Investigation of neocentromere 19 variation in *Equus asinus*

A thesis submitted

by

Tyler Medina

to

The Discipline of Bioinformatics,
School of Mathematics, Statistics & Applied Mathematics
National University *of* Ireland, Galway

in partial fulfilment of the requirements for the degree of

M.Sc. in Biomedical Genomics

August 10th 2018

Thesis Supervisors: Prof. Kevin F. Sullivan and Dr. Aaron Golden

Declaration of Academic Honesty

I, Tyler Medina, declare that this thesis, titled "Exploration of neocentromere 19 variation in *Equus Asinus*," submitted to the Discipline of Bioinformatics, School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway in partial fulfilment of the requirements for the degree of M.Sc. in Biomedical Genomics, is entirely my own work.

I have acknowledged all main sources of help. I agree freely that the library may lend or copy this thesis upon request.

Tyler Medina
August 10, 2018

Acknowledgements

My sincerest thanks and gratitude to Professor Kevin Sullivan for introducing me to some of the most thought-provoking things I've ever had to ponder. By some twist of fate, two Gauchos ended up in the same tiny part of the world to talk about centromeres, donkeys, and cheap tacos, and I couldn't be more grateful for that.

A special thanks to Dr. Aaron Golden, who was always quick to provide guidance and the eternal wisdom of proverbs. An example of the true modern Renaissance man, no one else can walk around a double helix in the morning, drop by Mars for lunch, and use a space laser to fill his kettle in time for tea. The amount of knowledge passed down to me in this short year from Dr. Pilib Ó Broin, Prof. Cathal Seoighe, and Dr. Derek Morris is, frankly, staggering. Thank you all so much for all that you do, and for all that you've given me.

I owe the NUIG Biomedical Genomics Class of 2018 a huge thanks. I honestly believe that I could not have done it without the support we've all given each other, and I couldn't have asked for a better group of people to be locked in a room with all year.

To my family, near and far, thanks for always being there for me when I needed it most. Your support means the world to me, and makes the world feel like a small place in the best possible way.

And above all else, thanks Bríd. I really could not have done this without you being there every step of the way. This one's for you.

Abstract

Highly repetitive satellite DNA at the centromere precludes the ability to sequence and assemble eukaryotic centromere domains using standard short-read sequencing techniques. However, evolutionary new centromeres formed on unique DNA sequences are more readily sequenced and assembled, making further genomic inquiries into the centromere possible. The domestic donkey, *Equus asinus*, possesses as many as 16 such satellite-free centromeres, making it an ideal model for centromeric sequence investigation.

Despite the centromere being a highly conserved structure among all eukaryotes, the sequence underlying the centromere, as well as centromeric position, can vary without disrupting centromere function. As shown in data of chromosome 19 sequence from four individual donkey isolates, neocentromere 19 appears to exhibit variation and polymorphism in the population, with evidence for the ongoing evolution of new repetitive elements. Here, an identification and comparison of the centromeric sequence structure between centromere 19 of four donkey isolates is performed, using a combination of quantitative and alignment methods to propose two models of centromeric structure that exist within the population.

Contents

List of Figures

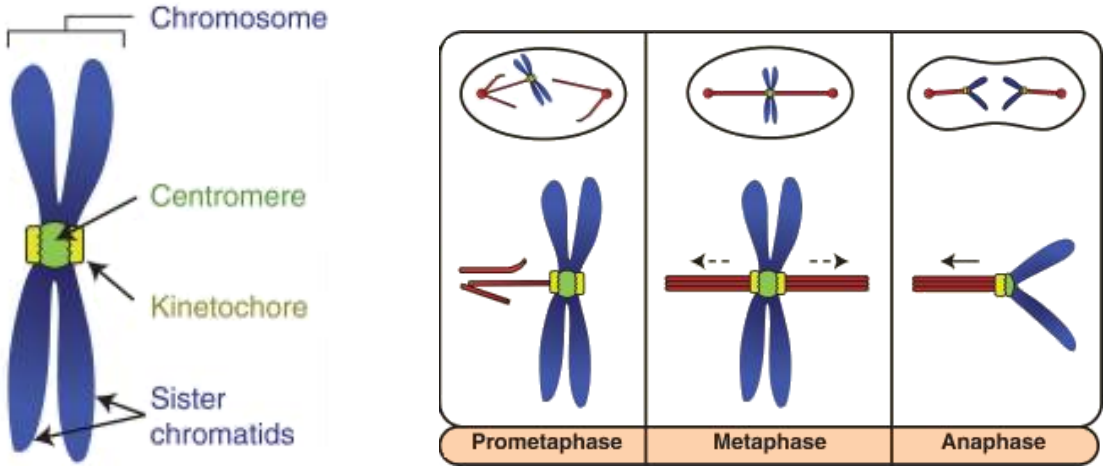
List of Tables

Chapter 1

Introduction

1.1 Function of the centromere

During chromatin replication in all eukaryotic genomes, replicated sister chromatids are held together at a specific locus in each chromatid known as the centromere. The centromere is a highly condensed region of chromatin that binds to the cohesin protein complex, which joins and holds together sister chromatids from S phase through the early phases of mitosis and meiosis.[1] This action forms the structure known as the primary constriction and the characteristic "X" shape commonly seen in eukaryotic metacentric human chromosome karyotypes (Fig. 1.1).



(a) General structure of a chromosome

(b) The centromere during stages of chromosome segregation. The inset at top shows a zoomed out image of the cell at each stage.

Figure 1.1: A centromere in context. Adapted from Westhorpe and Straight (2014).

The centromere is a vital component of the cell cycle due to its role in chromosome segregation. During the cell cycle, a protein structure known as the kinetochore forms on the outer surface of each cohesin-bound sister chromatid's centromere.[2] In the course of metaphase, spindle fiber microtubules, originating from the centrioles at the poles of the cell, bond to the kinetochore of each sister chromatid. These microtubules pull at the centromeres from either side, causing each pair of chromatids to come to rest at the metaphase plate due to the competing tension.[3] Kinetochore proteins that have not yet been bound by microtubules continuously delay the cell cycle at what is known as the spindle assembly checkpoint, inhibiting the onset of anaphase. Once all kinetochores have been properly bound by spindle fiber microtubules and all pairs of chromatids have aligned at the metaphase plate, the mitotic spindle assembly checkpoint is satisfied, and entry into anaphase begins. At this point, the cohesin protein complex is cleaved, and the connection between sister centromeres separates as the spindle microtubules pull the two sister chromatids to opposite poles of the dividing cell.[4] Centromere or kinetochore formation failure leads to problems in chromosome segregation, resulting in aneuploidies and genome instability, and is a common occurrence in cancer cells.

1.2 Centromeric chromatin

Nucleosomes are the basic sub-unit of eukaryotic chromatin. Typical nucleosomes found throughout eukaryotic genomes are composed of an octamer of histone proteins, consisting of two each of histone proteins H2A, H2B, H3, and H4.[5] This core of histone proteins is wrapped by a sequence of approximately 147 bp of DNA to form one nucleosome.[6] In condensed chromatin, DNA is present in a "beads on a string" formation, with nucleosome "beads" separated by stretches of non-nucleosome-associated linker DNA.

One of the hallmarks of centromeric chromatin is the presence of specific nucleosome variants which contain H3-like Centromere Protein A (CENP-A) in the place of histone H3.[7,8] The CENP-A histone fold domain shows approximately 62% conservation with H3, but contains a CENP-A targeting domain that allows it to localize to the centromere, as shown by H3 nucleosomes modified to contain this targeting domain.[9] CENP-A nucleosomes also differ from H3 nucleosomes in that they only wrap 133 bp of DNA, instead of 147 bp.[10] Although typical H3 nucleosomes are still the majority of nucleosomes throughout the centromere, the location of a small number of CENP-A nucleosomes is the deciding factor of centromere location and formation.[11] In humans, this is approximately 200 CENP-A nucleosomes per centromere, making up only 2 - 4% of total nucleosomes in the centromere.[10,12] CENP-A functions by recruiting a complex of 16

other centromere proteins comprising the constitutive centromere-associated network (CCAN) to the centromere site. Specifically, CCAN proteins CENP-C and CENP-N act to specifically identify CENP-A nucleosomes. Other CCAN proteins subsequently bind to the kinetochore, completing kinetochore localization to the centromere.[13] CENP-C is a vital component of kinetochore formation, as experiments inducing CENP-C binding to locations away from CENP-A nucleosomes cause kinetochore formation at these non-CENP-A locations.[14] Under normal circumstances, CENP-C and CENP-A co-purify, and CENP-C is only present at active centromeres, making both CENP-A and CENP-C protein markers of the centromere.[15,16] Similar results have been shown by ChIP-seq and ChIP-chip experiments, including in the domestic horse.[17,18]

1.3 DNA sequence of the centromere

Nearly all eukaryotic centromeres are associated with large arrays of tandemly repeating satellite DNA sequences (Fig. 1.2). In humans, these are composed of 171 bp monomers arranged in tandem higher-order repeat (HOR) arrays known as alpha satellites, and span up to 5 Mbp of DNA at centromeres.[19,20,21,22] These structures present a major challenge to current Next-Generation Sequencing assembly methods due to the difficulty of accurately assembling highly repetitive sequences into linear, contiguous scaffolds.[23,24] Overall, this prevents large scale, base-pair resolution inquiry into the structure of centromeric regions. Although long-read sequencing technologies such as PacBio SMRT and synthetic long-reads are promising, low sequencing quality and low throughput still remain obstacles for satellite areas. However, extremely long reads generated by Oxford Nanopore minIon supplemented with short-reads have recently been used successfully to build centromeric contiguous scaffolds of human chromosome 9.[25]

Despite the nearly ubiquitous presence of tandem repeat satellite DNA in centromeres, evidence shows that satellite DNA is not actually required for centromere formation.[26,27] Although the centromere protein CENP-B binds the specific alpha satellite sequence known as the CENP-B box, CENP-B is not essential for centromere formation, and CENP-A nucleosomes do not bind to specific DNA sequences or motifs in this manner.[28] Furthermore, while HOR satellite DNA arrays show high conservation within species, centromeric satellite sequences between species are not conserved, and evolve rapidly.[29]

1.4 Neocentromeres

Centromeres are known to be able to relocate to new chromosomal positions, forming "neocentromeres". The spontaneous formation of neocentromeres in non-

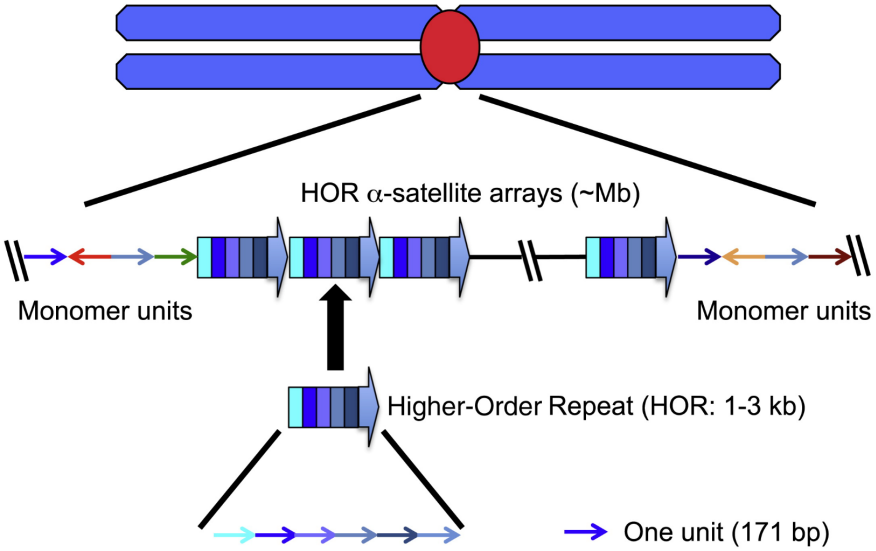


Figure 1.2: Higher-order repeat organization of alpha satellite DNA. Adapted from Fukagawa and Earnshaw (2014).

satellite DNA has been well-documented in humans since its first observation in 1993 in a clinically-discovered marker chromosome, formed through the cyclical rearrangement of chromosome 10. In this case, a chromosome 10 cross-over event with itself produced a linear, "mardel10" chromosome. The resultant mardel10 chromosome was acentric, meaning that it did not contain any of chromosome 10's original centromere locus. Despite this, the mardel10 chromosome was "rescued" by the spontaneous formation of a novel, functioning neocentromere.[15,30]

Though neocentromere formation in humans is typically seen through chromosomal rearrangement events, centromeres also have the ability to reposition elsewhere on their chromosomes, forming neocentromeres on non-satellite DNA without chromosomal rearrangement events. As evidenced by the comparative locations of centromeres in primates,[31] as well as in many other mammalian species, centromeres have relocated through the process of evolution and speciation without chromosomal rearrangement, forming evolutionary new centromeres (ENC).[32] At these new loci, neocentromeres behave similarly to traditional centromeres, forming kinetochores and binding most centromere proteins in the same way, despite being located at very different, non-satellite sequences.[33] Over time, the sequence under ENCs begins to accumulate repetitive elements until they come to resemble typical centromeres, perhaps as a result of the centromere complex and function.[34,35] While new centromere regions accumulate satellite arrays, the repetitive nature of their former loci quickly degrades.[31]

While ENCs gain satellite repeats over time, the process may take millions

of years, resulting in some neocentromeres remaining satellite-free for some time. Three such centromeres have been identified in chickens and one has been identified in the orangutan.[36,37,38] In addition, members of the Equus genus have been shown to have large numbers of satellite-free neocentromeres.[39,40] These satellite-free neocentromeres present unique opportunities for studying the centromere, as they are able to be sequenced and assembled by short-read sequencing. They also allow for investigation into the purpose and formation of centromeric satellite DNA, which seems to be very evolutionarily conserved at centromeres, yet does not seem to be essential for centromere formation. This high conservation of the centromere structure itself, despite the rapid evolution and change of the underlying sequence, has been referred to as the centromere paradox, and contributes to the strong support for epigenetic determination of the centromere locus.[29] However, the fact that evolutionary new centromeres on unique DNA sequences subsequently develop typical patterns of satellite repeat arrays suggests that while satellite repeats are not necessary for centromere formation, they do provide some evolutionarily-conserved advantage or function to the centromere.[34,41]

The continued presence of any particular CENP-A nucleosome at its particular locus is determined by its location during DNA replication. As DNA is replicated, CENP-A nucleosomes are distributed evenly between the daughter strands. New CENP-A nucleosomes are deposited after mitosis in G2 by HJURP chaperonin proteins, which use the centromere targeting domain of CENP-A to recognize the correct location.[42] Because centromere positions are only propagated epigenetically by this CENP-A redistribution during DNA replication and deposition in G2, and are not held in place by sequence identity, it has been demonstrated that centromere positions can slowly drift or slide.[43,44,45] One theory suggests that long satellite regions effectively "capture" centromeres, containing them within a long stretch of non-coding DNA to prevent them from drifting into gene sequences where the chromatin structure of the centromere would mute transcription (Fig. 1.3).[23,37]

1.5 Satellite-free centromeres in *Equus*

Members of the rapidly and relatively recently evolved Equus genus, comprising horses, asses, and zebras, have particularly high numbers of satellite-free evolutionary new neocentromeres. The most recent common ancestor of the genus is estimated to have existed 2 - 4 million years ago, with current species appearing approximately 1 million years ago, showing relatively rapid speciation.[46,47] In this currently "young" genus, the well-annotated genome of the domestic horse, *E. caballus* or alternatively *E. ferus caballus*, contains one satellite-free centromere on chromosome 11,[18] and Piras et al. (2010) demonstrated the presence of additional

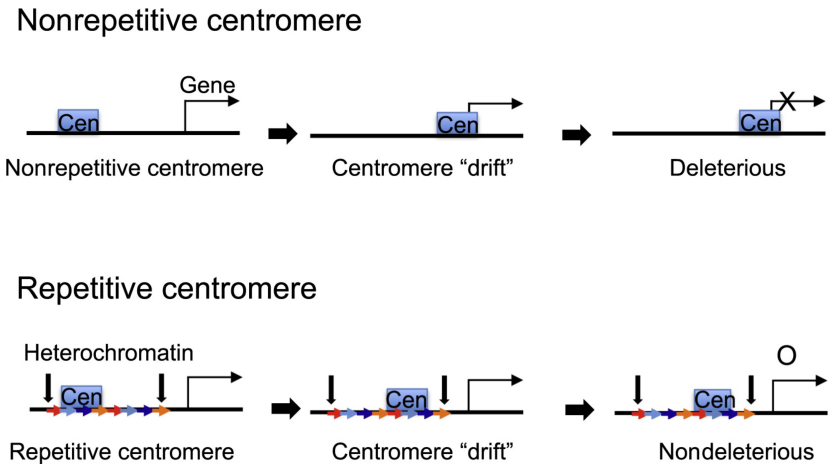


Figure 1.3: Model of satellite DNA preventing centromere drift into a gene. Adapted from Fukagawa and Earnshaw (2014).

satellite-free centromeres in other members of the genus.[40]

In their study, two satellite sequences were identified from the horse, which were not found in animals such as fruit flies, cows, mice, or humans, but were found in other equids. These satellite sequences were used as probes in fluorescent *in situ* hybridization (FISH) experiments to identify the location of said satellites on chromosomes from the horse, the domestic donkey *E. asinus* or alternatively *E. africanus asinus*, Grevy's zebra *E. grevyi*, and Burchelli's zebra *E. burchelli*. In addition, total genomic DNA FISH probes were used for each species to identify the presence of any other tandem repeat satellites besides the two horse sequences used. Satellite presence was identified by the differences in hybridization behavior between repetitive elements compared to single-copy regions. Results of these FISH experiments are shown in Figure 1.4. While centromeres of each chromosome were initially identified by karyotype observation of the primary constriction, these were later confirmed via anti-CENP-A FISH. The comparison of centromere locations with the presence or absence of satellite sequences revealed many centromeres without satellite sequences, including all nine evolutionary neocentromeres previously identified in the donkey, horse, and Burchelli's zebra.[39,48]. Furthermore, based on comparisons of centromere sequence structure between chromosomes and species in the genus, Piras et al. (2010) suggest a model for the evolution of satellite sequences at the centromere, shown in Figure 1.5.[40]

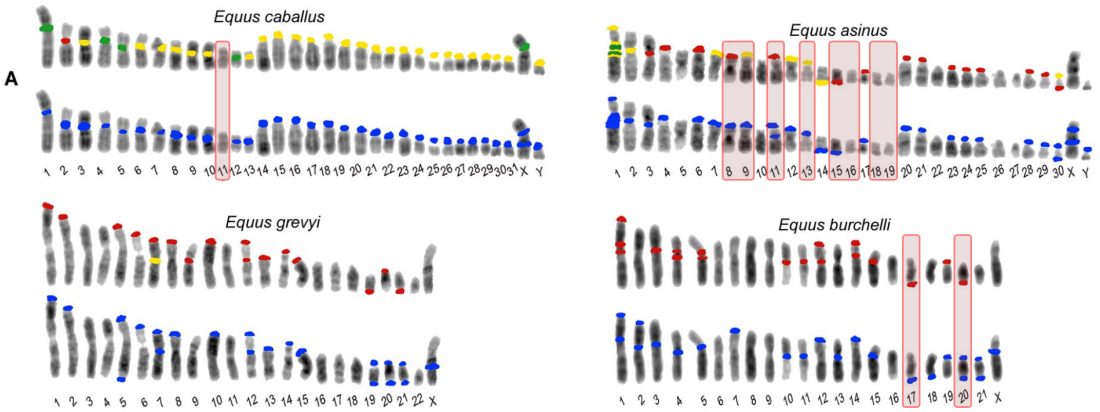


Figure 1.4: Satellite sequence FISH results in equid chromosomes. Top rows show hybridization to two different horse satellites in red and green, or yellow for hybridization to both. Bottom rows show whole genomic DNA self-hybridizations identified as satellite regions. Chromosomes with previously identified neocentromeres that are shown to be satellite-free are boxed in red. Adapted from Piras et al. (2010).

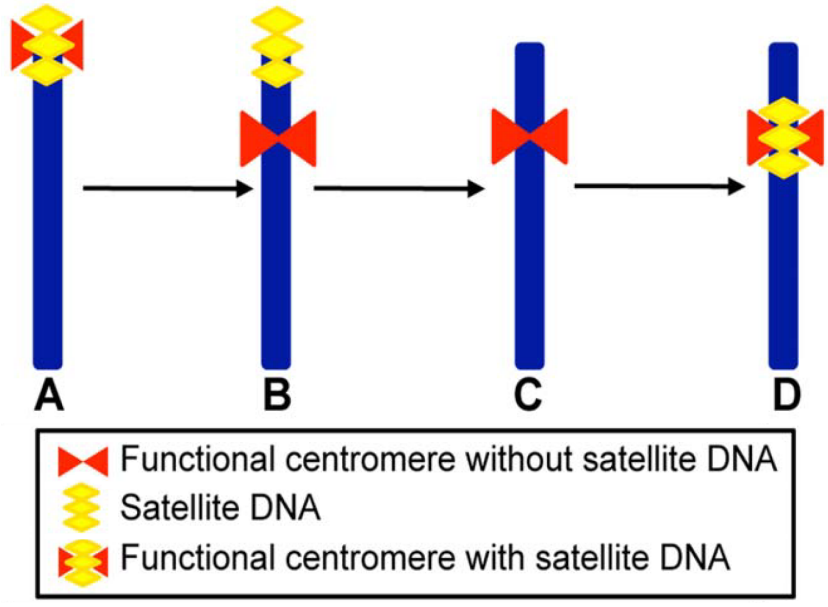


Figure 1.5: Model of neocentromere evolution. Adapted from Piras et al. (2010).

1.6 Sequencing *E. asinus*

The domestic donkey's large number of satellite-free centromeres make it an excellent model organism for further inquiries into centromeric structure and function. While donkey genome sequencing has been performed, no chromosome-level fully assembled donkey genome is currently available.

Whole genome shotgun sequencing experiments have been performed on two individual *E. asinus* individuals. In 2015, Huang et al. of the Inner Mongolian Agricultural University in Hohhot, China published the genome of a local male Guanzhong breed donkey named Maral Har.[49] This genome was sequenced to an average 42.4X depth using a combination of Illumina MiSeq 251 bp paired-end reads and Illumina HiSeq 100 bp long-insert mate-pair reads, and assembled using Roche's Newbler software. The resultant 2.36 Gbp genome consists of 2,166 scaffolds with an N50 value of 3.8 Mbp. Huang et al. also investigated the relationship between the donkey and horse genomes through synteny analysis, and found similar results to previous investigations regarding centromere repositioning and satellite-free neocentromeres in donkey chromosomes.

In 2018, a second whole genome shotgun sequencing project was published by Renaud et al. of the University of Copenhagen, Denmark on a local male zoo donkey named Willy, improving on a previous draft assembly of the same isolate produced by the same group.[50,51] Chicago library preparation was used, in which very large DNA fragments are linked and tagged via in vitro nucleosome reconstitution and formaldehyde fixation, similar in theory to Hi-C methods.[52] Using Illumina HiSeq paired-end reads and complementary PCR-free reads, this donkey genome was sequenced at 61.2X depth and assembled into 9,021 scaffolds with an N50 value of 15.4 Mbp. A comparison of scaffold sizes for the current Willy genome, the previous Willy genome, and the Maral Har genome is shown in Figure 1.6.

