).
35. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
36. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
37. Eizenga, J. SeqPrep2: tool for stripping adaptors and/or merging paired reads with overlap into single reads. *GitHub* [https://github.com/eizenga/SeqPrep2](https://github.com/eizenga/SeqPrep2) (2011).
38. Li, H. Toolkit for processing sequences in FASTA/Q formats. *GitHub* [https://github.com/fli3/seqtk](https://github.com/fli3/seqtk) (2012).
39. Green, R. E. *et al.* A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* **134**, 416–426 (2008).
40. 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).
41. White, J. R., Roberts, M., Yorke, J. A. & Pop, M. Figurine: a novel statistical method for vector sequence removal. *Bioinformatics* **24**, 462–467 (2008).
42. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
43. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
44. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011).
45. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
46. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
47. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
48. Kalbfleisch, T. S. EqCab3 single nucleotide error correction. *GitHub* [https://github.com/kalbfleisch/eqcab3/singleNucleotideErrorCorrection](https://github.com/kalbfleisch/eqcab3/singleNucleotideErrorCorrection) (2018).
49. Morgulis, A., Geritz, E. M., Schäffer, A. & Aparawala, R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J. Comput. Biol.* **13**, 1028–1040 (2006).
50. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenome sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
51. Robinson, E. R. & Edwards, R. Quality control and preprocessing of metagenomic databases. *Bioinformatics* **27**, 863–864 (2011).
52. Kalbfleisch, T. S. EqCab3 intermediate assemblies. *CyVerse Data Commons* [https://doi.org/10.7946/P20348](https://doi.org/10.7946/P20348) (2018).

**Acknowledgements**

This work was supported by Morris Animal Foundation Grant D1SEQ-019 and the NRRP58 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) (1K12OD015134 and L40 TR001136). The FAANG data (C.J.F., J.L.P., R.R.B.) was generated with the support of the Grayson Jockey Club Foundation. This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. 681605 - PEGASUS). 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 Wen Mei for providing horses and compilation of Illumina libraries. Support for C.J.F. and E.S.R. was provided by the horse, *Equus caballus*: extensive heteroplasmy of the control region. *Gene* **148**, 357–362 (1994).

**Author contributions**

T.S.K., E.S.R., and B.P.W. performed the assembly. J.N.M. designed and led the project. 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 Wen Mei for providing horses and compilation of Illumina libraries. Support for C.J.F. and E.S.R. was provided by the horse, *Equus caballus*: extensive heteroplasmy of the control region. *Gene* **148**, 357–362 (1994).

**Competing interests**

I.T.F. is an employee of 10× Genomics, Inc. R.E.G. is a co-founder and scientific adviser of Dovetail Genomics, LLC. The other authors declare no competing interests.

**Reprints and permission information**

Reprints and permission information is available online at [http://nrpg.nature.com/reprintsandpermissions/](http://nrpg.nature.com/reprintsandpermissions/).

**Publisher’s note**

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.