In addition to whole genome sequencing, sequencing targetting donkey centromeres has also been performed. Chromatin immunoprecipitation sequencing, or ChIP-seq, is an experimental technique used to target and sequence only regions of DNA that bind to particular proteins, such as transcription factors or histones. Nergadze et al. (2018) published the findings of ChIP-seq experiments targeting CENP-A nucleosomes in two male donkey isolates: an Italian donkey named Asino Nuovo, and a donkey named Blackjack from the Cornell University stables in New York.[45,53] While ChIP-seq data targeting centromeric sequences would normally yield very repetitive reads that would be very difficult to map back to a reference of equally repetitive satellite DNA, the satellite-free nature of many donkey centromeres enables the production of mappable reads for these loci. However, the lack of a chromosome-level assembled donkey genome prevents easy mapping of ChIP-seq reads. Alignments were instead performed against the pub-

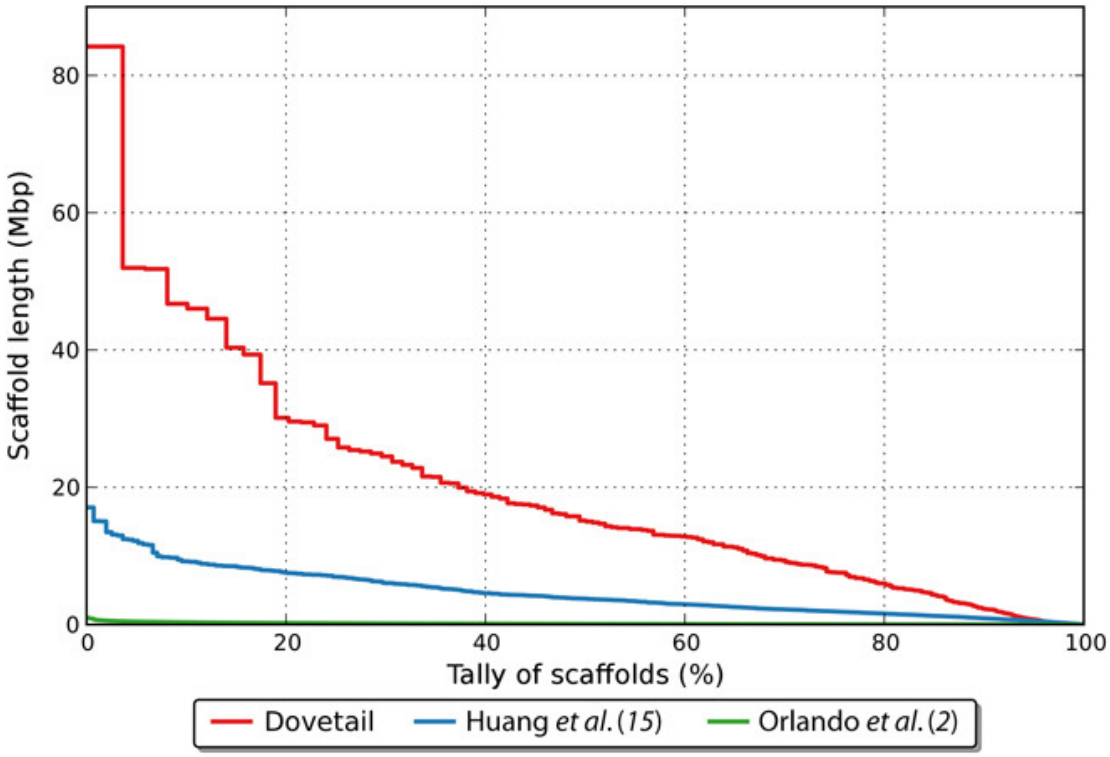


Figure 1.6: Comparison of donkey assembly scaffold sizes. The current Willy assembly using Dovetail Genomics' Chicago assembly method is shown at top in red, the Maral Har assembly method is shown center in blue, and the original Willy genome draft assembly with very low N50 is shown at bottom in green. Adapted from Renaud et al. (2018)

lished and well-annotated horse reference genome EquCab2, which has established chromosomal orthology and over 98% genomic identity with the donkey.[18,51,54,55] Alignment to the horse confirmed the cytogenetically proposed presence of repositioned, satellite-free neocentromeres in donkeys by sequence-level examination. While sequence rearrangements at donkey centromeric loci relative to their homologous non-centromeric horse sequences were observed, in some cases causing gaps in alignment profiles, of particular note were the presence of both bimodal and very narrow "spike" alignment profiles to the horse genome. Bimodal alignments imply the presence of two regions associated with the bulk of CENP-A nucleosome-associated DNA sequences. Because the centromere is a continuous region of the chromosome defined by a distribution of CENP-A nucleosomes, it is unlikely that two separate yet very nearby sequences would each contain this distribution independently. Instead, an epiallelic model emerges, in which individual chromatids in the population may each have a somewhat shifted centromeric locus, which has been documented in horse chromosome 11.[43]

The narrow spike distributions seen, such as in Asino Nuovo centromeres 8, 9, 16, 18, and 19, are of particular note due to the fact that the observed satellite-free centromere on horse chromosome 11 contains CENP-A nucleosomes distributed over 100 kbp anywhere within a 500 kbp range.[43] The bounds of these narrow spike distributions are much narrower than the range seen in horse chromosome 11, suggesting that these narrow regions would not be able to physically accommodate a normal compliment or distribution of CENP-A nucleosomes. In addition to being much narrower than the Gaussian-like and bimodal distributions seen in other donkey chromosome alignments, the spike alignments also exhibited much taller peaks, suggesting both a narrower and more concentrated alignment of reads while not differing from the overall numbers of reads aligning to the position. While satellite sequences were not detected in the 16 centromeres aligned by Nergadze et al., the difference in pileup at these spike alignment sites immediately provides evidence for the presence of some kind of repetitive element. Detection of reads spanning the junctions between tandem repeat elements provided further evidence of this. In efforts to quantify the number of repeats present at these spike alignments, qPCR was performed with two primers against the repeat unit locus in Asino Nuovo, Blackjack, a Blackjack-horse hybrid offspring (mule), and a horse (Fig. 1.7). As depicted, the horse sequence at these centromeres shows no evidence of tandem repeats, while Asino Nuovo and Blackjack, as well as the Blackjack mule offspring, show strong evidence of repeats. In all three shown mule centromeres, copy number appears to be approximately half of that seen in Blackjack, which is expected due to the hybrid nature of the mule genome, essentially consisting of both a haploid donkey genome and a haploid horse genome. In addition to the variation observed between the donkeys, horse, and mule, donkey centromere 19 shows intraspecific variation between Blackjack and Asino Nuovo, with Blackjack

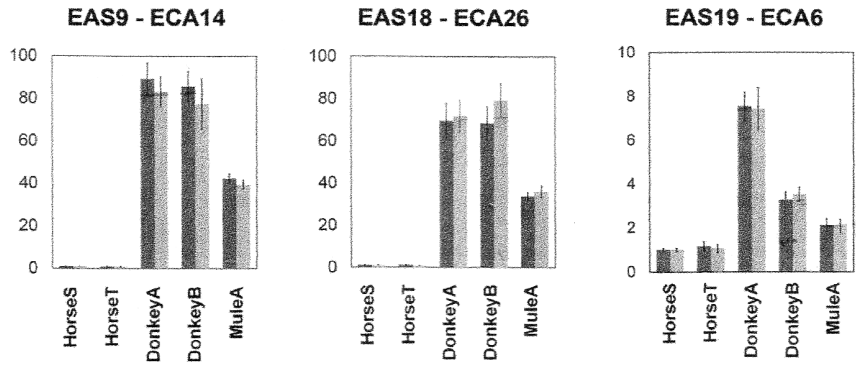


Figure 1.7: qPCR results quantifying tandem repeats. Two different probes are shown in dark and light gray, each targeting a portion of the repetitive sequence observed in the alignment of each animal isolate. Mule genomes were obtained from Blackjack-horse crosses. Adapted from Nergadze et al. (2018)

showing approximately half of the copy number of Asino Nuovo. Overall, these results demonstrate the simultaneous lack of satellite DNA sequences at these centromeres and the presence of non-satellite DNA amplification at a subset of these centromeres. In addition to these analytical results, Nergadze et al. also published model assemblies of Asino Nuovo and Blackjack centromeric sequences using a combination of ChIP-seq reads, their associated input reads, and homologous horse sequence. Due to the tandem repeats shown in five of the centromeres, their assemblies do not reflect the true and full sequence of the centromere, but represent sequence models. In the case of Asino Nuovo centromere 19, the published sequence contains two tandem repeat units to represent the n repeats truly present.

Further work investigating centromeres in *E. asinus* was carried out in the doctoral thesis of Dr. J. G. W. McCarter.[45] Using immortalized cells from Asino Nuovo, this experiment investigated the native distribution of CENP-A nucleosomes through micrococcal nuclease (MN)-digested ChIP-seq, rather than through formaldehyde crosslinking and subsequent sonication. Fragments were size selected via electrophoresis to produce three separate sequencing runs: two sets with fragment size in the 150 to 200 bp range, equivalent to the size of DNA wrapped in a single CENP-A nucleosome, and one set with fragment size of 450 to 500 bp, equivalent to the size of three consecutive CENP-A nucleosomes. While size selection was ultimately not entirely successful, these ChIP-seq experiments still provide centromeric sequencing coverage. In addition to the CENP-A ChIP-seq runs performed, two CENP-C ChIP-seq runs were also performed, creating two additional centromere-targeted sequencing sets. All alignments performed for these data sets supported previous donkey centromere alignments.

1.7 Project scope

For this project, donkey centromere 19 was chosen for investigation as an example of a potentially polymorphic satellite-free centromere. The aim of this project is to identify, explore, and where possible quantify the variation seen in *Equus asinus* centromere 19 of four individual donkey isolates, building on evidence previously reported. A combination of approaches are considered, including multiple sequence alignment, ChIP-seq read alignment, k-mer matching, read depth analysis, and CNV calling to create a body of evidence supported by multiple facets of the observed variation. In addition to data published by Huang et al. (2015) on the Maral Har isolate, Renaud et al. (2018) on the Willy isolate, and Nergadze et al. (2018) on the Asino Nuovo and Blackjack isolates, as of yet unpublished data on Asino Nuovo was provided by Dr. McCarter and the Sullivan research group at the Centre for Chromosome Biology of the National University of Ireland, Galway, in conjunction with the Giulotto research group of the University of Pavia, Italy.

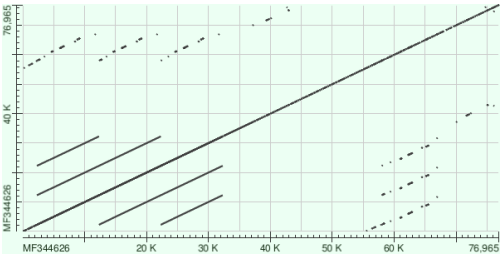
Chapter 2

Methods

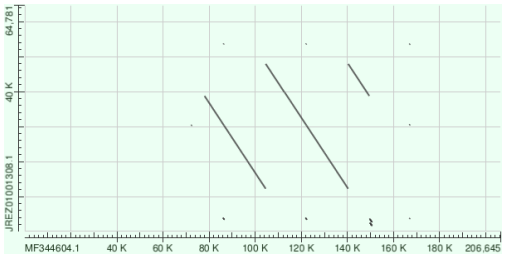
The dry lab methods for producing the results of this project follow. Example scripts and explanations of parameters used are shown where appropriate. A table of software versions and hardware used, as well as scripts created for this project can be found in Appendices A and E.

2.1 Homologous sequence alignment

For homologous sequence comparisons and exploration, both NCBI's Basic Local Alignment Search Tool (BLAST) and the UCSC Genome Browser's BLAST-Like Alignment Tool (BLAT) were used.[56,57] The BLAST nucleotide alignment webtool takes as input a query DNA sequence and a subject DNA sequence, either uploaded by the user or accessed through the NCBI repositories. The default "megablast" algorithm option was used for most BLAST queries, which optimizes results for highly similar sequences. Because all comparisons performed were either between *E. asinus* individuals or between *E. asinus* and *E. caballus*, this narrow-scope algorithm was preferred over the broad "blastn" algorithm. In addition to a list of homologous alignments ranked by score and E-value, BLAST also outputs a dot-plot comparison of the query and reference sequences (Fig. 2.1). All default BLAST algorithm values were used, with the exception of self BLAST searches for large megabase-scale scaffolds, for which the max target sequences were reduced from 100 to 10, word size was increased to 64, and the short queries option was removed. This was done to remove heavy noise in the form of many short self matches.



(a) Blackjack centromere 9, self-comparison



(b) Maral Har 1308 vs. Asino Nuovo centromere 19

Figure 2.1: BLAST-produced dot-plots

Although BLAST provides base-by-base sequence alignments, BLAT was used preferentially to streamline pipelining through its command line version. Default parameters were used for all alignments, producing PSL files containing information on numbers of matches, mismatches, gaps, and matching sequence coordinates per alignment. An example BLAT command is shown here:

```
# BLAT
blat reference.fasta query.fasta output.psl
```

The Broad Institute's Integrative Genomics Viewer (IGV) was used to visualize all output PSL files.[58,59] An example PSL file IGV visualization is shown in Figure 2.2.

2.2 Centromere Identification

In donkey individuals Asino Nuovo and Blackjack, centromeric regions were identified by ChIP-seq targeting of CENP-A and subsequently assembled.[53] Because scaffold assemblies for donkey individuals Willy and Maral Har were obtained via whole genome sequencing,[49,50] centromeric regions were not identified experimentally. Scaffolds containing likely centromere locations in Willy and Maral Har were identified by their sequence homology to Asino Nuovo and Blackjack centromeric sequences.

However, because Asino Nuovo and Blackjack show evidence of structural variation in the centromeric region, the known single-copy horse homologous sequence was used instead to establish likely centromeric loci in Willy and Maral Har. This horse locus homologous to the donkey centromere was first identified through homology to Asino Nuovo and Blackjack centromeres. The horse sequence at this locus was then extended past the bounds of the donkey centromere loci by 10% of the overall length at both the 5' and 3' ends. In this way, a horse sequence

probe sequence was created containing both a 5' and 3' sequence of unique DNA not known to be associated with the donkey centromere in either Asino Nuovo or Blackjack, which served as a non-centromeric control during alignment. This horse sequence was then aligned to both the Maral Har and Willy genome assemblies.

2.3 Quality control of Willy scaffolds

Alignments of centromeric sequences to the published Willy scaffolds revealed many small runs of undetermined bases in the identified Willy scaffold loci. To determine whether this was a local phenomenon potentially caused by centromeric activity, or a global phenomenon occurring throughout the Willy genome's assembly, five random horse sequences were generated using BEDTools random and getfasta, and were then aligned to the Willy genome to compare the quality of alignments.[60]

```
# GENERATION OF RANDOM HORSE LOCUS SEQUENCES
# Produces 5 (-n) random loci of length 75055 (-l) from the equCab2
    genome (-g), outputting them (>) in BED format
bedtools random -l 75055 -n 5 -g equCab2.fa > random_coordinates.bed
# Extracts sequences from the input equCab2 fasta (-fi), using a list
    of coordinates in BED format (-bed) and outputting the results in
    fasta format (-fo)
bedtools getfasta -fi equCab2.fa -bed random_coordinates.bed -fo
    random_sequences.fasta
```

Sequence size was selected to match that of the Asino Nuovo repeat unit including its deletion, plus an additional 10% of flanking sequence from either end for a total length of 75,055, equivalent to the probe length used to identify the homologous Willy centromere scaffolds.

2.4 Sequencing read alignment

Because the donkey genome has not been assembled to the level of chromosomes, donkey sequencing reads were aligned to the horse reference genome EquCab2, as the horse genome is well annotated and closely related to the donkey.[18,51] Data from three individual experiments was used: Asino Nuovo micrococcal nuclease-digested ChIP-seq and input reads; Asino Nuovo and Blackjack formaldehyde cross-linked ChIP-seq and input reads; and whole genome sequencing reads from Willy. Raw read data for the remaining donkey isolate, Maral Har, was unavailable. See Appendix B for a complete listing of the reads used, including accession numbers where appropriate.

2.4.1 Quality control

Quality control was performed on all FASTQ files using FastQC.[61] Quality reports and graphs were then generated using MultiQC, which aggregates FastQC output files to create a summary in HTML format.[62]

2.4.2 Alignment

Alignment was performed using both Bowtie2 and the Burrows-Wheeler Alignment Tool (BWA).[63,64] Bowtie2 was chosen to recapitulate the alignments used to generate the Asino Nuovo and Blackjack centromeric scaffolds[53], while BWA was chosen to compare and confirm alignment results.

For Bowtie2, "bowtie2-build" was used to index the FASTA format equCab2 genome. After indexing the horse reference genome, each run of paired-end sequences in FASTQ format was aligned using default parameters. In the example below, Bowtie2 is called to align the two paired FASTQ files using the horse index files, and outputs the result in sequence alignment map (SAM) format.[65]

```
# BOWTIE2 ALIGNMENT
# Creates genome index files for the reference genome (-f) with a
    chosen prefix ("horse" in this case)
bowtie2-build -f equCab2.fa horse
# Uses the index files with specified prefix (-x) to align the paired
    reads (-1 and -2), producing a SAM file (-S)
bowtie2 -x horse -1 reads_1.fastq -2 reads_2.fastq -S alignment.sam
```

BWA usage is similar to Bowtie2 and begins by using "bwa index" to index the equCab2 genome. The Burrows-Wheeler Transform Smith-Waterman (BWTSW) algorithm was chosen for indexing over the Induced Sorting (IS) linear-time algorithm, as the IS algorithm is not well-suited to the large size of mammalian genomes. Following indexing, bwa mem was used for paired-end alignment.

```
# BWA ALIGNMENT
# Indexes the equCab2 genome using the bwtsw algorithm (-a)
bwa index -a bwtsw equCab2.fa

# Uses 8 computing threads (-t) and the equCab2.fa index files to align
    paired-end reads from two input files, outputting (>) to a SAM file
bwa mem -t 8 equCab2.fa reads_1.fastq reads_2.fastq > alignment.sam
```

2.4.3 Post-processing

After alignment, the resultant SAM files from both Bowtie2 and BWA were processed in the same way using SAMtools.[65] SAMtools was used to first convert the human-readable SAM files to binary BAM files to facilitate faster computation time. After this, each BAM file was sorted by alignment name by using SAMtools sort, before using SAMtools fixmates to ensure that mate-pairs were correctly flagged in each file. Reads were then re-sorted by alignment coordinate to identify any reads aligning to the exact same coordinates. These reads are assumed to be PCR duplicates of the same DNA fragments, and are thus removed using SAMtool rmdup. Finally, SAMtools is used to index each processed BAM file for later use and create summary mapping statistics. An example of these steps is shown here:

```
# POST-PROCESSING
# View converts SAM input (-S) to BAM output (-b)
samtools view -S -b alignment.sam > alignment.bam

# Sort reorganizes the BAM file by name (-n), and automatically appends
   ".bam" to the output file
samtools sort -n alignment.bam alignment_name_sorted

# Fixmate ensures that read-pairs possess the correct flags to indicate
   their mates
samtools fixmate alignment_name_sorted.bam alignment_fix.bam

# Sort again reorganizes the BAM file, this time by alignment
   coordinate and chromosome (default)
samtools sort alignment_fix.bam alignment_chr_sorted

# Rmdup removes any duplicate reads
samtools rmdup alignment_chr_sorted.bam alignment_rmdup.bam

# Index provides an index file for later use, such as IGV visualization
samtools index alignment_rmdup.bam

# Flagstat provides a summary of how well the FASTQ files were mapped
samtools flagstat alignment_rmdup.bam > stats.txt
```

Note that in the case of the micrococcal nuclease-digested Asino Nuovo reads, each sample was split and run over multiple lanes, producing multiple pairs of FASTQ files per sample. Because alignment and mate-pairing are only dependent per run, and not per sample, alignment and mate flagging were performed on

each run independently. However, because PCR duplicates are created during library preparation, duplicates may exist distributed across individual runs of the same sample. To ensure that all PCR duplicates are removed, these multi-lane samples were combined using SAMtools merge prior to removing duplicates, as shown here:

```
# MERGING
# Merge combines multiple BAM files, which in this case are currently
   sorted by name (-n)
samtools merge -n combined.bam lane1.bam lane2.bam
```

2.4.4 Visualization

All processed BAM files were visualized in IGV using the BAM indexes generated by SAMtools. An example visualization of a region of interest is shown in Figure 2.2.

2.5 Coverage depth calculations

Sequencing depth at the centromeric region of the non-ChIP-seq aligned donkey reads was calculated to estimate copy number of each individual. SAMtools bedcov outputs the sum of the depth of sequencing of each base in the specified region. For each input or whole genome sequencing run alignment, bedcov was used to calculate the total depth of 6 regions, using coordinates of horse chromosome 6: all of chromosome 6, the 5' end of chromosome 6 to the 5' end of the suspected repeat region, the 5' region of the suspected repeat, the suspected deleted region in the center of the repeat region, the 3' region of the suspected repeat, and from the 3' end of the suspected repeat to the 3' end of the chromosome (Fig. 2.3).

To compare the differences in coverage between the repeat regions and the non-repeat regions, the total depth of repeat regions were combined, and the total depth of non-repeat regions were combined. These two sums were then normalized by the lengths of their respective sequences, producing an average depth per base of the two areas. The ratio between the repeat and non-repeat average depth per base was then compared to estimate copy number.

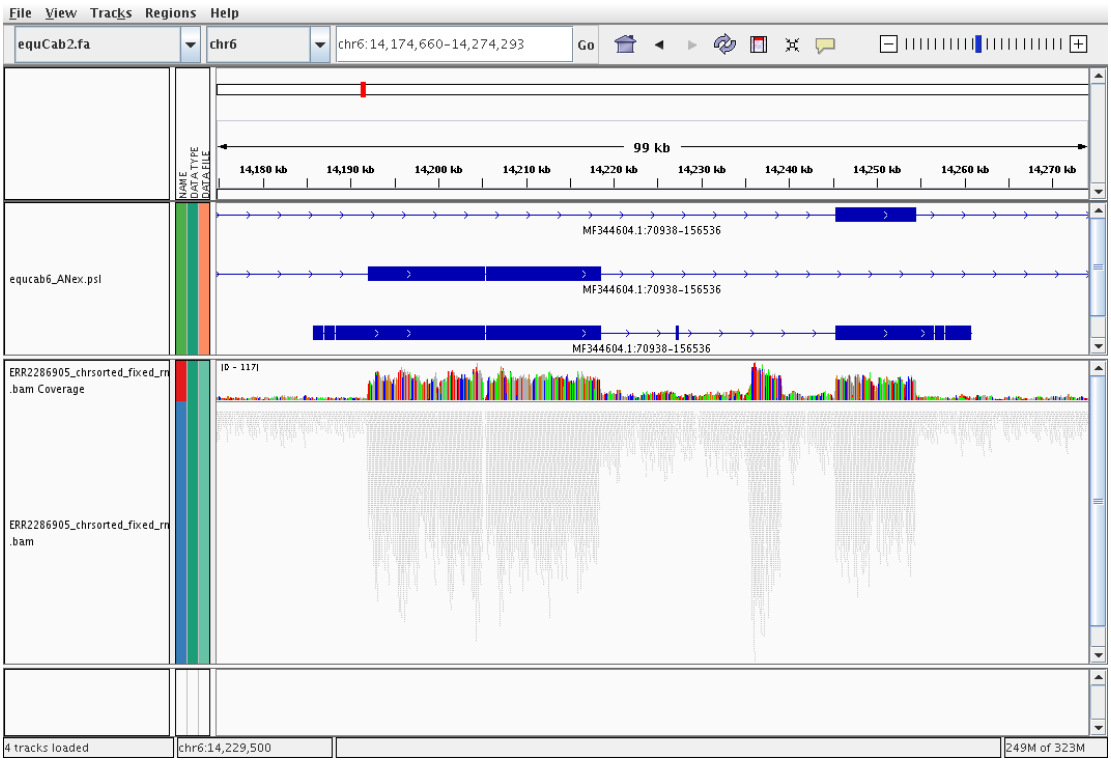


Figure 2.2: An example of IGV Visualization. The current reference genome file in use is indicated in the top left (equCab2.fa), followed by the current chromosome and viewing window (chr6:14,174,660-14,274,293). Below, the relative location on the current chromosome is highlighted by a red box. Between the relative chromosome location and the alignment tracks in the main window, the scale shows approximate window size (99kb) and chromosome coordinates. In the main window, two separate lanes show different alignments. The top alignment, named "equcab6_ANex.psl" at left, is an example of a BLAT produced PSL file visualized in IGV. Each thick blue line represents a run of largely continuous aligned sequence, with gaps represented as thin lines. Each individual continuous blue line indicates an alternate alignment, and arrows indicate the direction of the query sequence relative to the reference genome. The bottom lane, with names beginning with "ERR2286905", is an example of a sequencing read alignment in IGV. The top track displays normalized read depth coverage per locus, while the bottom track displays actual read alignments. Colored loci in the coverage track represent read nucleotides that differ from the reference genome.

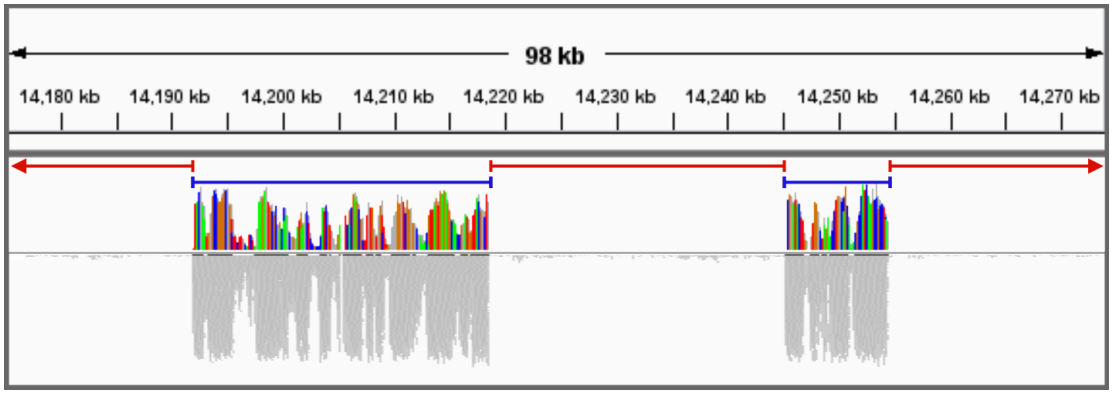


Figure 2.3: Regions for calculating sequencing depth. Regions in red are considered single-copy regions, while regions in blue are considered repeat regions.

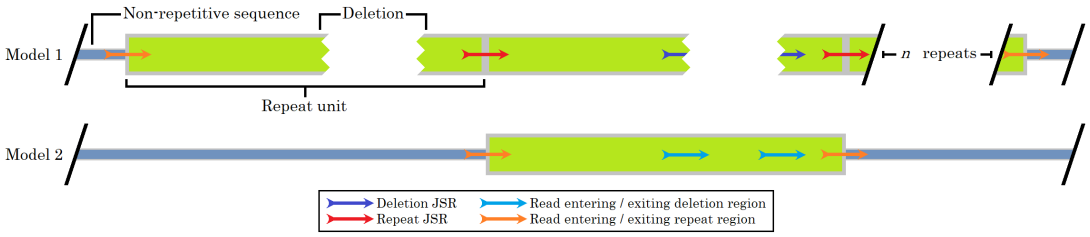


Figure 2.4: K-mer template loci. Four types of k-mer template sequences were created from loci representing the depicted transitions from one region to another.

2.6 Identification of junction-spanning reads by k-mers

The repeat regions at donkey centromere 19 likely cause difficulties and errors in scaffold assembly. Because of this, the actual structure of centromere 19 in the donkey likely differs from published scaffolds. To build evidence of large-scale deletions and repeats, individual reads containing sequences representative of deletions (deletion junction spanning reads, or deletion JSR) or repeats (repeat junction spanning reads, or repeat JSR) were identified by k-mer matching (Fig. 2.4). For this purpose, 50 bp reference sequences were constructed that represent deletion JSR and repeat JSR, with 25 bp from either side of the junction site. Sequences representing non-repeat junction sites and non-deletion junction sites were also constructed as controls.

BBDuk (Bestus Bioinformaticus Decontamination Using K-mers), part of the Joint Genome Institute's BBTools software suite, can be used to filter reads based on k-mer matching. Using BBDuk with each FASTQ file from the 3 donkeys, any reads containing a match to the constructed reference sequences were filtered

and saved in new FASTQ files. To ensure that filtered reads contained sequence representing the specific junction sites, the minimum matching k-mer length was set to 26 out of the 50 bp length, such that any continuous 26mer in the 50 bp constructed sequence must cross over the junction site.

BBDuk usage is shown here. Note that paired-usage was used, such that if any individual read contains a matching k-mer, its mate is also filtered. In addition, the default "maskmiddle" is left at its default value, causing each generated k-mer to have a wildcard at its middle base to increase sensitivity. Because each 26 k-mer generated from a 50 bp sequence has significant overlap with other k-mers, each wildcard would still be covered by a non-wildcard base in another sequence. Furthermore, because only the middle bases in each k-mer is masked, the 13 bp at each end of the 50 bp reference would not be queried by any wildcards, and so would not cause a loss in specificity for junction-spanning reads.

```
# BBDUK KMER MATCHING
# Uses a sliding window to generate length 26 (k) k-mers from a
    reference sequence (ref). Any bases in the reference k-mer template
    that are masked as lowercase are unmasked in the generated k-mers
    (touppercase). By default, reads and mates from the input FASTQ
    files (in, in2) containing at least 1 matching k-mer are considered
    a match. Matching reads and mates are each output to new paired
    FASTQ files (outm, outm2). Basic matching rates and statistics are
    output to a text file (stats) containing 5 columns of information
    (cols).
bbduk.sh \
in=reads.fastq in2=mates.fastq \
ref=kmer_reference.fasta \
touppercase=t \
outm=matching_reads.fastq \
outm2=matching_mates.fastq \
stats=matching.stats \
cols=5 \
k=26
```

A k-mer length of 26 allows for any individual read crossing the junction site to be matched and filtered. However, this also includes reads that may contain 99 bp upstream of the junction, and only 1 bp downstream of the junction, or vice versa. This single-base cross would be considered low-confidence, due to the match being dependent on the quality of a single base. This was of particular concern due to the fact that no FASTQ files were trimmed prior to alignment, and prompted consideration of longer k-mers for higher specificity at the cost of lower sensitivity. However, because the probability of aligning to a particular 1 bp locus

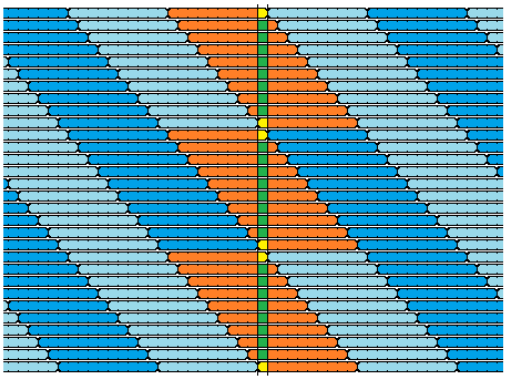


Figure 2.5: Location of an arbitrary locus within simulated reads. Above, individual reads are shown in blue or orange. Orange reads intersect an arbitrary green locus, while 5'-start or 3'-end bases that cross the locus are shown in yellow. In this model, 6 out of 30 length-10 reads cross the locus at their extreme ends, equal to a $2/length$ frequency.

in a genome via input or whole-genome sequencing is quite low, high sensitivity was required to detect the few reads that may be present at a particular locus in low-coverage or low-depth sequencing. In addition, in a hypothetical uniform distribution of simulated reads, an arbitrary position passes through the last base of a read with a frequency of $2/l$, where l is the length of the reads, equivalent to the probability of randomly and independently selecting 2 objects out of l possible objects. Therefore, assuming that reads are in large part randomly distributed without significant bias, with read lengths over 100 bp at least 98% of random reads that contain a particular arbitrary locus will contain the locus at some base closer to the center of the read, rather than at one of the extreme ends (Fig. 2.5). This indicates that a larger k-mer length to provide higher specificity was not necessary to identify reads containing junction sites, and that matching reads would likely contain enough overlap across the junction to be treated with confidence. Furthermore, if this were not the case, all matching reads could be further filtered anyway using a longer k-mer length after first using a shorter k-mer length, should higher specificity be required.

Ratios between deletion JSR, repeat JSR, and their control counterparts were calculated to observe any distribution patterns that would indicate centromere structure in each individual animal.

2.7 CNV calling

The software package CNVcaller was used to detect copy number variation between the aligned donkey sequences and chromosome 6 of the horse reference genome.[66] CNVcaller estimates copy number by comparing aberrant sequencing read depth to a reference with known copy number. In addition, CNVcaller compares multiple samples to detect candidate CNV loci.

Note that CNVcaller was built on Python 3.6 and has several Python package dependencies. As computationally-intensive steps for CNVcaller were run on an HPCC with Python 2, several modifications had to be made to the CNVcaller Python scripts to ensure backwards compatibility. While Python 3 was available on the HPCC, necessary Python 3 packages were not.

Before applying CNVcaller to detect CNVs in the donkey samples, a duplicated window record file was first created for horse reference genome chromosome 6. Chromosome 6 was first separated into short k-mer sequences using CNVcaller's Kmer_Generate script, as shown here. As the suspected CNV sequence was relatively large, and both input and whole-genome sequences contain relatively low-coverage, a window size of 1600 was selected, as recommended by the program for sub-10 X coverage data.

```
# KMER GENERATE
# Divides the reference genome into k-mers of designated length,
    storing them in FASTA format
python 0.1.Kmer_Generate.py chr6.fa 1600 chr6kmer.fa
```

Blasr was then used to align these short k-mer sequences back against the reference. Blasr's sawriter tool was used to build a reference suffix array for indexing purposes, prior to the alignment. Blasr was run with all CNVcaller recommended parameters, which primarily optimize alignment for speed, since the generated k-mers are exact matches to the genome.

```
# SA WRITER
# Creates a suffix array for the reference genome, blasr's required
    indexing format
sawriter chr6.fa

# BLASR
# Aligns the k-mers generated to the reference genome.
blasr chr6kmer.fa chr6.fa --sa chr6.fa.sa --out chr6kmer.aln -m 5
    --noSplitSubreads --minMatch 15 --maxMatch 20 --advanceHalf
    --advanceExactMatches 10 --fastMaxInterval --fastSDP
    --aggressiveIntervalCut --bestn10
```

The duplicated window file is then created using CNVcaller's Kmer_Link script, which establishes a baseline copy number for windows in horse chromosome 6.

```
# DUPLICATE WINDOW FILE
# Creates the prerequisite duplicate window file for CNVcaller
python 0.2.Kmer_Link.py chr6kmer.aln 1600 chr6window1600.link
```

Horse chromosome six was then indexed using CNVcaller's CNVRefrenceDB script. As stated, a window size of 1600 was chosen based on CNVcaller's recommendation. Optional GC content and gap content parameters were left at default settings.

```
# REFERENCE INDEXING FOR CNVCALLER
# References the reference genome file with the designated window file
perl CNVReferenceDB.pl -w 1600 chr6.fa
```

Because only chromosome 6 was used for CNV detection, alignments from the fully processed BAM files were filtered, producing BAM files containing only reads aligned to chromosome 6. Each BAM header was editted to reflect this by removing all other sequence (@SQ) lines, and a read group line (@RG) was added as well for CNVcaller compatibility. An example of this process is shown here:

```
# SAM TO BAM
# Converts BAM to SAM, including the header (-h), outputting (>) the
    contents to the designated file
samtools view -h ERR2286905_rmdup.bam > ERR2286905_rmdup.sam

# EXTRACT SAM HEADER
# Returns the header only (-H) of a SAM file (-S), outputting the
    contents to a new text file
samtools view -SH ERR2286905_rmdup.sam > ERR2286905.head

# ADD HD LINE TO NEW HEADER
# Returns the first (-n 1) line of the file, outputting the contents to
    a new head file
head -n 1 ERR2286905.head > ERR2286905_new.head

# ADD SEQUENCE LINE CHR6 TO NEW HEADER
Searches and return lines starting with (^) @SQ, appending the results
    to the new head file
grep "^@SQ.SN:chr6" ERR2286905.head >> ERR2286905_new.head

# ADD NEW RG LINE TO NEW HEADER
# Appends the specified read group string to the new head file
```

```
echo "@RG      ID:ERR2286905      SM:ERR2286905" >> ERR2286905_new.head

# ADD PG LINE TO NEW HEADER
Searches and return lines starting with (^) @PG, appending the results
    to the new head file
grep "^@PG" ERR2286905.head >> ERR2286905_new.head

# EXTRACT MATCHES TO CHR6 FROM SAM FILE
# Returns lines (print $0) where the 3rd column ($3) is (==) "chr6",
    outputting the matching lines to a new text file
awk '$3 == "chr6" { print $0 }' ERR2286905_rmdup.sam >
    ERR2286905.chr6matches

# ADD NEW HEADER AND MATCHES TO NEW SAM FILE
# Outputs the full contents of the head file and the chr6 matches to a
    new SAM file
cat ERR2286905_new.head ERR2286905.chr6matches > ERR2286905_6.sam

# NEW BAM TO SAM
Converts the new input SAM (-S) file to a bam file (-b)
samtools view -Sb ERR2286905.sam > ERR2286905_6.bam
```

With all prerequisite files produced, CNV calling then began on the donkey sample input and whole-genome sequencing BAM files. Read depth was first calculated by CNVcaller's Individual.Process script. This creates various read depth count files, using the created duplicated window file to correct absolute copy number, and ultimately outputting a file containing normalized read depth per window.

Note that although required, the sex option is not used in this processing. This option is used by CNVcaller to correct copy number on the sex chromosomes, given that human males, for example, would be calculated as having half copy number compared to autosomes if this were not performed. Because copy number estimation was only performed on chromosome 6 here, this correction is not necessary, but the argument must still be passed to the script for the program to run.

```
# GENERATING READ DEPTH COUNTS
# Creates normalized read depth counts from an input BAM (-b) file with
    the specified sample group designation in the header (-h), using
    the specified duplicated window file (-d). The name of the sex
    chromosome is specified as well (-s).
bash Individual.Process.sh -b ERR2286905_6.bam -h ERR2286905 -d
    chr6window1600.link -s X
```

CNV detection then proceeds by using CNV.Discovery to compile these read depth files for comparison to one another, establishing candidate CNV regions from each sample and merging significantly correlated adjacent CNV regions. CNV candidate windows are filtered based on meeting either minimum presence in the population of samples, or minimum absolute number of individuals homozygous for the CNV. Limits for correlation coefficients between adjacent windows, CNV population frequency, and number of CNV homozygous individuals were chosen based on the provided sample size and CNVcaller recommendations.

Normalized read depth files to process are provided in a list. A list of samples to exclude from said list must also be provided, but is empty in this case.

```
# CNV CANDIDATE DISCOVERY
# Creates a list of candidate CNV regions for the list of samples
    provided (-l), excluding samples on the exclusion list (-e). CNV
    regions present in at least 10% of individuals in the population
    (-f) are included, as are CNV regions estimated to be homozygous in
    at least 2 individuals (-h). Adjacent CNV regions are merged if
    they meet a Pearson's correlation coefficient of at least 0.5 (-r).
    Results are output to primary and merged result files (-p and -m).
bash CNV.Discovery.sh -l ListOfSampleFiles.txt -e
    ExcludedSampleFiles.txt -f 0.1 -h 2 -r 0.5 -p primaryCNVR -m
    mergedCNVR
```

In the final step, each CNV region determined from the population of samples is genotyped for each individual sample using CNVcaller's Genotype script, creating an output VCF file.

```
# GENOTYPE CNVR RESULTS
# Creates a VCF file named "AllDonk.vcf" containing genotype
    information for each individual at each CNV region in the
    mergedCNVR file
python Genotype.py --cnvfile mergedCNVR --outprefix AllDonk
```

An excerpt of this VCF file was then visualized in IGV after modification to follow VCF standards.

Chapter 3

Results

3.1 BLAT homology of centromeric sequences

Homologous sequence alignments were performed to locate homologous loci between the ChIP-seq identified donkey chromosome 19 sequence and the two published donkey genomes, as well as the reference horse genome. The most significant results show high homology between horse chromosome 6 near 14.2 Mbp; Asino Nuovo centromere 19 (AN19); Blackjack centromere 19 (BJ19); Willy scaffolds 6729, 7405, and 5647; and Maral Har scaffolds 350, 1308, and 133.* Table 3.1 shows the top alignments coordinates and query base coverage for Maral Har and Willy.

Scaffold	Match%	QStart	QStop	RSize	RStart	RStop
JREZ01000350	38.98	0	86079	2222255	2135204	2222255
JREZ01001308	17.05	86079	124199	64781	0	38743
JREZ01001308	9.19	129862	156252	64781	2510	64781
JREZ01000133	30.98	148625	217242	4507428	0	75974
PSZQ01006729	48.96	0	115282	14060657	0	114045
PSZQ01007405	2.59	115202	122754	7019	0	7019
PSZQ01005647	30.76	149282	217242	5280021	0	77194

Table 3.1: BLAT alignment statistics. Match% refers to the fraction of query bases matching the reference, QStart and QStop are query sequence alignment coordinates, RSize is the reference sequence size, and RStart and RStop are the reference sequence alignment coordinates.

*Note that scaffold numbers for Maral Har and Willy used here refer to the last digits of their accession numbers, and not necessarily to the names used within their respective genome

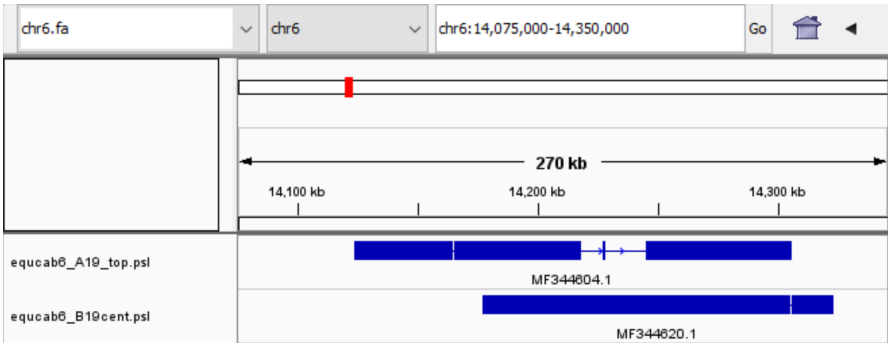


Figure 3.1: BLAT alignment of AN19 and BJ19 against horse chromosome 6

Figure 3.1 shows the top BLAT alignments for the full AN19 and BJ19 sequences against the horse reference genome. Both alignments occurred between 14.1 Mbp and 14.4 Mbp on horse chromosome 6. The AN19 alignment ran from start to end of the Asino Nuovo contig, and aligned with 63% matching bases and less than 1% mismatching bases, with gaps contributing 36% of the alignment length. While two other AN19 alignments were also produced by BLAT (not shown), these were very poor, diffuse alignments with only 15% and 6% matching bases. The BJ19 alignment resulted in 92% matching bases, 1% mismatched bases, and 7% gaps.

Figure 3.2 shows the top four BLAT alignments of this homologous region on horse chromosome 6 against Maral Har. Note that scaffold 1308 has been reversed to follow the same orientation as the query sequence. The horse sequence used did not align continuously to any one scaffold on Maral Har. Although over 1000 alignments were produced, the top three non-overlapping alignments to scaffolds 350, 1308, and 133 together contained the 5' and 3' ends of the query sequence, and combined had 87% matching bases and 1% mismatching bases. Of note is that the alignment to scaffold 350 covers the last base of the scaffold, two largely non-overlapping contiguous alignments to scaffold 1308 cover both ends of the scaffold, and alignment to scaffold 133 covers the first base of the scaffold, each with at least 1 kbp of highly contiguous sequence. In addition, alignments to scaffolds 350 and 1308 end and begin at the same base in the query sequence, as shown in Table 3.1.

Figure 3.3 shows the top three BLAT alignments of the same query horse sequence against the Willy scaffolds. Note that scaffolds 6729 and 7405 have been reversed to follow the same orientation as the query sequence. While there is a 1.2 kbp gap in scaffold 7405, the underlying sequence in the Willy scaffold

assemblies. This is particularly important to note for the Maral Har assembly. While both Maral Har scaffold accession numbers and scaffold names are numbered sequentially, scaffold names skip numbers, while accession numbers do not, sometimes resulting in slightly offset numbering.

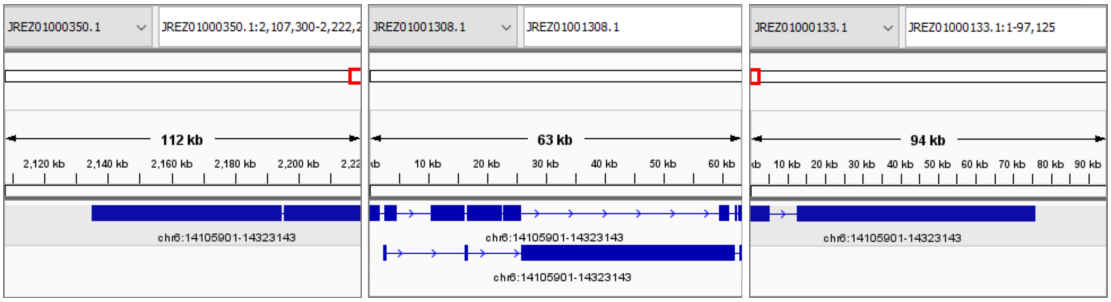


Figure 3.2: BLAT alignment of centromeric sequence against Maral Har, using the AN19 centromere homologous horse sequence as a probe.

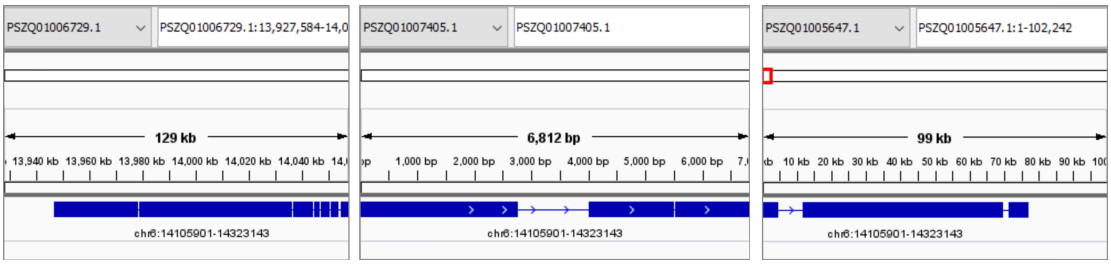
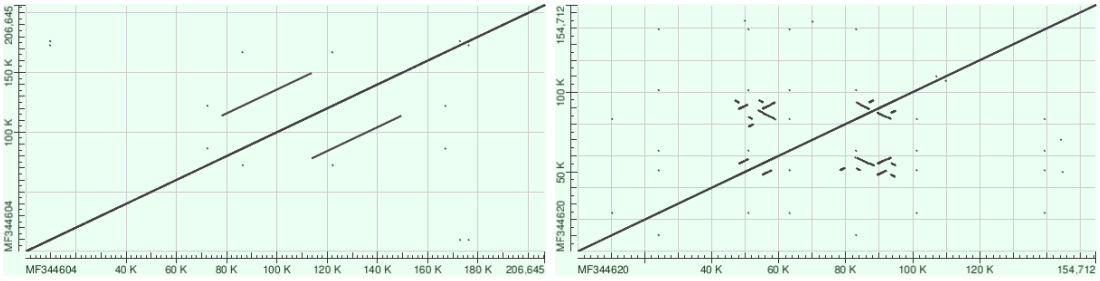


Figure 3.3: BLAT alignment of centromeric sequence against Willy, using the AN19 centromere homologous horse sequence as a probe.

is a run of undetermined bases. As in Maral Har, alignments were scattered over many scaffolds, generating over 1500 alignments. A pattern similar to the Maral Har alignment emerged, with 3 major scaffolds containing the bulk of the alignment. Although scaffolds 6729 and 7405 overlap by 80 bp, the alignment over these three scaffolds produces 82% matching bases and 1% mismatching bases in the query sequence, and again consists of alignments that cover the ends of their scaffolds. Although the alignment covers the 3' end of scaffold 7405 and the 5' end of scaffold 5647, no single large continuous alignment was found that covered the 24 kbp between the two scaffolds, though many smaller overlapping alignments from many scaffolds were found to cover portions of this region.

3.2 Self-similarity BLAST searches

BLAST multiple sequence alignment shows a significant stretch of self-similarity in AN19, present in two copies (Fig 3.4a). This region is a large repetitive element identified during the assembly of the AN19 centromeric region, and represents some n number of repeats in the Asino Nuovo sequence. The Blackjack centromeric region does not have this structure present in its assembly, as shown in its own self-similarity dot-plot (Fig. 3.4b). The repeat unit in AN19 is also visible



(a) AN19 **(b)** BJ19

Figure 3.4: BLAST self-similarity dot-plots of centromeric sequences

in BLAT self-searches of AN19. Figure 3.5 shows two possible matches to a region in AN19, each of which were then extracted and re-aligned, each resulting in the same two alignments.

Like Blackjack, the homologous assembled scaffolds of Maral Har and Willy do not show any large regions of self-similarity through BLAST dot-plot (Fig. 3.6).

3.3 BLAT homology of AN19 repeat unit

The relationship between the repeat unit evident in Asino Nuovo centromere 19 and the sequence of the other donkeys was explored by BLAT homology with the other donkey isolates. The horse sequence homologous to the Asino Nuovo repeat unit was used as a probe against all four donkey isolates, using an additional sequence of unique horse DNA flanking the homologous region to identify donkey sequences homologous to the same locus as the repeat unit. The resultant sequences from all four donkeys were each aligned against the horse to compare sequence and position. Results are shown in Figure 3.7. The extracted Asino Nuovo repeat unit aligned to the horse sequence is also shown at top for reference.

All alignments concur with the homologous full centromere region alignments in Figures 3.1 through 3.3. The Asino Nuovo alignment produces three separate alignments, the two shorter alignments together reflecting the bounds of the repeating unit in AN19.

Note that the alignment of Maral Har scaffold 1308 to the horse sequence at this locus does not produce one alignment with a gap, but rather two separate, non-overlapping alignments shown on the same line, with the upstream alignment showing the 3' end of the query scaffold, and the downstream alignment showing the 5' end of the query scaffold. A dot-plot comparison was produced between the Asino Nuovo repeat unit and Maral Har scaffold 1308, due to their similar, structurally different, alignments in BLAT. Results are shown in Figure 3.8, indicating similar sequences that are rearranged with respect to one another, as also shown

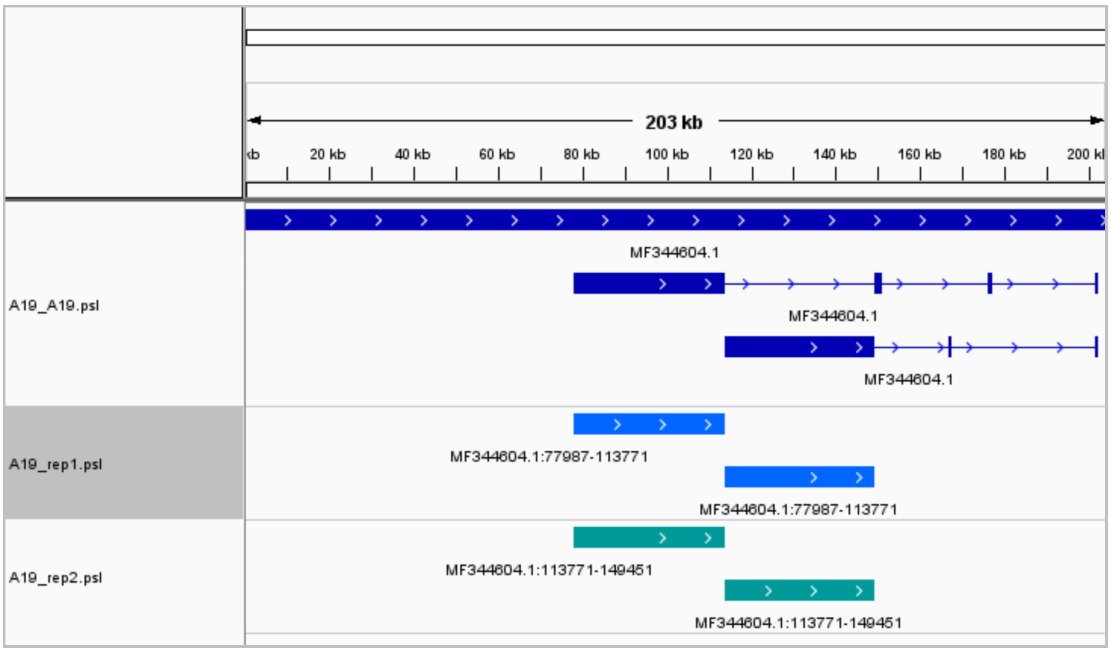
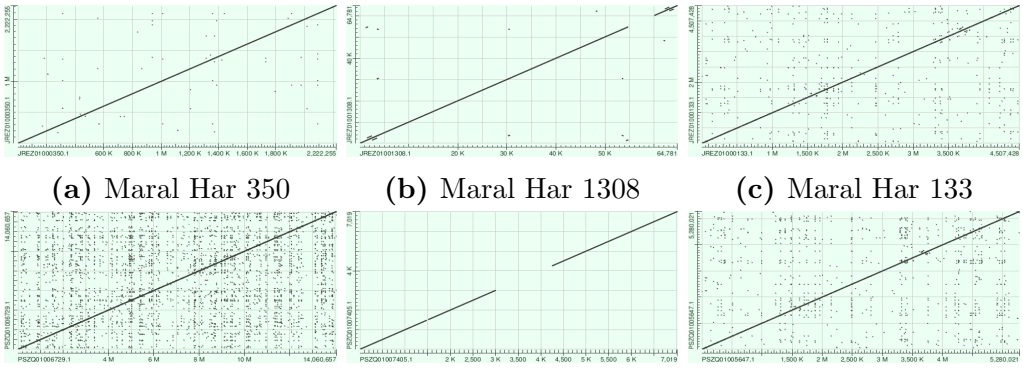


Figure 3.5: BLAT self-alignment of AN19 and its repeat. Full self-similarity is shown in the top alignment, while two shorter alignments each align to a different place. Each of the aligned region were then extracted and re-aligned. Both sequences show alignments back to the same two region, definitively showing that the two sequences are the same.



(a) Maral Har 350 **(b)** Maral Har 1308 **(c)** Maral Har 133

(d) Willy 6729 **(e)** Willy 7405 **(f)** Willy 5647

Figure 3.6: BLAST self-similarity dot-plots of Willy and Maral Har scaffolds

Chromosome	Start	Stop	Length
chr2	8228001	8303056	75055
chr8	5985953	6061008	75055
chr14	87845316	87920371	75055
chr28	33054195	33129250	75055
chrUn	17345505	17420560	75055
chrUn	17345505	17420560	75055

Table 3.2: Random horse sequence coordinates

in the BLAT alignment. Also of note is the fact that while the entirety of the Asino Nuovo repeat unit is present within Maral Har scaffold 1308, scaffold 1308 also contains additional sequence that extends beyond the bounds of the repeat unit at both ends.

The structure of Maral Har scaffold 1308 appears to possibly represent a junction between the end and start of a new repeat, so a re-structured scaffold was created by first dividing the scaffold into two pieces at the site identified by its homology to the Asino Nuovo repeat unit, then by switching the 5' and 3' fragments. A new alignment was then performed against the horse using this novel scaffold, producing a single high scoring continuous alignment at the same locus, rather than two separate alignments with lower scores (Fig. 3.9). Of particular interest is the fact that the rearranged scaffold and the repeat identified by self-similarity alignment in Asino Nuovo align to the exact beginning and ending bases of horse sequence.

Compared to the alignments of the other three donkey isolates, the Willy scaffold 6729 alignment to this region exhibits many relatively small gaps. Like the gap in Willy scaffold 7405 previously mentioned, these gaps tend to be represented by large runs of N's in the Willy scaffold. To examine whether this was a common occurrence in the Willy assembly or rather a feature of this particular region of the genome, five sequence alignments were carried out against the Willy genome using random horse sequences.

The size of the five random horse sequences was chosen to match the size of the horse sequence previously used to identify the Willy scaffolds homologous to the Asino Nuovo repeat unit. Table 3.2 shows the coordinates of the random horse sequences generated, and Figure 3.10 compares the structure of the resultant alignments to the original centromeric probe alignment. Gaps consisting of runs of undetermined bases are indicated in red. Note that the random horse sequence generated from chromosome did not produce a significant continuous alignment to any one scaffold in Willy, and is thus not shown.

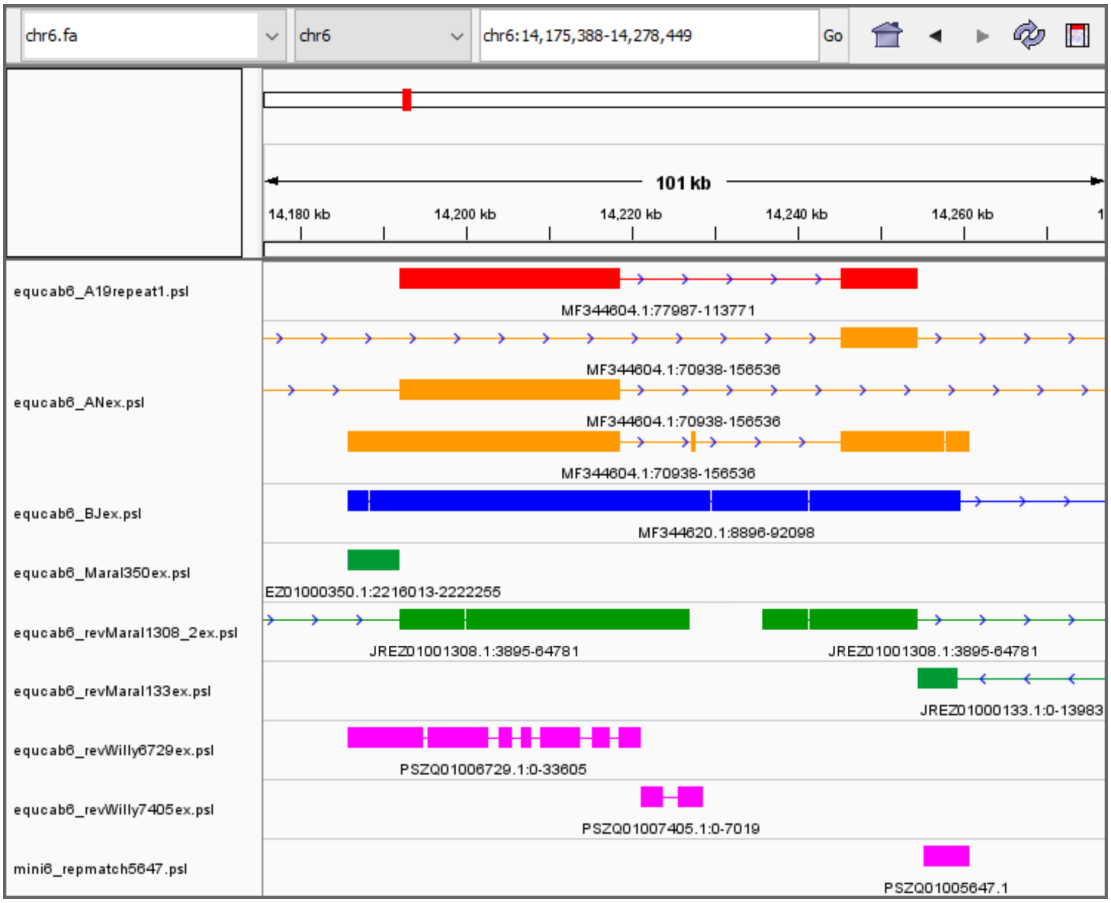


Figure 3.7: BLAT alignments of four donkeys against horse chromosome 6. AN19 repeat unit is shown in red, AN19 full sequence in orange, Blackjack in blue, Maral Har in green, and Willy in pink.

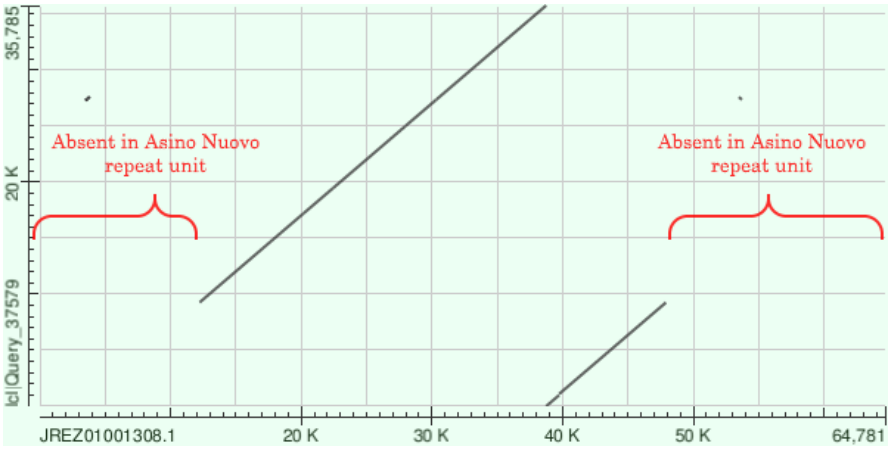


Figure 3.8: BLAST dot-plot of Maral 1308 vs. AN19 repeat unit. Maral Har 1308 is on the x-axis, while AN19 is on the y-axis.

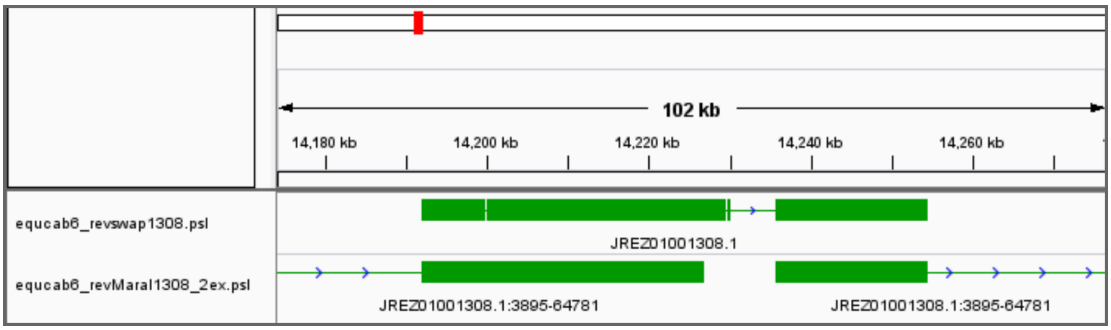


Figure 3.9: BLAT alignment of the rearranged and reversed Maral Har 1308 scaffold, at top. The original alignment is shown below.

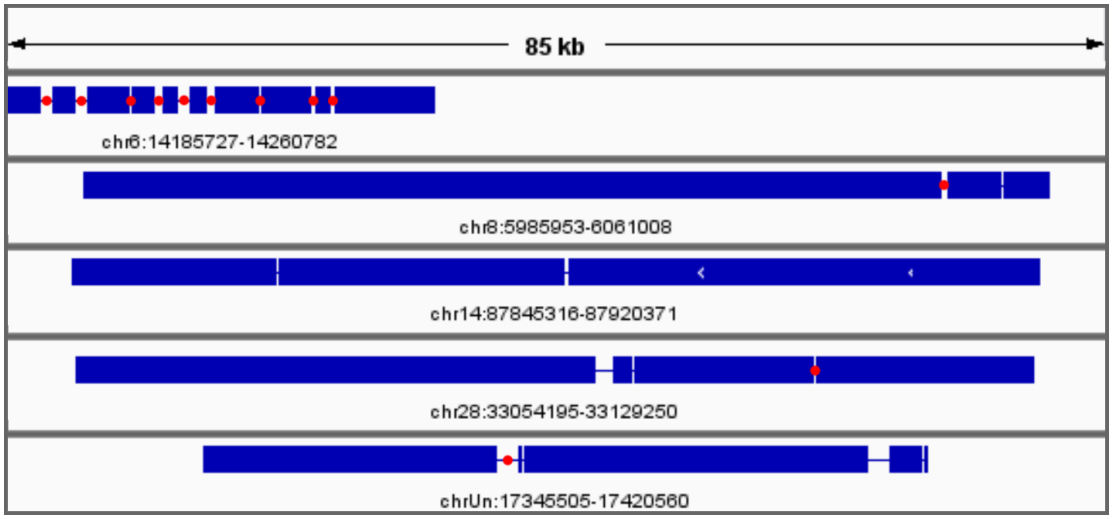


Figure 3.10: Alignment of random horse sequences against Willy. The alignment of scaffold 6729 is shown at top. Runs of undetermined bases (N's) are indicated in red. The random sequence from horse chromosome 2 did not produce a continuous alignment, and is not shown. Each alignment is shown at the same resolution 85 kbp resolution in the context of its own alignment.

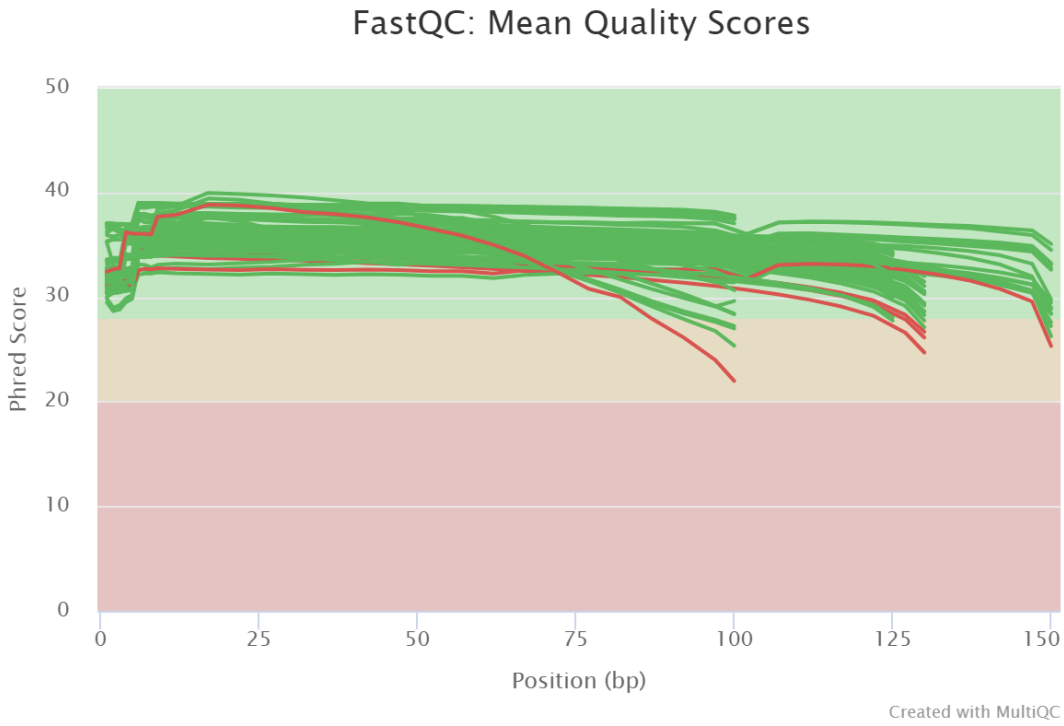
3.4 Raw sequencing read alignment

3.4.1 Quality control

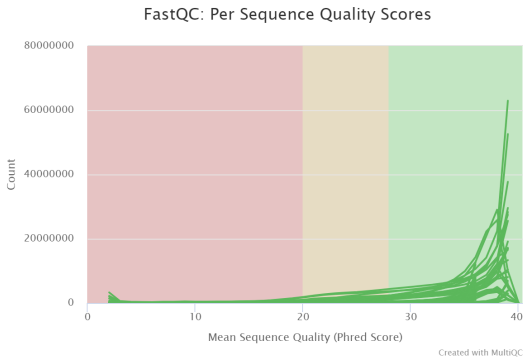
Raw sequencing reads were obtained for three donkey isolates: ChIP-seq and associated input reads for Asino Nuovo and Blackjack, and the whole-genome sequencing reads from Willy that were used to assemble the Willy genome. Quality scores of the raw reads are presented in Figure 3.11, showing overall high quality sequencing, although five out of 82 FASTQ files showed trailing quality scores below 28. Additionally, several FASTQ files from the formaldehyde cross-linked data set showed high numbers of undetermined bases (N's) at their extreme ends, as shown by the spike in Figure 3.11c.

3.4.2 Alignment performance

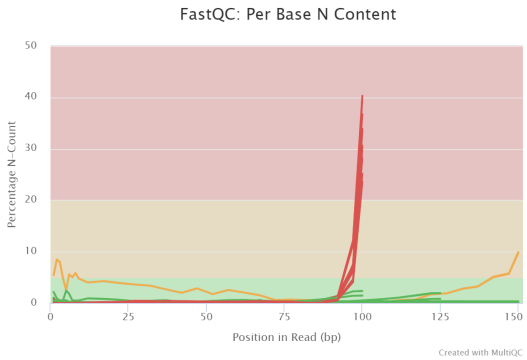
Paired-end alignments with both Bowtie2 and BWA of each data set performed well overall against the horse reference genome EquCab2. Due to an error in pair name recognition, files obtained from the NCBI SRA database were initially run as single-end, rather than paired-end reads. Later realignments were performed for the two SRA data sets with altered parameters to avoid error, resulting in



(a) Mean quality scores



(b) Per sequence quality scores



(c) Per base N content

Figure 3.11: Summary MultiQC graphics assembled from all FastQC analyses.

Sample	Reads	Mapped %	Paired %
MN1	49396978	93.23	89.63
MN2	38167459	89.78	80.63
MN4	42501680	93.63	89.11
MN5	43945873	93.34	88.76
MN6	46612101	91.53	55.07
MN7	20377233	79.24	49.75
MN8	21914479	86.99	73.20
MN9	22650523	67.90	54.15
MN10	17211881	82.23	63.57
MN11	18574433	86.14	66.94

Table 3.3: Bowtie mapping statistics for MNase ChIP-seq runs

actual paired alignment. Summary alignment statistics are shown in Tables 3.3 through 3.8 for each alignment. Overall, all BWA alignments achieved higher proportions of aligned reads than their Bowtie2 counterparts. Notably, aligning Willy sample ERR2286912 with BWA approximately doubled the amount of mapped reads compared to Bowtie2, increasing the mapped proportion from approximately 48% to 96%. For the formaldehyde cross-linked data set and the Willy whole-genome sequencing data set, the paired-end alignments achieved slightly higher mapping rates compared to the single-end alignments for both Bowtie2 and BWA.

3.4.3 Alignment visualization

All alignments were visualized in IGV. BWA and Bowtie2 alignments all resulted in nearly identical overall alignments to the region of interest on horse chromosome 6. Alignments shown in Figure 3.12 are representative, showing the common structure seen across all similar experimental samples, e.g. the BWA Asino Nuovo formaldehyde cross-linked CENP-A ChIP-seq alignment shown mirrors the alignment profile of all other Asino Nuovo ChIP-seq alignments performed in both BWA and Bowtie2. In addition, the formaldehyde cross-linked and Willy whole-genome sequencing data set alignments shown still show the same alignment profiles when aligned in single-end mode. Images of coverage profiles for all alignments performed are available in Appendix C.

3.5 Coverage depth analysis

Coverage depth of the raw sequence alignments against horse chromosome 6 was calculated for loci corresponding to the Asino Nuovo repeat unit, as well as for

Sample	Reads	Mapped %	Paired %
ERR2286905	159176224	90.35	0.00
ERR2286906	162427356	90.44	0.00
ERR2286907	109836640	89.87	0.00
ERR2286908	112702044	90.07	0.00
ERR2286909	246656166	84.41	0.00
ERR2286910	128063686	71.00	0.00
ERR2286911	306265144	75.21	0.00
ERR2286912	223611276	48.51	0.00
SRR5515970	28937922	79.55	0.00
SRR5515971	77041628	90.01	0.00
SRR5515972	37434334	92.53	0.00
SRR5515973	44267364	93.34	0.00
SRR5515974	20012978	92.11	0.00
SRR5515975	25308774	92.45	0.00
SRR5515976	22546630	93.02	0.00
SRR5516015	247035998	91.27	0.00

Table 3.4: Bowtie single-end mapping statistics for formaldehyde cross-linked ChIP-seq runs

Sample	Reads	Mapped %	Paired %
ERR2286905	156171243	90.17	74.21
ERR2286906	159340661	90.26	73.53
ERR2286907	108186242	89.71	75.01
ERR2286908	110970400	89.91	74.58
ERR2286909	240935445	84.04	74.92
ERR2286910	126341358	70.61	41.61
ERR2286911	299114029	74.62	59.86
ERR2286912	221813338	48.09	17.09
SRR5515970	22201327	73.34	63.36
SRR5515971	51809968	85.14	64.79
SRR5515972	35917518	92.22	86.34
SRR5515973	42477573	93.06	86.95
SRR5515974	19699805	91.99	80.79
SRR5515975	24519489	92.21	69.33
SRR5515976	21722658	92.75	67.57
SRR5516015	204157588	89.44	81.82

Table 3.5: Bowtie paired-end mapping statistics for formaldehyde cross-linked ChIP-seq runs

Sample	Reads	Mapped %	Paired %
MN1	48245131	98.07	94.54
MN2	37583097	97.32	89.47
MN4	41570036	98.4	94.54
MN5	42860683	98.3	94.07
MN6	45545993	97.59	92.36
MN7	19455774	89.47	83.56
MN8	21696382	98.19	93.38
MN9	21835833	85.16	73.13
MN10	17084269	94.57	90.37
MN11	18466266	96.33	92.47

Table 3.6: BWA mapping statistics for MNase ChIP-seq runs

Sample	Reads	Mapped %	Paired %
ERR2286905	161315820	96.71	0.00
ERR2286906	164699815	96.85	0.00
ERR2286907	111314223	96.23	0.00
ERR2286908	114278070	96.46	0.00
ERR2286909	250065475	96.04	0.00
ERR2286910	129582715	96.42	0.00
ERR2286911	310269423	96.05	0.00
ERR2286912	225791428	96.00	0.00
SRR5515970	29105343	87.81	0.00
SRR5515971	77546683	95.03	0.00
SRR5515972	37564742	97.54	0.00
SRR5515973	44419579	97.37	0.00
SRR5515974	20107871	97.11	0.00
SRR5515975	25468034	96.62	0.00
SRR5515976	22699676	97.07	0.00
SRR5516015	248144717	95.68	0.00

Table 3.7: BWA single-end mapping statistics for formaldehyde cross-linked ChIP-seq runs

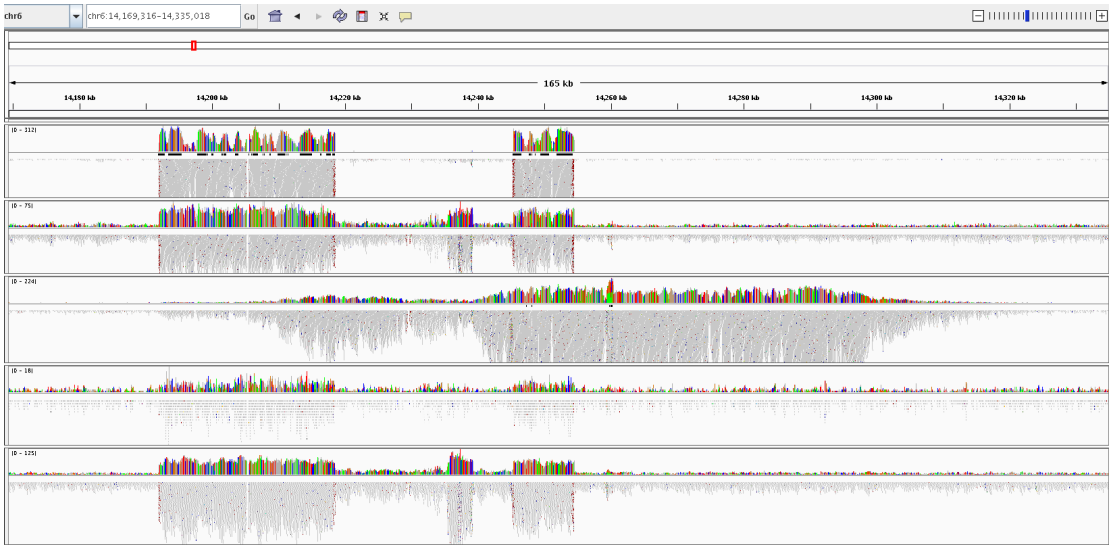


Figure 3.12: Summary visualization of read alignments to horse chromosome 6. All alignments shown are BWA paired-end, and are representative samples of all other alignments performed for the same input/ChIP/WGS run of each isolate. From top to bottom: Formaldehyde cross-linked Asino Nuovo ChIP-Seq, Formaldehyde cross-linked Asino Nuovo ChIP-Seq input, Formaldehyde cross-linked Blackjack ChIP-Seq, Formaldehyde cross-linked Blackjack ChIP-Seq input, Willy WGS

Sample	Reads	Mapped %	Paired %
ERR2286905	153354354	97.76	90.84
ERR2286906	156339531	97.81	90.77
ERR2286907	106522162	97.41	90.19
ERR2286908	109166421	97.53	90.22
ERR2286909	232962173	98.22	93.49
ERR2286910	120994289	98.57	96.08
ERR2286911	285643398	98.50	95.11
ERR2286912	208373905	98.49	96.44
SRR5515970	19907899	85.55	77.19
SRR5515971	45605873	93.21	85.99
SRR5515972	35627697	98.52	94.60
SRR5515973	41980596	98.37	94.54
SRR5515974	19500046	98.00	92.73
SRR5515975	24464293	97.51	91.96
SRR5515976	21575427	97.91	93.00
SRR5516015	198867119	95.84	91.58

Table 3.8: BWA paired-end mapping statistics for formaldehyde cross-linked ChIP-seq runs

regions outside of the repeat unit loci. Only input and whole-genome sequences were included, as uneven non-random read depth can be indicative of copy number. ChIP-seq reads were not included due to the expected pileup as a result of ChIP enrichment. Summary results of the ratio between depth at the Asino Nuovo repeat unit loci and depth of the rest of the chromosome from all non-ChIP alignments are shown in Table 3.9. Complete results showing total depth per region and mean depth per base are available in the appendix.

3.6 Kmer matching to identify JSR

Kmer-matching using BBDuk was used to search for individual reads within the raw FASTQ files. Six sequences were constructed from various loci in the horse genome homologous to parts of the Asino Nuovo repeat. These constructs were used as kmer templates to search for four types of reads: reads spanning the possible deletion in the Asino Nuovo repeat unit, reads entering or exiting the possibly deleted locus, reads spanning the tail-to-head junction between tandem repeat units, and reads entering the repeat unit from non-repeat sequence or exiting the repeat unit into non-repeat sequence, as show in Methods, Figure 2.4.

Table 3.10 shows the number of reads found matching k-mers from three tem-

Sample	Isolate	Depth Ratio
MN1	Asino Nuovo	6.08
MN4	Asino Nuovo	6.24
MN6	Asino Nuovo	6.83
MN8	Asino Nuovo	7
MN10	Asino Nuovo	4.74
SRR5516015	Asino Nuovo	5.66
SRR5515972	Blackjack	3.71
SRR5515975	Blackjack	3.31
ERR2286905	Willy	7.35
ERR2286906	Willy	7.02
ERR2286907	Willy	7.49
ERR2286908	Willy	7.25
ERR2286909	Willy	6.76
ERR2286910	Willy	6.44
ERR2286911	Willy	6.54
ERR2286912	Willy	6.46

Table 3.9: Repeat vs. single-copy region sequencing depth ratios

plates reflecting aspects of the suspected Asino Nuovo deleted sequence. The first template was built from the homologous horse sequence flanking the suspected Asino Nuovo repeat unit deletion locus. The second two templates each contain horse sequence from the suspected deletion locus, as well as continuous sequence from one side of the Asino Nuovo repeat unit. Reads matching the first template are referred to here as xJSR, while reads matching the latter pair of templates are referred to here as non-xJSR. Read counts have been normalized per million reads per run, and individual lane splits from the micrococcal mononuclease ChIP-seq data set have been combined. Table 3.12 shows combined average counts of xJSR for each read type. Note that for the pair of control templates, k-mers were matched against two templates, doubling the probability of a random match to these templates compared to the xJSR template. Counts shown have not been altered to account for this.

Table 3.11 shows the number of reads found matching kmers from three templates built to represent aspects of the repeating structure of the Asino Nuovo repeat unit. The first template consists of horse sequence homologous to the 3' sequence of the repeat unit continuing onto the 5' end of the repeat unit, representing the end of one repeat and the start of the next. The second two templates each contain horse sequence flanking the Asino Nuovo repeat unit homologous locus and sequence homologous to the ends of the Asino Nuovo repeat unit, rep-

Sample	Isolate	Method	xJSR	Non-xJSR
MN07	Asino Nuovo	ChIP	15.92	0.23
SRR5515970	Asino Nuovo	ChIP	14.17	0.14
SRR5515971	Asino Nuovo	ChIP	4.69	0.21
MN11	Asino Nuovo	ChIP	3.7	0.46
MN02	Asino Nuovo	ChIP	1.42	0.12
MN09	Asino Nuovo	ChIP	1.09	0.07
MN05	Asino Nuovo	ChIP	0.88	0.34
MN08	Asino Nuovo	Input	0.48	0.09
ERR2286911	Willy	WGS	0.31	0.23
ERR2286906	Willy	WGS	0.26	0.15
ERR2286909	Willy	WGS	0.24	0.2
ERR2286910	Willy	WGS	0.22	0.12
ERR2286905	Willy	WGS	0.19	0.16
ERR2286907	Willy	WGS	0.16	0.24
ERR2286908	Willy	WGS	0.16	0.18
ERR2286912	Willy	WGS	0.15	0.12
SRR5515976	Blackjack	ChIP	0.13	5.99
MN06	Asino Nuovo	Input	0.12	0.23
SRR5516015	Asino Nuovo	Input	0.11	0.09
MN04	Asino Nuovo	Input	0.09	0.2
MN01	Asino Nuovo	Input	0.08	0.08
SRR5515975	Blackjack	Input	0.08	0.16
SRR5515973	Blackjack	ChIP	0.07	2.39
MN10	Asino Nuovo	Input	0.05	0.43
SRR5515974	Blackjack	ChIP	0	3.5
SRR5515972	Blackjack	Input	0	0.05

Table 3.10: Normalized xJSR read counts

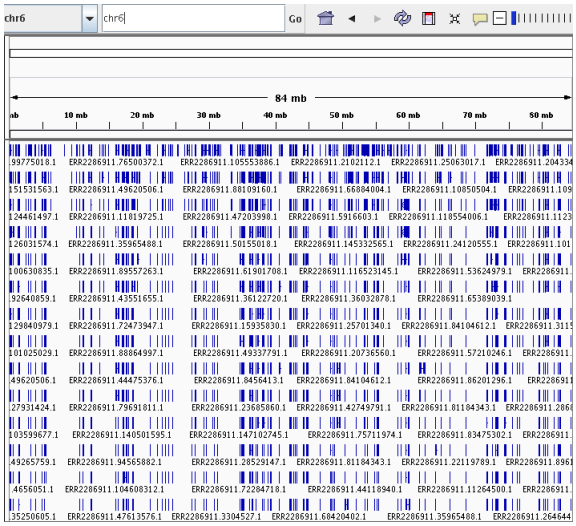


Figure 3.13: Alignment of k-mer matches to the "3'-exiting" sequence. The full horse chromosome 6 is shown in the window.

resenting the start and end of a potential repetitive region. Reads matching kmers from the first template are referred to here as nJSR, while reads matching kmers from the latter two templates are referred to here as non-nJSR. As before, counts are normalized per million reads per run, and split lanes from the micrococcal mononuclease ChIP-seq data set have been combined. Due to the high number of reads found matching the "3'-exiting" template, the number of non-nJSR reads for each template are shown separately as "5'-entering" and "3'-exiting". Table 3.12 shows combined average counts of nJSR for each read type. The high results found for "3'-exiting" prompted an alignment of all matching reads to horse chromosome, the results of which are shown in Figure 3.13.

3.7 CNV calling

The software package CNVcaller was used to estimate copy number at the centromere regions of Asino Nuovo, Blackjack, and Willy, treating the horse sequence as a baseline single-copy reference. CNVcaller uses a sliding window to measure local depth, reporting area containing possible copy-number variants in a modified VCF file. Because ChIP-seq reads are enriched, only the input and whole-genome sequencing runs were passed through CNVcaller. Approximately 1400 possible CNV sites were identifed by the program. Table 3.13 shows loci called by CNVcaller on horse chromosome 6 at 14 - 15 Mbp, the location of alignment of the Asino Nuovo repeat unit. Columns show the average copy number estimated by the program for each isolate. Two regions within this range were reported to have

Sample	Isolate	Method	nJSR	5'-entering	3'-exiting
MN07	Asino Nuovo	ChIP	16.3	0.04	1.92
SRR5515970	Asino Nuovo	ChIP	5.43	0	2.56
SRR5515971	Asino Nuovo	ChIP	4.98	0.01	3.04
MN11	Asino Nuovo	ChIP	3.85	0	2.43
MN02	Asino Nuovo	ChIP	1.52	0	4.13
MN05	Asino Nuovo	ChIP	0.45	0.09	3.96
MN09	Asino Nuovo	ChIP	0.45	0	1.95
MN06	Asino Nuovo	Input	0.31	0.04	2.96
ERR2286907	Willy	WGS	0.3	0.03	3.62
ERR2286909	Willy	WGS	0.27	0.03	2.83
MN10	Asino Nuovo	Input	0.27	0	1.5
MN08	Asino Nuovo	Input	0.26	0.04	2.68
ERR2286905	Willy	WGS	0.25	0.03	3.43
ERR2286908	Willy	WGS	0.25	0.11	3.31
ERR2286910	Willy	WGS	0.24	0.05	3.21
ERR2286906	Willy	WGS	0.23	0.04	3.47
ERR2286911	Willy	WGS	0.21	0.03	2.68
MN01	Asino Nuovo	Input	0.19	0.04	3.52
SRR5516015	Asino Nuovo	Input	0.17	0.02	2.02
SRR5515974	Blackjack	ChIP	0.15	0.05	3.4
ERR2286912	Willy	WGS	0.11	0.02	2.32
SRR5515972	Blackjack	Input	0.11	0	2.48
MN04	Asino Nuovo	Input	0.05	0.02	4.14
SRR5515975	Blackjack	Input	0.04	0	2.45
SRR5515973	Blackjack	ChIP	0.02	0.07	3.98
SRR5515976	Blackjack	ChIP	0	0.04	7.36

Table 3.11: Normalized nJSR read counts

Combined Sample	xJSR	non-xJSR	nJSR	5'-entering	3'-exiting
Asino Nuovo ChIP	5.98	0.22	4.71	0.02	2.86
Asino Nuovo Input	0.16	0.19	0.21	0.03	2.8
Willy	0.21	0.18	0.23	0.04	3.11
Blackjack ChIP	0.07	3.96	0.06	0.05	4.91
Blackjack Input	0.04	0.11	0.08	0	2.47

Table 3.12: Average k-mer match counts per isolate

Chr6 Locus	Asino Nuovo	Blackjack	Willy
1403201-1407200	1.78	1.51	2.22
14113601-14117600	1.98	1.66	2.25
14191201-14218400	12.78	7.02	13.23
14221601-14255200	8.13	3.44	9.77
14293601-14297600	2.03	2.47	2.13
14520801-14524000	2.2	3.35	2.11
14752001-14756000	1.73	1.65	2.01

Table 3.13: CNV candidate windows at 14 Mbp

a much higher copy number than the surrounding area. The VCF file indicating these regions is shown in IGV, along with a representative Asino Nuovo ChIP-Seq alignment, a representative Asino Nuovo input alignment, and the Asino Nuovo repeat unit alignment, in Figure 3.14.

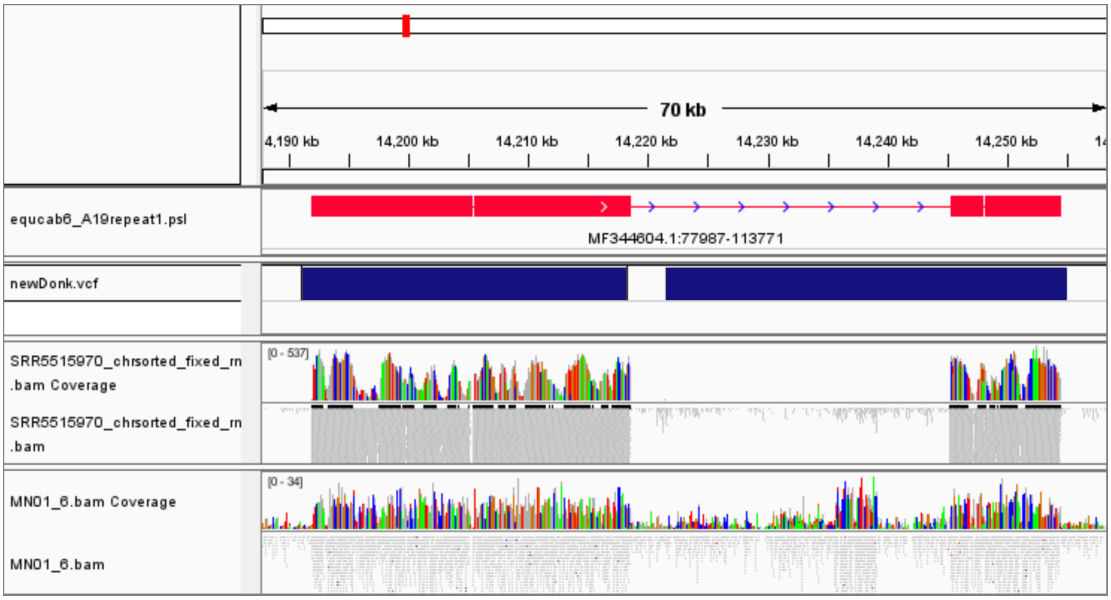


Figure 3.14: Comparison of AN19 repeat locus, detected CNV region, Asino Nuovo ChIP-seq alignment, and Asino Nuovo input alignment

Chapter 4

Discussion

4.1 Centromere location in Willy and Maral Har

Comparisons of the AN19, BJ19, Willy, Maral Har, and horse sequences reveals clear evidence of variation between donkey individuals. While most comparisons in this project use the horse genome as a baseline comparison, it is also possible that the horse sequence has undergone rearrangements in the time since the donkey and horse species diverged, causing some of the alignment differences observed. However, the similarity between the Blackjack and horse sequences make this very unlikely, as both the Blackjack lineage and horse lineage would have had to acquire the same rearrangements independently. Thus, variation between the donkey individuals can be considered to be population polymorphism that has emerged since the divergence of the horse and donkey. For example, the large gap seen in the AN19 alignment to the horse is likely the result of deletion in the donkey genome, rather than an insertion in the horse genome, since the Blackjack alignment retains this sequence.

The Asino Nuovo and Blackjack centromeres were experimentally identified via ChIP-seq targeting CENP-A and CENP-C; however, the Willy and Maral Har sequences were obtained via whole genome sequencing, and thus did not have determined centromeric sequences. Because centromere location is not primarily determined by the underlying DNA sequence, it cannot be said for certain where the centromeres of Willy and Maral Har are located within their published sequences. However, assuming that no great centromere relocations or chromosomal rearrangements commonly exist within the current donkey population, it is likely that the genomic locus containing the centromere in one donkey is the same as the centromeric locus in another donkey. This likelihood is increased in donkeys in particular due to their long history of domestication, which may increase rates of homozygosity through human selection and inbreeding of individuals with favorable traits.[47,50,67]

The homologous sequence alignment performed between the AN19 centromere scaffold, the horse, and both the Willy and Maral Har genomes revealed that the sequence associated with the centromere in Asino Nuovo and Blackjack is present across multiple scaffolds in both Willy and Maral Har. Once these scaffolds were identified, they were aligned to the homologous region of the horse genome alongside both the Asino Nuovo and Blackjack sequences, further showing the relative structure between the WGS-obtained scaffold sequences and the experimentally determined centromeric sequences.

Despite the high homology shown between the sequences of the four individuals and the horse, the cytologenetic evidence of tandem repeats in AN19 found by Nergadze et al. (2018) brings up the clear possibility that the centromere sequence homologues in Maral Har and Willy may not be accurate representations of the true sequence structure in these indviduals' genomes. Furthermore, it should be understood that this is absolutely the case of the AN19 sequence, as this scaffold was assembled as a model including the presence of a repeat unit, but not to necessarily reflect the exact number of repeats present in Asino Nuovo centromere 19. The self-similarity BLAST and BLAT results of AN19 shown in this project reflect this inclusion of two assembled repeat units in this published sequence.

4.2 Structure of Maral Har centromere homologue sequence

The scaffolds identified by homology in Maral Har clearly show a pattern of continuity at this region. Over 90% of bases in the query sequence are matched across scaffolds 350, 1308, and 133. Alignments to these scaffolds all cover the first or last bases of the scaffolds, indicating that the query sequence begins in scaffold 350, runs through scaffold 1308, and ends in scaffold 133. The fact that the alignments to scaffolds 350 and 1308 begin and end at the exact base in the query sequence provides very strong evidence of this continuity. Although there is overlap in the alignment at the 3' end of scaffold 1308 and the 5' end of scaffold 133, this can be explained through the likely repetitive nature of this region. In a scenario where the Maral Har genome contains multiple copies of a sequence similar to that of scaffold 1308, the 5' end of scaffold 133 may contain a second copy of the 3' end of such a repeat, explaining why their alignments overlap. The 3' end of scaffold 1308 then ends at the exact base that the 5' end of scaffold 133 begins. This implies that in Maral Har these three scaffolds are in fact somewhat continuous with each other, forming parts of one larger scaffold. However, the fact that these seemingly nearly-contiguous scaffolds were not assembled into one larger contig implies that there were difficulties in assembly at these regions. This can be explained again by the presence of repetitive sequences, such as tandem

repeats. If scaffolds 350 and 133 in actuality represent the single-copy flanks of a series of tandemly repeated sequences similar to scaffold 1308, assembly would not have been able to correctly identify a contiguous sequence beginning in scaffold 350, crossing through several copies of scaffold 1308, and terminating in scaffold 133.

The tail-to-head arrangement of scaffold 1308 compared to both the AN19 repeat unit and horse provides strong evidence that scaffold 1308 samples the junction between two repeats in Maral Har. As shown in Figure 3.8, scaffold 1308 contains the full AN19 repeat unit, split in two and rearranged. Rearranging the 1308 scaffold based on the structure of the AN19 repeat unit creates a sequence with a single improved alignment to the horse, requiring fewer rearrangement events to explain its structure. Figure 4.1 shows a suggested model for the repeat junction represented in scaffold 1308.

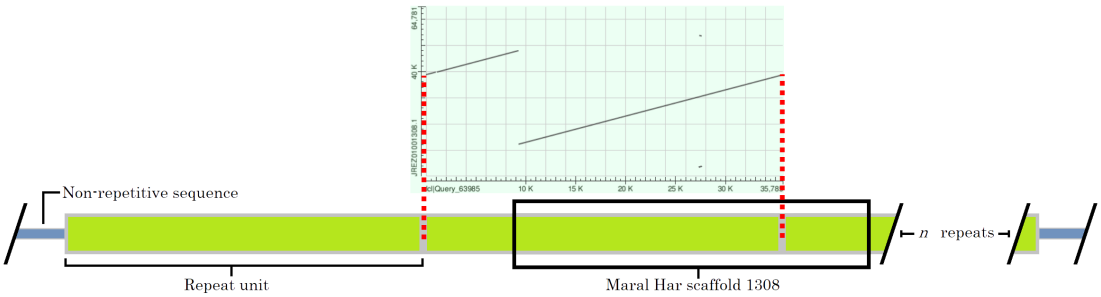


Figure 4.1: Model of Maral Har scaffold 1308. From the BLAST results comparing Maral Har scaffold 1308 to the sequence of the repeat unit seen in Asino Nuovo, a clear 3' to 5' structure emerges (top). The model shows the proposed location of scaffold 1308 within the context of repeating units without deletion in the Maral Har centromere.

Despite its increased similarity to both the horse and Asino Nuovo, the rearranged scaffold still shows significant variation compared to AN19. While the alignment of AN19 to the horse shows a deletion of approximately 26 kbp, the Maral Har scaffold 1308, when rearranged, only shows a much smaller gap of about 6 kbp. This indicates that scaffold 1308 contains sequence absent or moved in the Asino Nuovo genome. Combined with the evidence for repetition of a scaffold 1308-like sequence in the Maral Har genome, this may suggest that the deletion event occurred independently in the two donkeys' lineages, or perhaps to a varying degree. An alternate explanation, requiring fewer independent events, suggests that Maral Har possesses two different centromeric epialleles, and is a centromeric heterozygote. In this model, each Maral Har chromosome 19 possesses a different structure, with one Blackjack-type epiallele maintaining an ancestral structure similar to the single-copy horse sequence, and one Asino Nuovo-type

epiallele showing evidence of repeating units that each contain a large deletion relative to the horse. This model would suggest that the assembly of scaffold 1308 is incorrect, and represents a chimeric combination of the two epiallele sequences. However, without the ability to directly query the sequencing reads used for this assembly, any further elucidation or confirmation of the true structure remains hypothetical.

4.3 Structure of Willy centromere homologue sequence

The Willy scaffolds in this area show a somewhat similar continuity structure when compared to Maral Har. The 3' end of Willy scaffold 6729 overlaps with scaffold 7405 by 80 bp, which may suggest continuity in the same way as Maral Har scaffolds 350 and 1308. if this is the case, the location of scaffold 7405 compared to the other donkeys suggests that Willy, like Maral Har, may not have the same deletion structure as Asino Nuovo. However, analysis of Willy raw sequencing shows very close similarity to Asino Nuovo. Combined with the fact that scaffold 7405 is small in size with a lower matching rate and a significant quantity of undetermined bases, it may be that scaffold 7405 is not continuous here. Furthermore, no direct continuity was seen in the alignments from scaffold 7405 to scaffold 5647. Instead, a large span of approximately 26 kbp was only covered by many small overlapping scaffolds. As in Maral Har, the difficulty in assembling this region into contiguous scaffolds may be due to the presence of tandem repeats.

Compared to alignments of the other 3 donkeys, Willy scaffolds 6729 and 7405 contain a much higher content of N's, resulting in the gaps seen in the alignments. When compared to the alignments of random horse sequences, scaffold 6729 still contained a higher frequency of runs of undetermined bases. This may be a result of using Chicago chromatin capture technology in a region of large tandem repeats. Chicago library preparation involves forming intra-fragment cross-links and tags on fragments several hundred of kilobases long. These fragments are subsequently digested and ligated together before being released from their protein cross-links. This results in reads that have been tagged as being located on the same 100-kilobase-scale sequence in the genome. While this is generally very useful for establishing contiguous sequences and scaffolds in genome assembly, it may be that large tandem repeats confound this process by creating tagged long-range linked reads with conflicting insert sizes and order. Thus, while short assemblies would still be well established, their overall order and the distances between them could face inconsistencies between individual Chicago structures producing linked sequences from different locations within a tandem repeat region. In essence, the

runs of undetermined bases in Willy scaffold 6729 may represent indeterminate distances between linked sequences that are present in multiple tandem repeats. This is further supported by the fact that alignment of scaffold 6729 sequence flanking the region homologous to the AN19 repeat unit does not show this pattern of interspersed undetermined bases (Fig. 3.3. While there is less evidence for such repeat structure in the Willy scaffold assemblies and alignments, further evidence was found in the outcome of raw sequencing alignment.

4.4 Sequencing read alignments

All Bowtie2, BWA, single-end, and paired-end alignments all produced nearly identical alignment for each run, providing very strong evidence for accurate mapping results. Overall, BWA was found to map at least as well as Bowtie2 for all runs performed, and significantly better for several runs. While paired-end alignment did not perform vastly different in terms of the proportion of reads that were able to be mapped, visualization of BWA Asino Nuovo and Willy paired-end alignments in IGV revealed the presence of annotated read pairs aligning with abnormally large insert size to either side of the Asino Nuovo deletion site, as well as forward reads aligning at the 3' end of the overall alignment, while their reverse pairs aligned to the 5' end, providing definitive evidence for the repeat structure and deletion in both Asino Nuovo and Willy.

The raw sequencing alignment results for Asino Nuovo and Blackjack ChIP-seqs were consistent with their assembled sequences, aligning to the same areas shown in their scaffold alignments. Blackjack exhibits a somewhat typical tapering distribution at this locus, with high central read depth that slowly tapers, indicative of a normal sampling of reads across the CENP-A distribution in the centromere. Because Blackjack is expected to have a single-copy sequence, similar to the horse, the input Blackjack alignments are expected to be somewhat level and evenly distributed. However, a small level of pileup is noticeable at each of the three peaks seen in the Asino Nuovo alignment. In Asino Nuovo, these pileups directly parallel the AN19 repeat unit, which was not noted in the BJ19 assembly. Despite the lack of inclusion in the BJ19 sequence, evidence for tandem repeats in Blackjack centromere 19 was found in the qPCR experiments performed by Nergadze et al., so the appearance of these peaks is not entirely unexpected. What may be more surprising is the fact that CENP-A does not appear to be associated with this stretch of amplified DNA, the centromere instead localizing somewhat downstream.

The Asino Nuovo ChiP-seq alignment occurred directly at the site of the AN19 scaffold alignment, arranged into two distinct regions as expected. Very few, if any, reads were found outside of this domain. Furthermore, the same characteristic

deletion is noted in the gap between read alignments. The oddly specific alignment region is unusual and does not follow the same tapering distribution seen in the Blackjack alignment, but does resemble the "spike" distribution observed in earlier Asino Nuovo alignments. The blocky distribution without tapering ends can be explained by the CENP-A distribution being completely contained within a sequence of tandem repeats. A model of this is shown in Figure 4.2. As described previously, this domain does not cover the expected range that would normally compose a fully functioning centromere, such as the 100 kbp range seen in the Blackjack alignment. However, as modeled, this narrow alignment is likely the result of a much longer region of multiple tandem repeats aligning to the same single-copy sequence in the horse.

Although input sequences are not subjected to immunoprecipitation and read selection, each Asino Nuovo input run nevertheless mirrored the distribution seen by the ChIP runs. In the context of the Asino Nuovo model proposed, this structure is easily explained by the multiple copies of non-unique repeating structures in Asino Nuovo underlying this region being sequenced and aligned to the single copy in the horse genome, resulting in an increased depth proportional to copy number that can be compared to the otherwise level background.

Curiously, input reads are present throughout the region between the ChIP-seq pileups, and furthermore pileup occurs not only at the two regions identified by ChIP-seq, but also at a third location located within what was believed to be a sequence deleted from the Asino Nuovo genome. The fact that there is no ChIP-seq pileup between the two blocks and the fact that there are paired reads aligning to either side of this region imply that the two blocks comprise a continuous region in Asino Nuovo, with the ancestral sequence between them presumably deleted. However, the presence of input reads in this location suggest that the sequence does exist, perhaps translocated elsewhere in the genome. However, the amplification noted within this region implies that it was somehow amplified alongside the two amplified blocks flanking it. Somehow, this sequence has undergone the same amplification as the tandem repeats while simultaneously not being present with the tandem repeats associated with the centromere in Asino Nuovo. While some effort was made to clarify the nature of the additional pileup site, not much information was gained. BLAST search results did not offer any concrete insights, nor did BLAT searches against the Willy genome. While some low confidence matches were produced against sheep, cows, and yak, implying that this sequence in the horse is just a loosely-conserved intergenic region, BLAST also produced many small matches to roughly 800 bp at the 5' region of the pileup showing approximately 70 - 80% matching with horse satellite DNA. While it may be that this is a singular occurrence of a random sequence similar to satellite elements, further investigation may be warranted.

The Willy WGS sequencing alignment resulted in alignments nearly identical

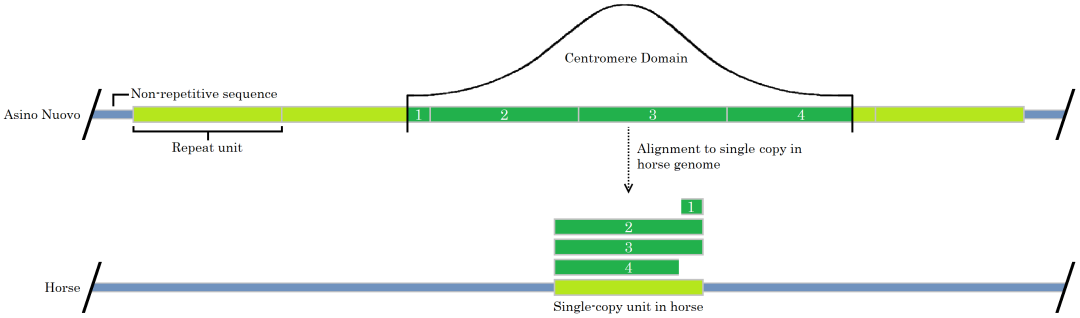


Figure 4.2: Asino Nuovo pileup model. Multiple copies from the Asino Nuovo ChIP-seq reads align against the single copy in the horse, producing a squarely defined region of alignment rather than the true CENP-A or CENP-C distribution.

to those of all Asino Nuovo input alignments. Because these were WGS reads, the pileup noted would normally be unexpected behavior. However, the matching pileup seen in Willy is clear evidence of a copy number difference relative to the horse, caused by the presence of the same tandem repeats seen in Asino Nuovo. The lack of ChIP-seq data for Willy makes it difficult to ascertain whether the centromere is actually associated with this amplified region, however, as it is possible that the centromere could instead be located downstream as seen in Blackjack, or anywhere else.

4.5 Tandem repeat quantification

With strong evidence for the presence of a repeating unit in the sequencing data of Willy, Asino Nuovo, and Blackjack, estimation of its copy number in each individual was achieved through alignment read-depth analysis. Quantification was performed by comparing the average sequencing depth of the tandem repeat loci to that of the rest of the chromosome. Results were very consistent, showing that the amplified region in Asino Nuovo was sequences approximately 6 times more than non-amplified regions. The same regions in Willy were sequenced at 7 times depth, while Blackjack showed only 3.5 times depth. These values obtained for Asino Nuovo and Blackjack directly agree with the qPCR results obtained by Nergadze et al., lending support to the theory that Blackjack may be a heterozygote exhibiting one amplified allele, and one ancestral allele, producing a total of half as many tandem repeats compared to Asino Nuovo. However, due to the fact that Blackjack mules show half again as many repeats through qPCR, it is more likely that Blackjack is also homozygous, and instead contains tandem

repeats at a lower quantity in both chromosomes.

The level of amplification seen in Willy seems to roughly agree with the amount seen in Asino Nuovo, implying that both individuals are homozygous for an amplified sequence and rejecting the idea that Willy scaffolds 6729 and 7405 are continuous. For these calculations, only the regions identified as associated with CENP-A in Asino Nuovo were considered as tandem repeat areas. Although amplification was definitely seen in the third pileup region in Asino Nuovo, this region was included in the single-copy counts due to the uncertainty surrounding its actual presence at this locus. However, it's consideration in either count did not significantly pull the average values in either direction due to its small size.

Further evidence for the presence and quantity of deletion and repeat structures was found by searching for specific reads that could only exist if these structures were present, comparable to the aberrant read pairs noted in the BWA paired-end alignments.

In searching for reads representing junctions between tandem repeats, junction reads crossing into or out of unique DNA were used as a control. Large numbers of reads corresponding to sequence exiting from the 3' end of the repeats into unique sequence were immediately obvious for all samples. This was an unexpected result, as these exiting reads should hypothetically be present in equal amount to reads entering into the repetitive region from unique sequence, based on evidence from sequence alignment. Even if this was not the case, it is highly unlikely that any arbitrary 50 bp locus would be randomly sequenced to such a high degree in either input or whole genome sequencing. This result prompted an additional BLAT search of all k-mer-matching reads against horse chromosome 6, which explained the unexpected result. For most sequences matching k-mers from the queried 50 bp reference sequences, the filtered reads aligned back to the sequences used to construct their kmer references, as expected. However, in the case of the high numbers of reads exiting the repeat region, BLAT alignment showed that these filtered reads were homologous to sequences throughout the entirety of horse chromosome 6 (Fig. 3.13. Upon closer investigation, the horse sequence at these many homologous sites, including the locus used to build the "exiting" reference sequence for k-mer matching, were masked regions of EquCab2, indicating that they are known repetitive sequences in the horse genome. While the total count of reads matching k-mers in the "exit" sequence is quite high because of this, the actual amount of reads matching the desired locus is much closer to that of the "entering" sequence in each run, as observed through IGV visualization.

Aside from these matches to a known small repetitive element, all other results exhibited expected totals. ChIP-seq runs for both Asino Nuovo and Blackjack produced more matches compared to their input counterparts, likely due to the fact that ChIP-seq reads are enriched for the queried region in general, whereas input reads only pileup here due to differences in copy number. Perhaps the most

interesting results are shown by the contrast between the counts shown in the ChIP-seq run of Asino Nuovo and Blackjack. Blackjack reads showing the presence of a continuous, undeleted sequence in the suggested amplified area occurred 50 times as often (25 times when accounting for the use of two templates) compared to the frequency of reads possibly spanning such a deletion. In Asino Nuovo ChIP runs, the opposite situation is found, where the frequency of reads spanning the deletion is 27 times higher than that of reads entering into the region, lending further evidence to the idea that this area is not present in Asino Nuovo. Willy and Asino Nuovo input k-mer matching produced very similar results, which did not show variation between read counts of the two scenarios. This may be due to low read counts at these locations compared to overall reads, and does not demonstrate evidence for or against a deletion at this locus.

Results for reads spanning repeat junctions in Asino Nuovo ChIP-seq were found with a frequency much higher than that of reads entering into a repeat unit from unique sequence. In Blackjack ChIP-seq, these values did not show variation. Despite the lack of variation seen between the read counts in Willy and the Asino Nuovo input runs, both had higher counts for repeat-spanning junctions compared to the presumed single-copy control. These results are in line with the idea that both Willy and Asino Nuovo contain multiple tandem repeats, each containing a common deletion relative to the horse sequence.

The presence of tandem repeats in the donkey genome compared to a single copy of similar sequence in the horse genome presents a situation similar to copy-number variant analysis. The results from applying CNV analysis software to the donkey sequencing reads directly support all other quantitative evidence of tandem repeat quantity. Two potential CNV regions were detected covering the Asino Nuovo centromeric region, which were determined jointly from the overall population of alignments supplied. The diploid average copy number determined for Asino Nuovo and Willy in the first region was estimated by the program to be 12.78 and 13.23, respectively, while Blackjack was estimated to have a copy number of 7.02. These values agree with the haploid number estimated for all three individuals by read depth analysis and by qPCR, in the case of Asino Nuovo and Blackjack.

4.6 Sequence model of centromere 19

From the combined quantification results from CNV calling, read depth analysis, and prior qPCR results, it appears likely that Asino Nuovo centromere 19 contains six copies of a tandem repeat unit on each of its two chromosomes, with Willy mirroring this layout. Furthermore, six tandem repeats of 35 kbp each creates a continuous sequence similar in length to that of the Blackjack CENP-A

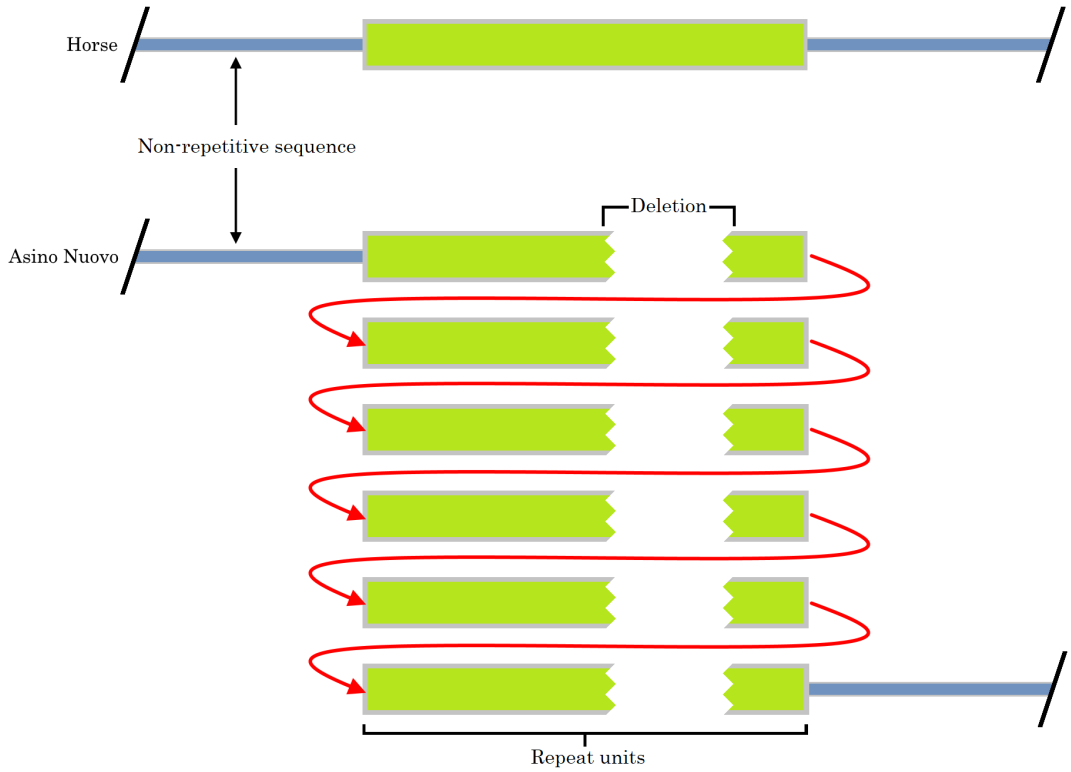


Figure 4.3: Model of Asino Nuovo centromere 19 sequence structure, compared to the homologous single-copy horse sequence.

domain. Blackjack instead appears to have 3 tandem repeat units on each copy of centromere 19, although its centromere is not associated with this region or repeats.

For Asino Nuovo, the presence of six copies per chromosome is directly in line with the number of reads found spanning deletions and the number of reads found spanning repeats. As shown in Figure 4.3, six repeats contain six deletions and five repeat junctions, paralleling the 5.98 to 4.71 ratio of deletion-spanning reads to repeat junction spanning reads found in Asino Nuovo.

Chapter 5

Conclusion

Based on a combination of homologous sequence alignment, raw sequencing read alignment, raw read k-mer matching, read depth analysis, CNV analysis, and previous qPCR data, an overall view of donkey centromere 19 and its inherent variation has been presented. A clear model emerges for the structure of Asino Nuovo centromere 19, which consists of a deletion-containing unit tandemly repeated six times, completely containing the centromere's distribution of CENP-A. The structure of the homologous location in Willy is likely very similar, despite differences seen in homology searches between the published sequences of the two isolates. Given its structural similarity to the Asino Nuovo sequence, this locus in Willy likely contains the centromere, though without experimental targeting of centromere proteins, this cannot be said for certain as variation in centromere location is clearly possible, as seen between Asino Nuovo and Blackjack. Although Blackjack centromere 19 is not located on the same locus as the Asino Nuovo centromere, the same tandem repeat units are found at the same locus as well, seemingly without deletion and repeated only three times. The structure of the Maral Har sequence at this locus, although speculated to be heterozygous, was not determined, as raw read data was unavailable.

Further work is required to elucidate the nature of the amplified unit that was clearly seen in both the Asino Nuovo input and Willy whole-genome sequencing data, but was not present in the Asino Nuovo ChIP-seq data. While its level of duplication suggests its involvement in some kind of amplification process and thus negates the notion that it may be the cause of spurious read alignments from transposons or other repetitive elements, its absence from ChIP data clearly shows its excision from this region. Rearrangements of the donkey sequence relative to the horse seem unlikely, as any rearrangement justifying its amplification by placing it within or adjacent to the tandem repeat units would also necessitate its presence in the centromere, and thus its presence in the ChIP-seq data. It may be that whatever mechanism duplicating the tandem repeat units also creates a

byproduct sequence elsewhere through some type of crossover event each time a repeat unit is duplicated, resulting in equivalent copy number between the two. The presence of a small sequence of satellite DNA within its bounds is also thought provoking, as perhaps this is evidence of the initial mechanism that produced this evolutionary new centromere in the donkey. Alternatively, the sequence is just a ubiquitous short satellite present throughout the genome and happens to be at this particular locus. Further cause to look into this unit is seen in a faint column of abnormal read alignments at its center.

Despite the ability to sequence satellite-free neocentromeres, many questions remain unanswered. The presence of tandem repeats at a neocentromere may point to the ongoing accumulation of repetitive elements at a young, developing centromere. The fact that Blackjack centromere 19 is not associated with these tandem repeats is cause for more confusion, however. Work by McCarter (2016) demonstrated Blackjack centromere positions sliding in a single generation. It may be that sequence amplification at centromere 19 did not proceed long enough in Blackjack to fully encompass and capture the centromere before it slid away from the amplified sequence, and the tandem repeats present are evidence of a former centromere location in the Blackjack lineage.

Despite being such a fundamental and conserved element of the eukaryotic genome, the centromere is clearly an extremely dynamic structure. It's self-deterministic nature makes it an enigma with no clear beginning or end. Nevertheless, the variation and polymorphism seen in the current evolutionary snapshot of donkey neocentromeres promise to provide new insights into the core developmental processes of neocentromeres in days to come.

Appendices

List of Figures in Appendices

List of Tables in Appendices

Appendix A

Software and hardware

A.1 Hardware

Most computationally-intensive procedures, such as short-read alignment, were performed on the NUI Galway High-Performance Computing Cluster "Syd".

Remote access to Syd and less demanding procedures, such as multiple sequence alignment, were performed on a Dell Inspiron laptop with an Intel i7-7700 2.8 GHz processor and 16 GB RAM, running Windows 10 as well as Ubuntu 16.04 LTS and 18.04 LTS Linux virtual environments.

Local network access to Syd was performed on an Apple iMac with an Intel i7 3.4 Ghz processor and 32 GB RAM, running iOS X Mountain Lion.

A.2 Software

The following table lists software versions used for this project. Multiple versions of certain software are included, and access type to each software version is listed.

Software and Versions Used		
Name	**Version**	**Access**
BBDuk	38.11	HPCC
bedtools	2.15.0 2.26.0	HPCC Local
Blasr	1.3.1 5.3	HPCC Local
BLAST		Web
BLAT	35 36x2	HPCC Local
BWA	0.7.12-r1039	HPCC
Bowtie2	2.1.0	HPCC
CNVcaller	4.0	Local/HPCC
FastQC	0.10.0	HPCC
IGV	2.1.17 2.4.0 2.4.4	HPCC HPCC Local
MultiQC	1.3	HPCC
Perl	5.20.1 5.26.1	HPCC Local
Python	2.7.6 3.6.5	HPCC Local
Samtools	0.1.18 1.8 1.7	HPCC HPCC Local
SRAtools	2.5.0	HPCC

Table A.1: Software

Appendix B

Accessions and data

B.1 Assembly accessions

Assembly Accession Numbers and Associated Values		
GenBank Accession	Isolate	Scope
PSZQ00000000.1	Willy	Whole Genome
JREZ00000000.1	Maral Har	Whole Genome
MF344604.1	Asino Nuovo	Centromere 19
MF344620.1	Blackjack	Centromere 19

Table B.1: Assembly accessions

B.2 Sequencing read accessions

Raw Read Accession Numbers and Associated Values				
Bioproject	SRA Run	Isolate	Strategy	Target
PRJEB24845	ERR2286905 ERR2286906 ERR2286907 ERR2286908 ERR2286909 ERR2286910 ERR2286911 ERR2286912	Willy	WGS	N/A
PRJNA385275	SRR5515970 SRR5515971 SRR5515972 SRR5515973 SRR5515974 SRR5515975 SRR5515976 SRR5516015	Asino Nuovo Asino Nuovo Blackjack Blackjack Blackjack Blackjack Blackjack Asino Nuovo	ChIP-seq	CENP-A CENP-A Input CENP-A CENP-A Input CENP-A Input

Table B.2: Sequencing read accessions

B.3 Asino Nuovo Micrococcal Nuclease Runs

Asino Nuovo MNase ChIP-seq File Directory		
File Name	Target	Method
01_AGAGGATG_L1 01_AGAGGATG_L2 01_AGAGGATG_L3 01_AGAGGATG_L5	CENP-A Mononucleosomes	Input
02_ACGCTTCT_L1 02_ACGCTTCT_L2 02_ACGCTTCT_L5	CENP-A Mononucleosomes	ChIP-seq
04_AGTCAGGT_L1 04_AGTCAGGT_L2 04_AGTCAGGT_L3 04_AGTCAGGT_L5	CENP-A Mononucleosomes	Input
05_TAGCAGGA_L3 05_TAGCAGGA_L5	CENP-A Mononucleosomes	ChIP-seq
06_CATGGATC_L3 06_CATGGATC_L5	CENP-A Trinucleosomes	Input
07_CTCGAACA_L3 07_CTCGAACA_L5	CENP-A Trinucleosomes	ChIP-seq
08_TCGACAAG_L3 08_TCGACAAG_L5	CENP-C	Input
09_AGTGCATC_L3 09_AGTGCATC_L5	CENP-C	ChIP-seq
10_TGGCTACA_L3 10_TGGCTACA_L5	CENP-C	Input
11_GCATAGTC_L3 11_GCATAGTC_L5	CENP-C	ChIP-seq

Table B.3: Asino Nuovo Micrococcal Nuclease Runs

Appendix C

Sequencing read alignments

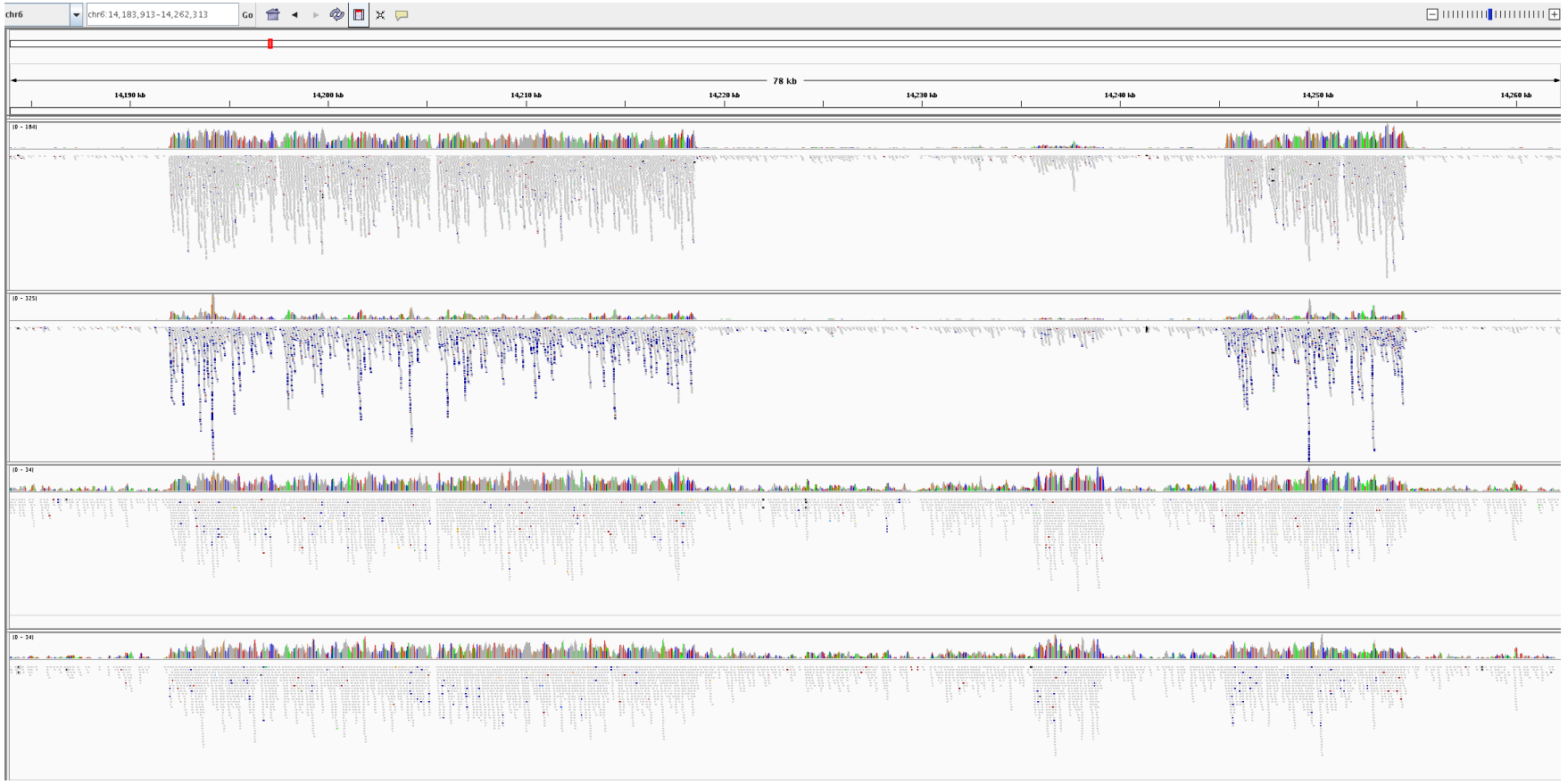


Figure C.1: Bowtie2 alignment, MNase ChIP-seq of Asino Nuovo CENP-A mononucleosomes. Samples from top to bottom are 02 (ChIP-seq), 05 (ChIP-seq), 01 (input), and 04 (input).

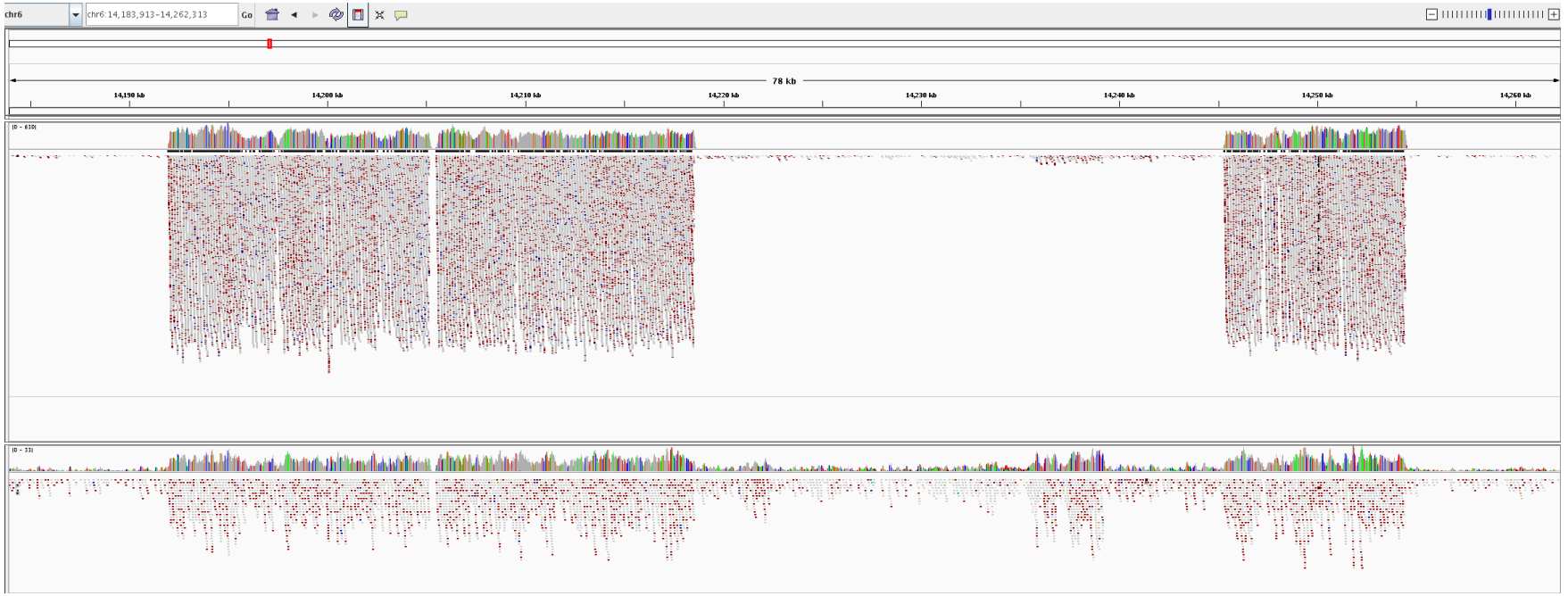


Figure C.2: Bowtie2 alignment, MNase ChIP-seq of Asino Nuovo CENP-A trinucleosomes. Samples from top to bottom are 06 (ChIP-seq) and 07 (input).

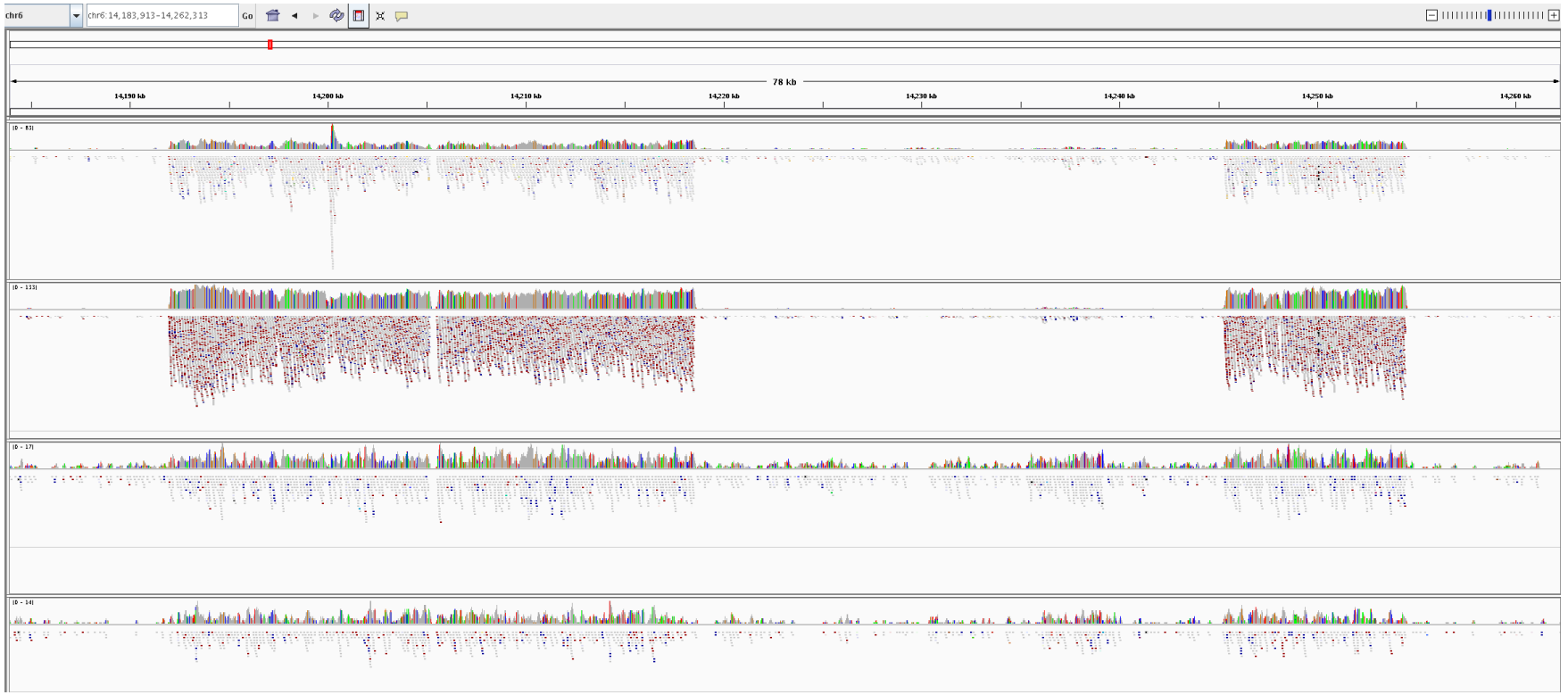


Figure C.3: Bowtie2 alignment, MNase ChIP-seq of Asino Nuovo CENP-C. Samples from top to bottom are 09 (ChIP-seq), 11 (ChIP-seq), 08 (input), and 10 (input).

Figure C.4: Bowtie2 alignment, Formaldehyde Cross-Linked ChIP-seq of Asino Nuovo and Blackjack CENP-A. Samples from top to bottom are Asino Nuovo SRR5515970 (ChIP-seq), SRR5515971 (ChIP-seq), SRR5516015 (input); and Blackjack SRR5515973 (ChIP-seq), SRR5515974 (ChIP-seq), SRR5515976 (ChIP-seq), SRR5515972 (input), and SRR5515975 (input).

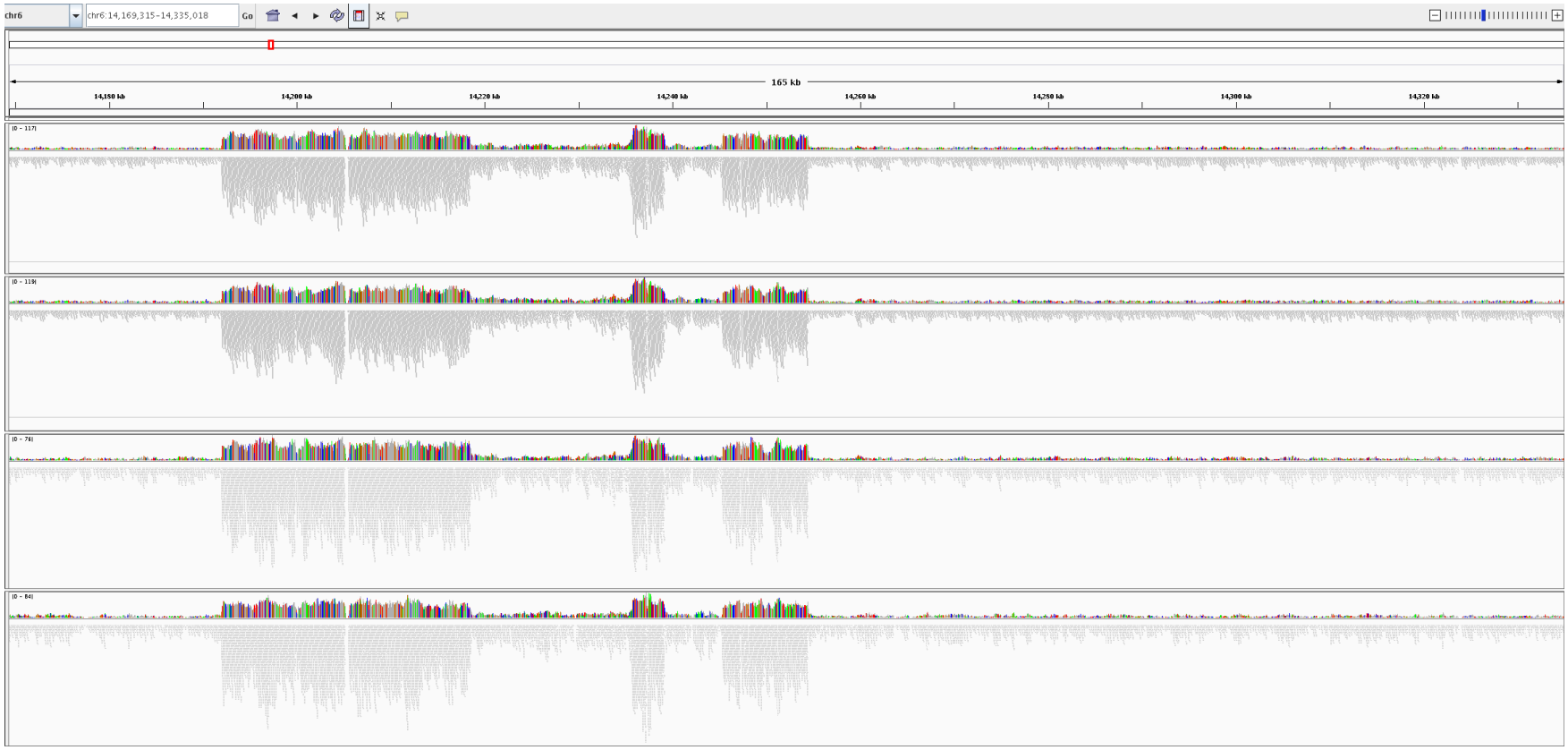


Figure C.5: Bowtie2 alignment, Willy whole genome shotgun sequencing, samples ERR2286905 through ERR2286908.

Figure C.6: Bowtie2 alignment, Willy whole genome shotgun sequencing, samples ERR2286909 through ERR2286912.

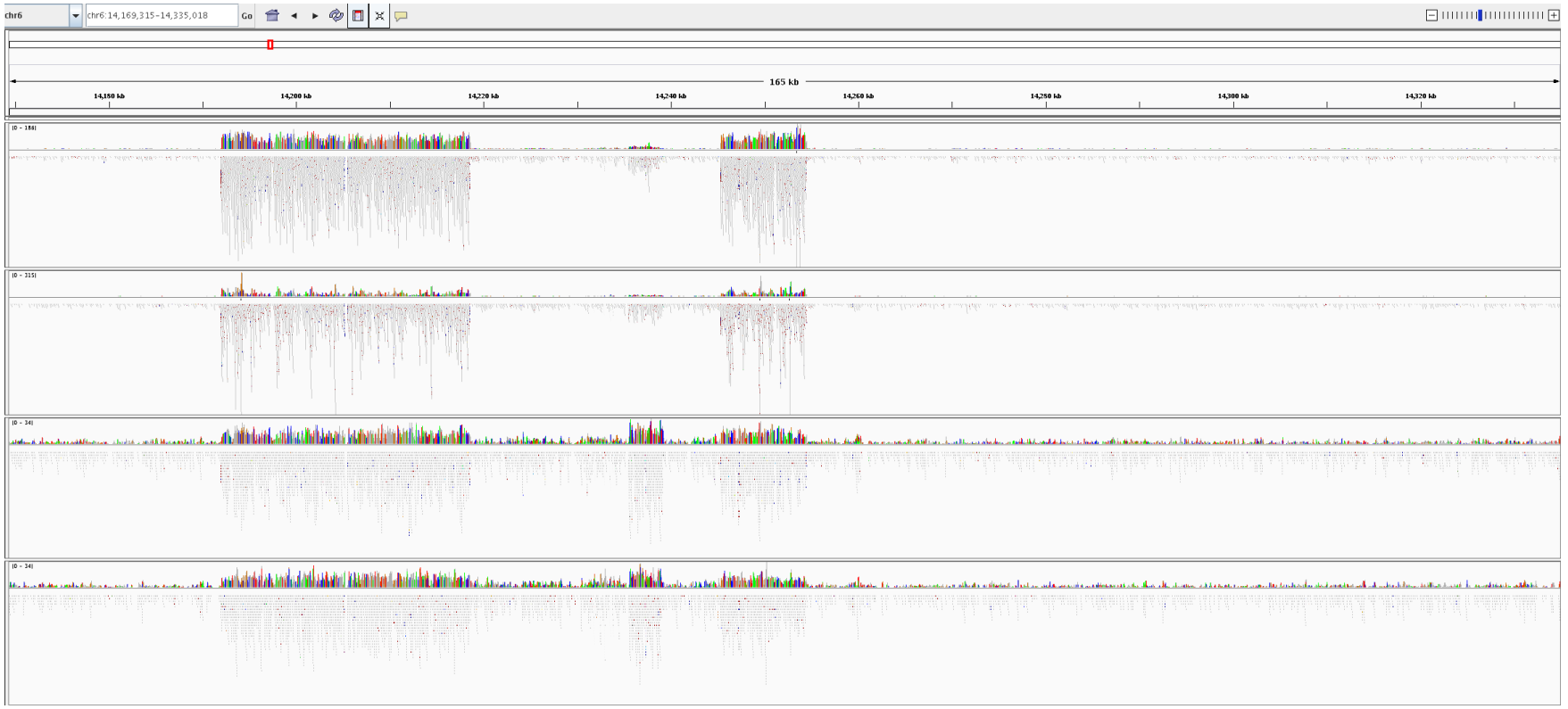


Figure C.7: BWA alignment, MNase ChIP-seq of Asino Nuovo CENP-A mononucleosomes. Samples from top to bottom are 02 (ChIP-seq), 05 (ChIP-seq), 01 (input), and 04 (input).

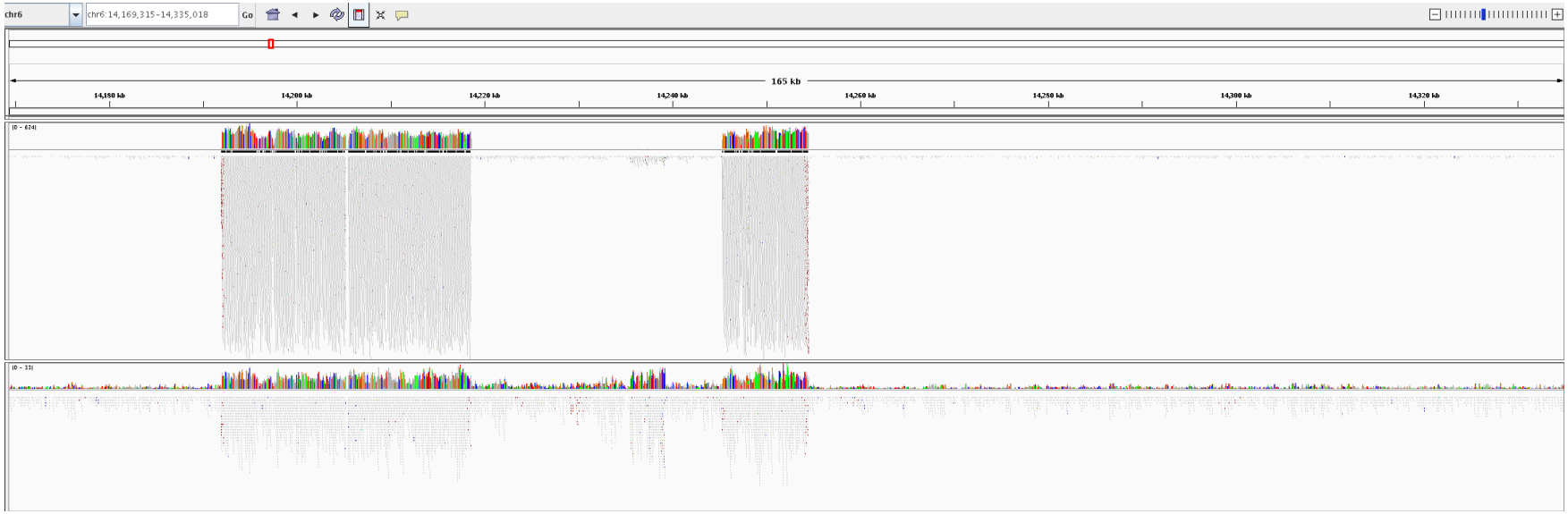


Figure C.8: BWA alignment, MNase ChIP-seq of Asino Nuovo CENP-A trinucleosomes. Samples from top to bottom are 06 (ChIP-seq) and 07 (input).

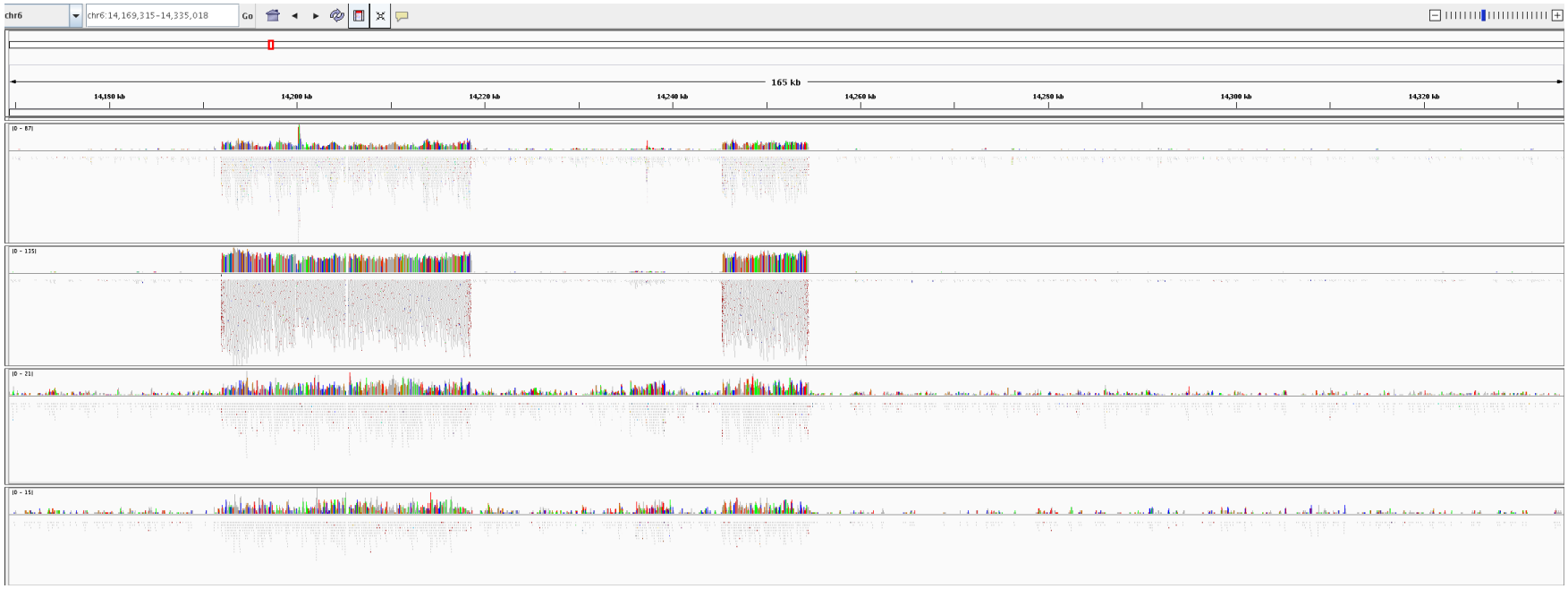


Figure C.9: BWA alignment, MNase ChIP-seq of Asino Nuovo CENP-C. Samples from top to bottom are 09 (ChIP-seq), 11 (ChIP-seq), 08 (input), and 10 (input).

Figure C.10: BWA alignment, Formaldehyde Cross-Linked ChIP-seq of Asino Nuovo and Blackjack CENP-A. Samples from top to bottom are Asino Nuovo SRR5515970 (ChIP-seq), SRR5515971 (ChIP-seq), SRR5516015 (input); and Blackjack SRR5515973 (ChIP-seq), SRR5515974 (ChIP-seq), SRR5515976 (ChIP-seq), SRR5515972 (input), and SRR5515975 (input).

Figure C.11: BWA alignment, Willy whole genome shotgun sequencing, samples ERR2286905 through ERR2286908.

Figure C.12: BWA alignment, Willy whole genome shotgun sequencing, samples ERR2286909 through ERR2286912.

Figure C.13: Bowtie2 paired-end re-run alignment, Formaldehyde Cross-Linked ChIP-seq of Asino Nuovo and Blackjack CENP-A. Samples from top to bottom are Asino Nuovo SRR5515970 (ChIP-seq), SRR5515971 (ChIP-seq), SRR5516015 (input); and Blackjack SRR5515973 (ChIP-seq), SRR5515974 (ChIP-seq), SRR5515976 (ChIP-seq), SRR5515972 (input), and SRR5515975 (input).

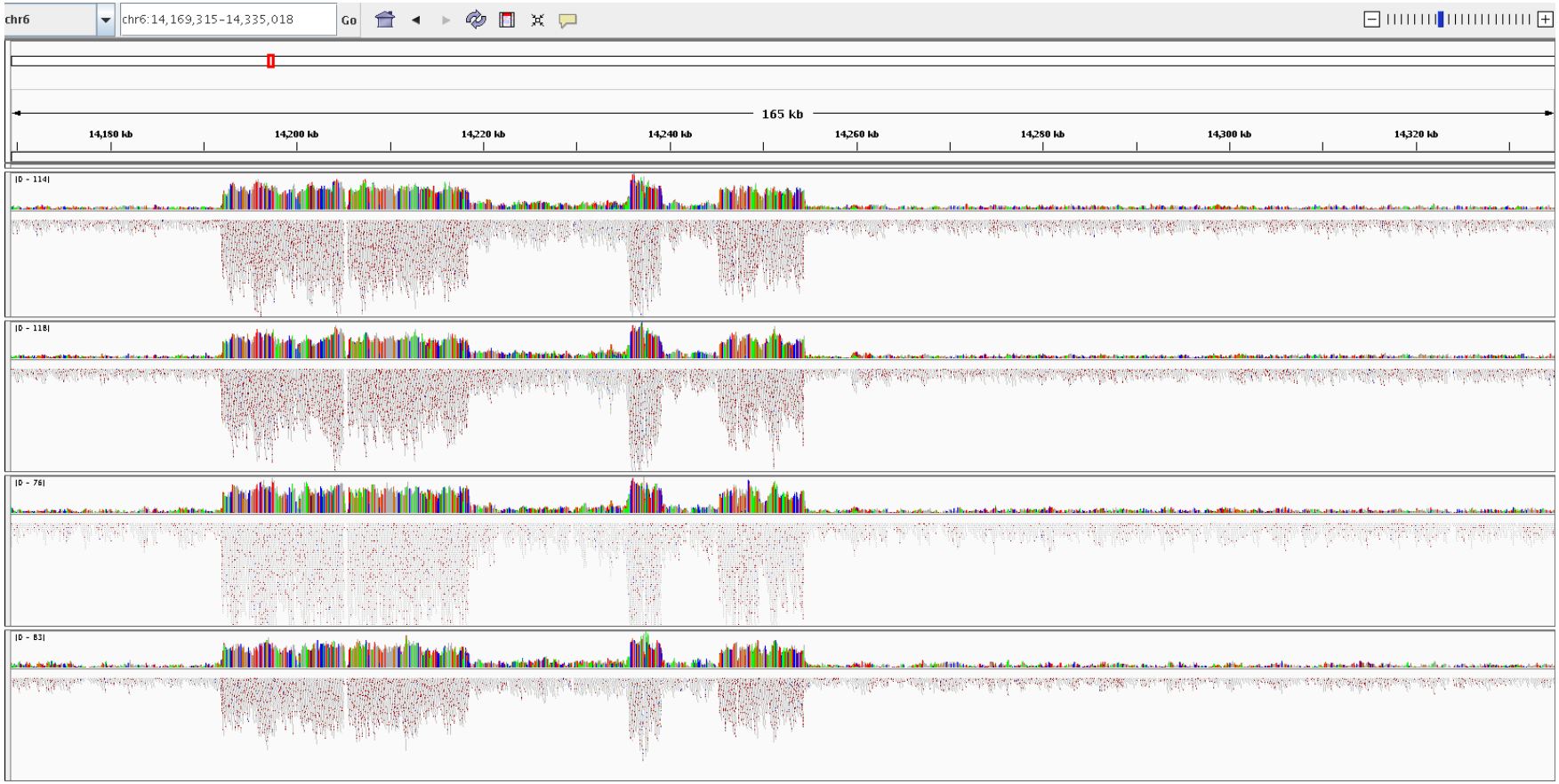


Figure C.14: Bowtie2 paired-end re-run alignment, Willy whole genome shotgun sequencing, samples ERR2286905 through ERR2286908.

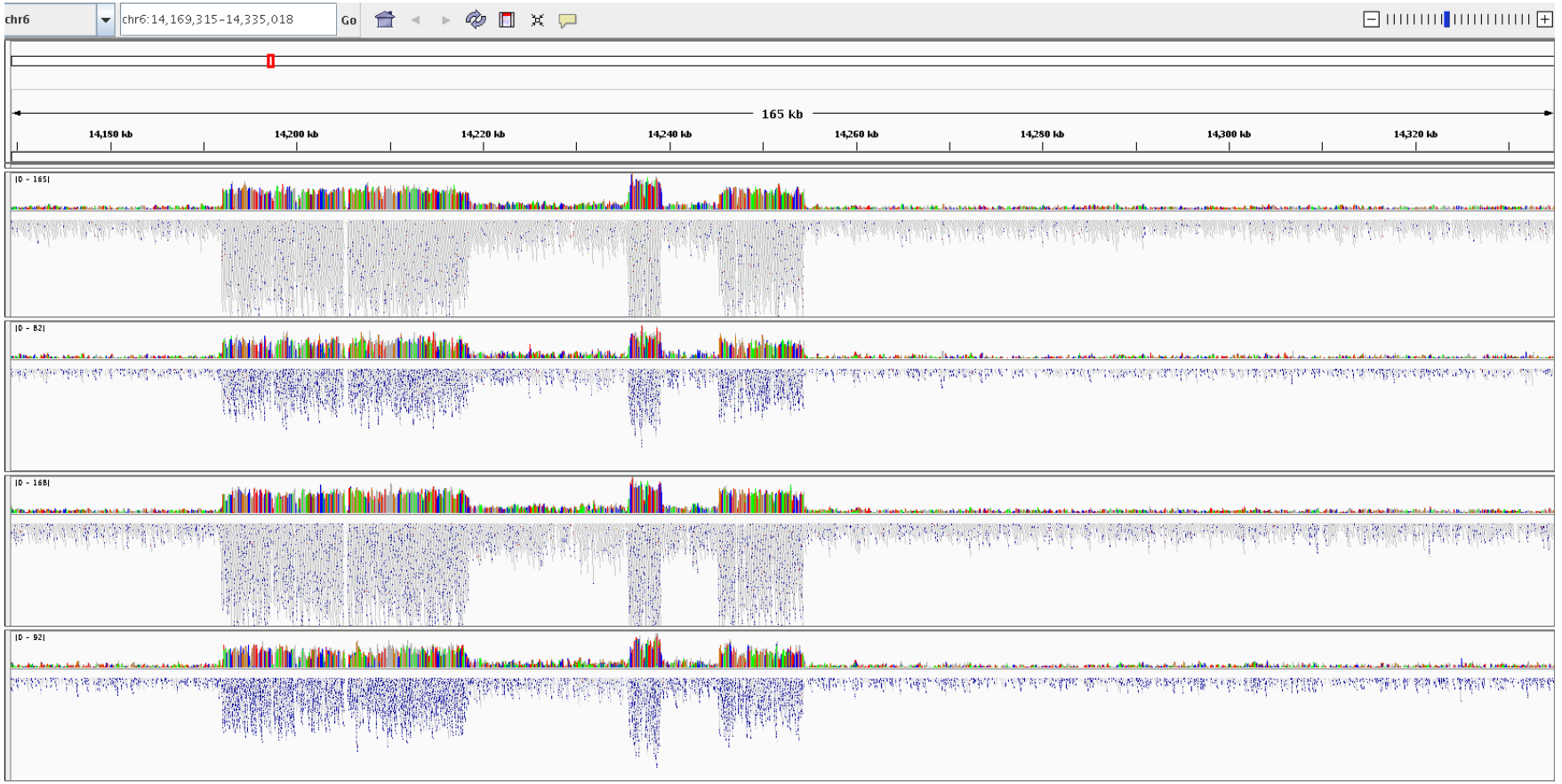


Figure C.15: Bowtie2 paired-end re-run alignment, Willy whole genome shotgun sequencing, samples ERR2286909 through ERR2286912.

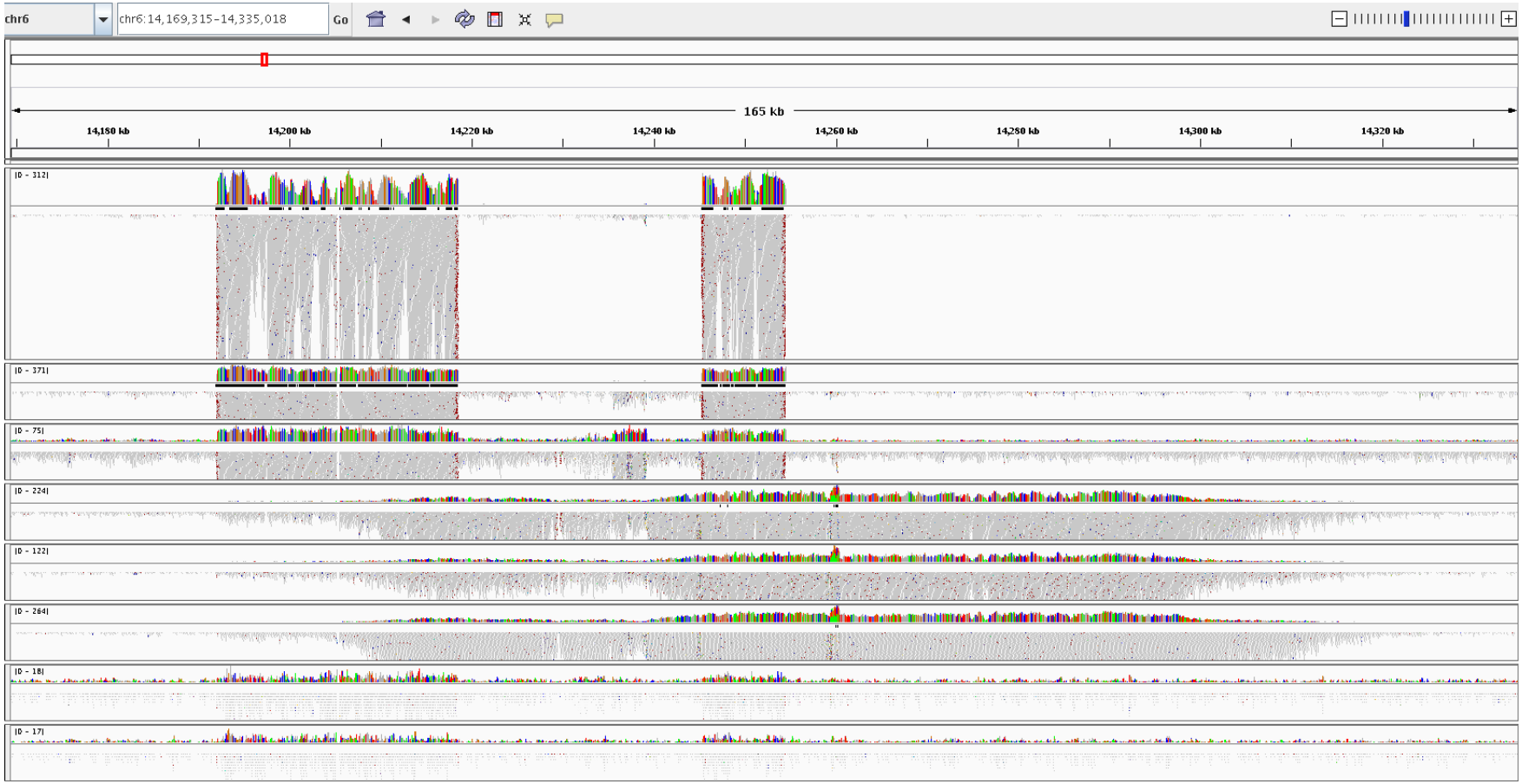


Figure C.16: BWA paired-end re-run alignment, Formaldehyde Cross-Linked ChIP-seq of Asino Nuovo and Blackjack CENP-A. Samples from top to bottom are Asino Nuovo SRR5515970 (ChIP-seq), SRR5515971 (ChIP-seq), SRR5516015 (input); and Blackjack SRR5515973 (ChIP-seq), SRR5515974 (ChIP-seq), SRR5515976 (ChIP-seq), SRR5515972 (input), and SRR5515975 (input).

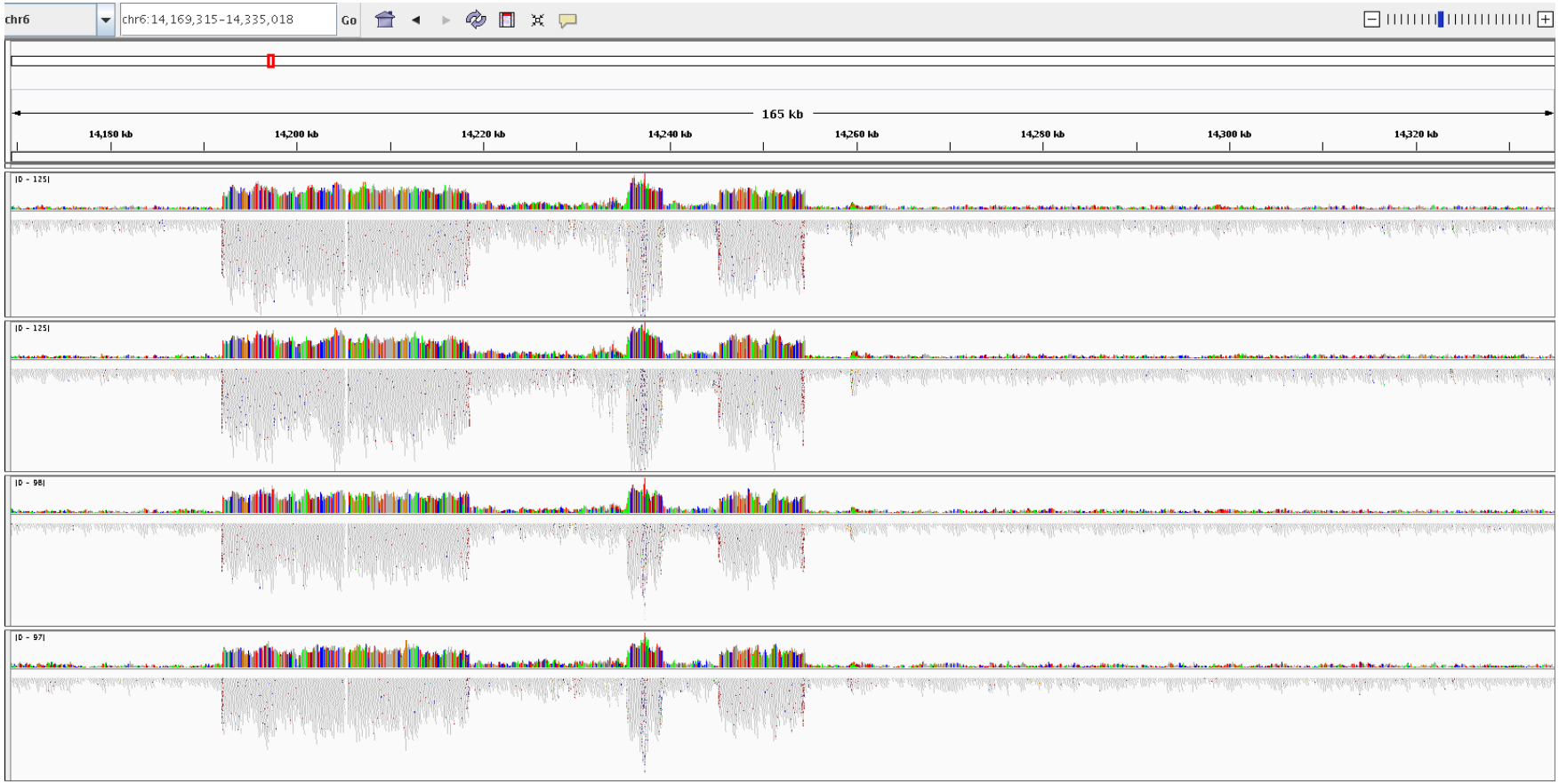


Figure C.17: BWA paired-end re-run alignment, Willy whole genome shotgun sequencing, samples ERR2286905 through ERR2286908.

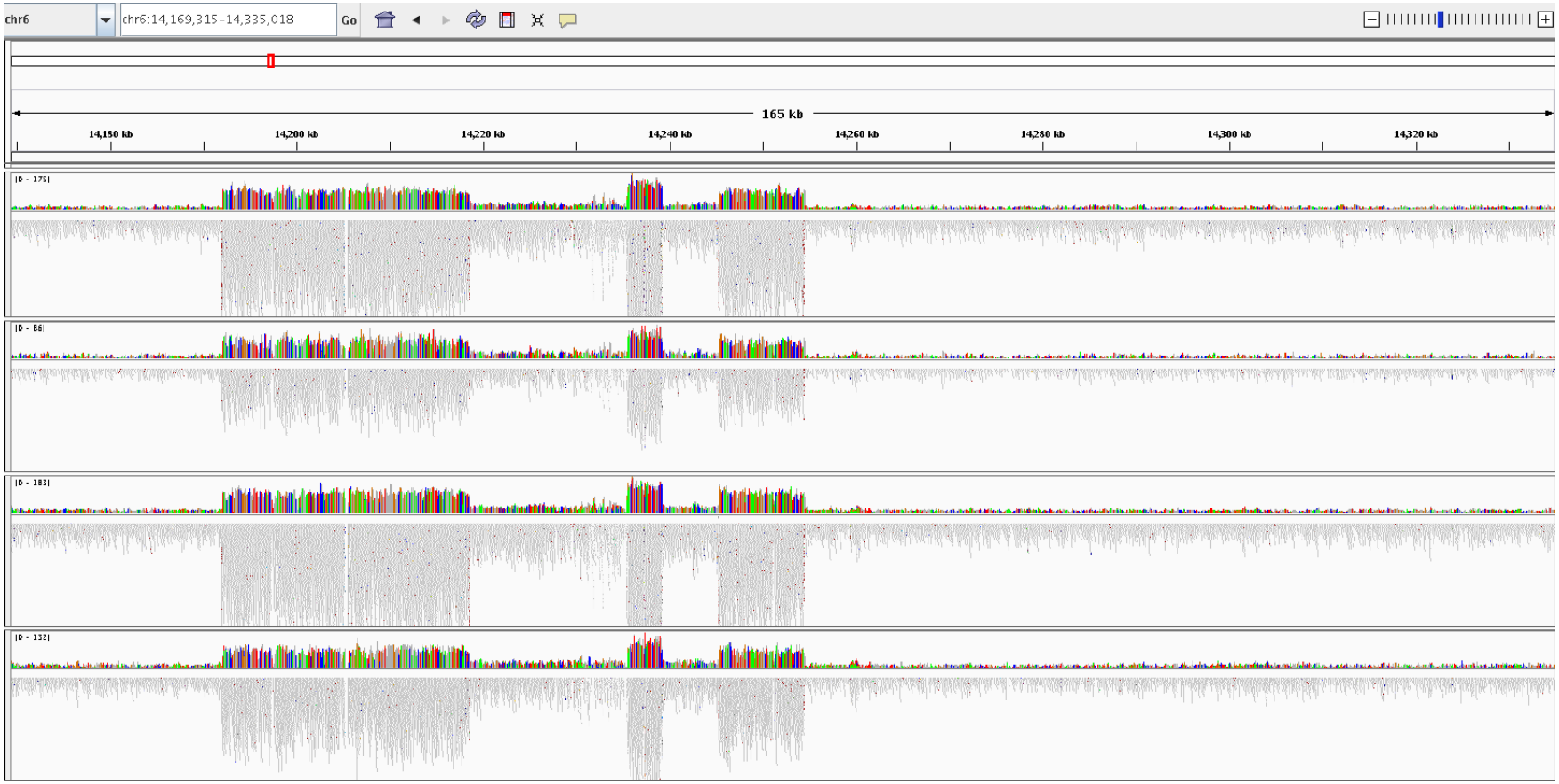


Figure C.18: BWA paired-end re-run alignment, Willy whole genome shotgun sequencing, samples ERR2286909 through ERR2286912.

Appendix D

Sequencing depth data

The following tables each show the depth of sequencing for each input or WGS sample aligned against horse chromosome 6. The depths of six regions were calculated. Then, read depth of the two regions identified as potential CNVs relative to the horse sequence were combined, and all non-CNV regions except the whole chromosome were combined. Average depth was calculated for each combined region, and the ratio of CNV region depth to non-CNV region depth was calculated.

Region	Start	Stop	Length	Depth	Mean	Ratio
Whole	0	84719076	84719076	183078303	2.161	
Start to 5'	0	14191980	14191980	30781264	2.169	
5'	14191981	14218611	26630	349746	13.134	
Deletion	14218612	14245307	26695	165248	6.19	
3'	14245308	14254527	9219	119896	13.005	
3' to End	14254528	84719076	70464548	151662136	2.152	
Non-Copy			84683223	182608648	2.156	
Copy			35849	469642	13.101	6.0753

Table D.1: Asino Nuovo Mononucleosome 1 Input Depth Stats

Region	Start	Stop	Length	Depth	Mean	Ratio
Whole	0	84719076	84719076	169925896	2.006	
Start to 5'	0	14191980	14191980	28561186	2.012	
5'	14191981	14218611	26630	342746	12.871	
Deletion	14218612	14245307	26695	152415	5.709	
3'	14245308	14254527	9219	105131	11.404	
3' to End	14254528	84719076	70464548	140764397	1.998	
Non-Copy			84683223	169477998	2.001	
Copy			35849	447877	12.493	6.2426

Table D.2: Asino Nuovo Mononucleosome 2 Input Depth Stats

Region	Start	Stop	Length	Depth	Mean	Ratio
Whole	0	84719076	84719076	182849982	2.158	
Start to 5'	0	14191980	14191980	30421467	2.144	
5'	14191981	14218611	26630	388018	14.571	
Deletion	14218612	14245307	26695	159133	5.961	
3'	14245308	14254527	9219	139311	15.111	
3' to End	14254528	84719076	70464548	151742031	2.153	
Non-Copy			84683223	182322631	2.153	
Copy			35849	527329	14.71	6.8322

Table D.3: Asino Nuovo Trinucleosome Input Depth Stats

Region	Start	Stop	Length	Depth	Mean	Ratio
Whole	0	84719076	84719076	65564063	0.774	
Start to 5'	0	14191980	14191980	9667914	0.681	
5'	14191981	14218611	26630	98842	3.712	
Deletion	14218612	14245307	26695	32936	1.234	
3'	14245308	14254527	9219	32425	3.517	
3' to End	14254528	84719076	70464548	55731936	0.791	
Non-Copy			84683223	65432786	0.773	
Copy			35849	131267	3.662	4.7389

Table D.4: Asino Nuovo CENP-C 1 Input Depth Stats

Region	Start	Stop	Length	Depth	Mean	Ratio
Whole	0	84719076	84719076	80870435	0.955	
Start to 5'	0	14191980	14191980	13373755	0.942	
5'	14191981	14218611	26630	175766	6.6	
Deletion	14218612	14245307	26695	65984	2.472	
3'	14245308	14254527	9219	63146	6.85	
3' to End	14254528	84719076	70464548	67191773	0.954	
Non-Copy			84683223	80631512	0.952	
Copy			35849	238912	6.664	6.9993

Table D.5: Asino Nuovo CENP-C 2 Input Depth Stats

Region	Start	Stop	Length	Depth	Mean	Ratio
Whole	0	84719076	84719076	121566983	1.435	
Start to 5'	0	14191980	14191980	19229625	1.355	
5'	14191981	14218611	26630	145059	5.447	
Deletion	14218612	14245307	26695	46469	1.741	
3'	14245308	14254527	9219	45669	4.954	
3' to End	14254528	84719076	70464548	102100155	1.449	
Non-Copy			84683223	121376249	1.433	
Copy			35849	190728	5.32	3.7119

Table D.6: Blackjack SRR5515972 Depth Stats

Region	Start	Stop	Length	Depth	Mean	Ratio
Whole	0	84719076	84719076	105234540	1.242	
Start to 5'	0	14191980	14191980	15954785	1.124	
5'	14191981	14218611	26630	113263	4.253	
Deletion	14218612	14245307	26695	35455	1.328	
3'	14245308	14254527	9219	34115	3.701	
3' to End	14254528	84719076	70464548	89096915	1.264	
Non-Copy			84683223	105087155	1.241	
Copy			35849	147378	4.111	3.3129

Table D.7: Blackjack SRR5515975 Depth Stats

Region	Start	Stop	Length	Depth	Mean	Ratio
Whole	0	84719076	84719076	818269425	9.659	
Start-¿5’	0	14191980	14191980	126450618	8.910	
5’	14191981	14218611	26630	1513266	56.826	
Deletion	14218612	14245307	26695	488006	18.281	
3’	14245308	14254527	9219	444039	48.166	
3’-¿End	14254528	84719076	70464548	689373413	9.783	
Non-Copy			84683223	816312037	9.640	
Copy			35849	1957305	54.599	5.6640

Table D.8: Asino Nuovo SRR5516015 Depth Stats

Region	Start	Stop	Length	Depth	Mean	Ratio
Whole	0	84719076	84719076	717876594	8.474	
Start to 5’	0	14191980	14191980	120177842	8.468	
5’	14191981	14218611	26630	1697399	63.74	
Deletion	14218612	14245307	26695	699940	26.22	
3’	14245308	14254527	9219	529078	57.39	
3’ to End	14254528	84719076	70464548	594772246	8.441	
Non-Copy			84683223	715650028	8.451	
Copy			35849	2226477	62.107	7.3492

Table D.9: Willy ERR2286905 Depth Stats

Region	Start	Stop	Length	Depth	Mean	Ratio
Whole	0	84719076	84719076	732817007	8.65	
Start to 5’	0	14191980	14191980	123156426	8.678	
5’	14191981	14218611	26630	1638810	61.54	
Deletion	14218612	14245307	26695	712737	26.699	
3’	14245308	14254527	9219	532633	57.776	
3’ to End	14254528	84719076	70464548	606776318	8.611	
Non-Copy			84683223	730645481	8.628	
Copy			35849	2171443	60.572	7.0204

Table D.10: Willy ERR2286906 Depth Stats

Region	Start	Stop	Length	Depth	Mean	Ratio
Whole	0	84719076	84719076	491750644	5.804	
Start to 5'	0	14191980	14191980	82936058	5.844	
5'	14191981	14218611	26630	1187807	44.604	
Deletion	14218612	14245307	26695	448610	16.805	
3'	14245308	14254527	9219	366401	39.744	
3' to End	14254528	84719076	70464548	406811700	5.773	
Non-Copy			84683223	490196368	5.789	
Copy			35849	1554208	43.354	7.4896

Table D.11: Willy ERR2286907 Depth Stats

Region	Start	Stop	Length	Depth	Mean	Ratio
Whole	0	84719076	84719076	505086493	5.962	
Start to 5'	0	14191980	14191980	85003801	5.99	
5'	14191981	14218611	26630	1164039	43.712	
Deletion	14218612	14245307	26695	477532	17.888	
3'	14245308	14254527	9219	381902	41.426	
3' to End	14254528	84719076	70464548	418059154	5.933	
Non-Copy			84683223	503540487	5.946	
Copy			35849	1545941	43.124	7.2524

Table D.12: Willy ERR2286908 Depth Stats

Region	Start	Stop	Length	Depth	Mean	Ratio
Whole	0	84719076	84719076	1041989621	12.299	
Start to 5'	0	14191980	14191980	176329607	12.425	
5'	14191981	14218611	26630	2224660	83.54	
Deletion	14218612	14245307	26695	997370	37.362	
3'	14245308	14254527	9219	747357	81.067	
3' to End	14254528	84719076	70464548	861690500	12.229	
Non-Copy			84683223	1039017477	12.269	
Copy			35849	2972017	82.904	6.7569

Table D.13: Willy ERR2286909 Depth Stats

Region	Start	Stop	Length	Depth	Mean	Ratio
Whole	0	84719076	84719076	455306441	5.374	
Start to 5'	0	14191980	14191980	76840264	5.414	
5'	14191981	14218611	26630	942121	35.378	
Deletion	14218612	14245307	26695	424528	15.903	
3'	14245308	14254527	9219	296649	32.178	
3' to End	14254528	84719076	70464548	376802831	5.347	
Non-Copy			84683223	454067623	5.362	
Copy			35849	1238770	34.555	6.4445

Table D.14: Willy ERR2286910 Depth Stats

Region	Start	Stop	Length	Depth	Mean	Ratio
Whole	0	84719076	84719076	1153180639	13.612	
Start to 5'	0	14191980	14191980	194467854	13.703	
5'	14191981	14218611	26630	2367414	88.9	
Deletion	14218612	14245307	26695	1085691	40.67	
3'	14245308	14254527	9219	816750	88.594	
3' to End	14254528	84719076	70464548	954442779	13.545	
Non-Copy			84683223	1149996324	13.58	
Copy			35849	3184164	88.822	6.5406

Table D.15: Willy ERR2286911 Depth Stats

Region	Start	Stop	Length	Depth	Mean	Ratio
Whole	0	84719076	84719076	543568397	6.416	
Start to 5'	0	14191980	14191980	91302211	6.433	
5'	14191981	14218611	26630	1112101	41.761	
Deletion	14218612	14245307	26695	502789	18.835	
3'	14245308	14254527	9219	370468	40.185	
3' to End	14254528	84719076	70464548	450280784	6.390	
Non-Copy			84683223	542085784	6.401	
Copy			35849	1482569	41.356	6.4605

Table D.16: Willy ERR2286912 Depth Stats

Appendix E

Scripts

Scripts used for pipelining alignment, alignment post-processing, and alignment filtering are shown below. For other processes performed on multiple files, such as k-mer matching and CNV calling, the parallel queuing script described here was used to submit simple scripts containing the commands described in the methods section.

E.1 Reference genome indexing script

```
#!/bin/bash

# SGE OPTIONS=====================================

#$ -N Horse_Indexing
#$ -q all.q
#$ -cwd
#$ -v PATH
#$ -v LD_LIBRARY_PATH
#$ -v PYTHONPATH
#$ -S /bin/bash

# COMMANDS========================================
# BUILD HORSE INDEX-----------------------------
bowtie2-build -f equCab2.fa horse
bwa index -a bwtsw equCab2.fa
```

E.2 Parallel queuing script

The following is an example script used to submit multiple jobs to the HPCC simultaneously, massively reducing runtime compared to linear back-to-back processing of multiple files. Target files are then passed to the desired script as command line arguments.

```
#!/bin/bash

for file in *.sra
do
   qsub alignment_script.sh "$file"
done
```

E.3 Alignment script

The following script includes both BWA and Bowtie2 alignment. For use, one or the other was left commented out.

```
#!/bin/bash

# SGE OPTIONS====================================

#$ -N Alignment_Script
#$ -q all.q
$ -cwd
#$ -v PATH
#$ -v LD_LIBRARY_PATH
#$ -v PYTHONPATH
#$ -S /bin/bash

# COMMANDS=======================================

# "$1" designates the first argument passed to this script on the
    command line, which should be an NCBI SRA archive file. This
    command stores the name of the SRA run, excluding its extension.
    Used to name and coordinate input and output files at each step.
name="${1%.*}"
# SRAtoolkit fastq-dump converts SRA files to FASTQ files. The -F
    removes SRA-created read names, which are incompatible with
    paired-end alignment tools. The --split-files option produces 2
    separate files for reads and mates.
```

```
fastq-dump -F --split-files $1

# ALIGNMENT-----------------------------------------
file="$name"_1.fastq
pair="$name"_2.fastq

### BOWTIE ALIGNMENT. Comment out if using BWA
sam="$name".sam
bowtie2 -x horse -1 "$file" -2 "$pair" -S "$sam"
###

### BWA ALIGNMENT. Uncomment following 3 lines to use BWA
# name="$name"_bwa
# sam="$name".sam
# bwa mem -t 8 equCab2.fa "$file" "$pair" > "$sam"
###

# SAM TO BAM CONVERSION ------------------------------
bam="$name".bam
samtools view -Sb "$sam" > "$bam"

# SORT ALIGNMENTS BY NAME-----------------------------
sorted="$name"_namesorted.bam
samtools sort -n -m 10000000000 "$bam" "${sorted/.bam}"

# FIX MATE FLAGS--------------------------------------
fixed="$name"_fixed.bam
samtools fixmate "$sorted" "$fixed"

# SORT ALIGNMENTS BY CHR AND COORDS-------------------
chr_sort="$name"_chrsorted.bam
samtools sort -m 10000000000 "$fixed" "${chr_sort/.bam}"

# REMOVE PCR DUPLICATE READS--------------------------
rmdup="$name"_rmdup.bam
samtools rmdup "$chr_sort" "$rmdup"

# INDEX FINAL BAM FILES AND GENERATE MAPPING STATS------
samtools index "$rmdup"
stats="$name"_mapstats.txt
samtools flagstat "$rmdup" > "$stats"
```

E.4 Alignment Filtering

The following script was used to produce BAM files containing only alignments to horse chromosome 6 for the purposes of CNV calling.

```
#!/bin/bash

# SGE OPTIONS=====================================

#$ -N Cthulu_Log
#$ -q all.q
#$ -cwd
#$ -v PATH
#$ -v LD_LIBRARY_PATH
#$ -v PYTHONPATH
#$ -S /bin/bash

# COMMANDS=======================================

#name="${1/*\/}"
name="${1/_*}"

#echo "$1"
#echo "$name"
#exit 0

# SAM TO BAM
samtools view -h $1 > "$name"_post.sam

# EXTRACT SAM HEADER
samtools view -SH "$name"_post.sam > "$name".head

# ADD HD LINE TO NEW HEADER
head -n 1 "$name".head > "$name"_new.head

# ADD SEQUENCE LINE CHR6 TO NEW HEADER
grep "^@SQ.SN:chr6" "$name".head >> "$name"_new.head

# ADD NEW RG LINE TO NEW HEADER
echo "@RG  ID:$name SM:$name" >> "$name"_new.head

# ADD PG LINE TO NEW HEADER
grep "^@PG" "$name".head >> "$name"_new.head
```

```
# EXTRACT MATCHES TO CHR6 FROM SAM FILE
awk '$3 == "chr6" { print $0 }' "$name"_post.sam > "$name".chr6matches

# ADD NEW HEADER AND MATCHES TO NEW SAM FILE
cat "$name"_new.head "$name".chr6matches > "$name"_6.sam

# NEW BAM TO SAM
samtools view -Sb "$name"_6.sam > "$name"_6.bam
```

Bibliography

[1] F. G. Westhorpe and A. F. Straight. "The centromere: epigenetic control of chromosome segregation during mitosis". In: *Cold Spring Harb Perspect Biol* 7.1 (Nov. 2014), a015818.

[2] H. Maiato et al. "The dynamic kinetochore-microtubule interface". In: *J. Cell. Sci.* 117.Pt 23 (Nov. 2004), pp. 5461–5477.

[3] J. S. Verdaasdonk and K. Bloom. "Centromeres: unique chromatin structures that drive chromosome segregation". In: *Nat. Rev. Mol. Cell Biol.* 12.5 (May 2011), pp. 320–332.

[4] A. Musacchio and K. G. Hardwick. "The spindle checkpoint: structural insights into dynamic signalling". In: *Nat. Rev. Mol. Cell Biol.* 3.10 (Oct. 2002), pp. 731–741.

[5] A. R. Cutter and J. J. Hayes. "A brief review of nucleosome structure". In: *FEBS Lett.* 589.20 Pt A (Oct. 2015), pp. 2914–2922.

[6] K. Luger et al. "Crystal structure of the nucleosome core particle at 2.8 A resolution". In: *Nature* 389.6648 (Sept. 1997), pp. 251–260.

[7] K. F. Sullivan, M. Hechenberger, and K. Masri. "Human CENP-A contains a histone H3 related histone fold domain that is required for targeting to the centromere". In: *J. Cell Biol.* 127.3 (Nov. 1994), pp. 581–592.

[8] K. Yoda et al. "Human centromere protein A (CENP-A) can replace histone H3 in nucleosome reconstitution in vitro". In: *Proc. Natl. Acad. Sci. U.S.A.* 97.13 (June 2000), pp. 7266–7271.

[9] B. E. Black et al. "Centromere identity maintained by nucleosomes assembled with histone H3 containing the CENP-A targeting domain". In: *Mol. Cell* 25.2 (Jan. 2007), pp. 309–322.

[10] Y. Nechemia-Arbely et al. "Human centromeric CENP-A chromatin is a homotypic, octameric nucleosome at all cell cycle points". In: *J. Cell Biol.* 216.3 (Mar. 2017), pp. 607–621.

[11] B. E. Black et al. "Structural determinants for generating centromeric chromatin". In: *Nature* 430.6999 (July 2004), pp. 578–582.

[12] D. L. Bodor et al. “The quantitative architecture of centromeric chromatin”. In: *Elife* 3 (July 2014), e02137.

[13] M. E. Stellfox, A. O. Bailey, and D. R. Foltz. “Putting CENP-A in its place”. In: *Cell. Mol. Life Sci.* 70.3 (Feb. 2013), pp. 387–406.

[14] K. E. Gascoigne et al. “Induced ectopic kinetochore assembly bypasses the requirement for CENP-A nucleosomes”. In: *Cell* 145.3 (Apr. 2011), pp. 410–422.

[15] L. E. Voullaire et al. “A functional marker centromere with no detectable alpha-satellite, satellite III, or CENP-B protein: activation of a latent centromere?” In: *Am. J. Hum. Genet.* 52.6 (June 1993), pp. 1153–1163.

[16] D. R. Foltz et al. “The human CENP-A centromeric nucleosome-associated complex”. In: *Nat. Cell Biol.* 8.5 (May 2006), pp. 458–469.

[17] K. M. Smith et al. “Heterochromatin is required for normal distribution of Neurospora crassa CenH3”. In: *Mol. Cell. Biol.* 31.12 (June 2011), pp. 2528–2542.

[18] C. M. Wade et al. “Genome sequence, comparative analysis, and population genetics of the domestic horse”. In: *Science* 326.5954 (Nov. 2009), pp. 865–867.

[19] B. Vissel and K. H. Choo. “Human alpha satellite DNA–consensus sequence and conserved regions”. In: *Nucleic Acids Res.* 15.16 (Aug. 1987), pp. 6751–6752.

[20] H. F. Willard. “Chromosome-specific organization of human alpha satellite DNA”. In: *Am. J. Hum. Genet.* 37.3 (May 1985), pp. 524–532.

[21] J. S. Waye and H. F. Willard. “Chromosome-specific alpha satellite DNA: nucleotide sequence analysis of the 2.0 kilobasepair repeat from the human X chromosome”. In: *Nucleic Acids Res.* 13.8 (Apr. 1985), pp. 2731–2743.

[22] C. Tyler-Smith and W. R. Brown. “Structure of the major block of alphoid satellite DNA on the human Y chromosome”. In: *J. Mol. Biol.* 195.3 (June 1987), pp. 457–470.

[23] T. Fukagawa and W. C. Earnshaw. “The centromere: chromatin foundation for the kinetochore machinery”. In: *Dev. Cell* 30.5 (Sept. 2014), pp. 496–508.

[24] Shannon M. McNulty and Beth A. Sullivan. “Alpha satellite DNA biology: finding function in the recesses of the genome”. In: *Chromosome Research* (July 2018). ISSN: 1573-6849. DOI: 10.1007/s10577-018-9582-3.

[25] M. Jain et al. “Linear assembly of a human centromere on the Y chromosome”. In: *Nat. Biotechnol.* 36.4 (Apr. 2018), pp. 321–323.

[26] S. Henikoff and Y. Dalal. “Centromeric chromatin: what makes it unique?” In: *Curr. Opin. Genet. Dev.* 15.2 (Apr. 2005), pp. 177–184.

[27] B. A. Sullivan, M. D. Blower, and G. H. Karpen. “Determining centromere identity: cyclical stories and forking paths”. In: *Nat. Rev. Genet.* 2.8 (Aug. 2001), pp. 584–596.

[28] M. Kapoor et al. “The cenpB gene is not essential in mice”. In: *Chromosoma* 107.8 (Dec. 1998), pp. 570–576.

[29] S. Henikoff, K. Ahmad, and H. S. Malik. “The centromere paradox: stable inheritance with rapidly evolving DNA”. In: *Science* 293.5532 (Aug. 2001), pp. 1098–1102.

[30] O. J. Marshall et al. “Neocentromeres: new insights into centromere structure, disease development, and karyotype evolution”. In: *Am. J. Hum. Genet.* 82.2 (Feb. 2008), pp. 261–282.

[31] G. Montefalcone et al. “Centromere repositioning”. In: *Genome Res.* 9.12 (Dec. 1999), pp. 1184–1188.

[32] M. Rocchi et al. “Centromere repositioning in mammals”. In: *Heredity (Edinb)* 108.1 (Jan. 2012), pp. 59–67.

[33] R. Saffery et al. “Human centromeres and neocentromeres show identical distribution patterns of ¿20 functionally important kinetochore-associated proteins”. In: *Hum. Mol. Genet.* 9.2 (Jan. 2000), pp. 175–185.

[34] E. E. Eichler. “Repetitive conundrums of centromere structure and function”. In: *Hum. Mol. Genet.* 8.2 (Feb. 1999), pp. 151–155.

[35] D. J. Amor et al. “Human centromere repositioning ”in progress””. In: *Proc. Natl. Acad. Sci. U.S.A.* 101.17 (Apr. 2004), pp. 6542–6547.

[36] W. H. Shang et al. “Chickens possess centromeres with both extended tandem repeats and short non-tandem-repetitive sequences”. In: *Genome Res.* 20.9 (Sept. 2010), pp. 1219–1228.

[37] W. H. Shang et al. “Chromosome engineering allows the efficient isolation of vertebrate neocentromeres”. In: *Dev. Cell* 24.6 (Mar. 2013), pp. 635–648.

[38] D. P. Locke et al. “Comparative and demographic analysis of orang-utan genomes”. In: *Nature* 469.7331 (Jan. 2011), pp. 529–533.

[39] L. Carbone et al. “Evolutionary movement of centromeres in horse, donkey, and zebra”. In: *Genomics* 87.6 (June 2006), pp. 777–782.

[40] F. M. Piras et al. “Uncoupling of satellite DNA and centromeric function in the genus Equus”. In: *PLoS Genet.* 6.2 (Feb. 2010), e1000845.

[41] M. Ventura et al. “Evolutionary formation of new centromeres in macaque”. In: *Science* 316.5822 (Apr. 2007), pp. 243–246.

[42] D. R. Foltz et al. “Centromere-specific assembly of CENP-a nucleosomes is mediated by HJURP”. In: *Cell* 137.3 (May 2009), pp. 472–484.

[43] S. Purgato et al. “Centromere sliding on a mammalian chromosome”. In: *Chromosoma* 124.2 (June 2015), pp. 277–287.

[44] E. Giulotto, E. Raimondi, and K. F. Sullivan. “The Unique DNA Sequences Underlying Equine Centromeres”. In: *Prog. Mol. Subcell. Biol.* 56 (2017), pp. 337–354.

[45] J. G. W. McCarter. “Nucleosome organisation and CENP-C distribution at satellite-free centromeres in Equus asinus”. PhD thesis. National University of Ireland, Galway, Oct. 2016.

[46] E. A. Oakenfull and J. B. Clegg. “Phylogenetic relationships within the genus Equus and the evolution of alpha and theta globin genes”. In: *J. Mol. Evol.* 47.6 (Dec. 1998), pp. 772–783.

[47] L. Orlando. “Equids”. In: *Curr. Biol.* 25.20 (Oct. 2015), R973–978.

[48] F. M. Piras et al. “Phylogeny of horse chromosome 5q in the genus Equus and centromere repositioning”. In: *Cytogenet. Genome Res.* 126.1-2 (2009), pp. 165–172.

[49] J. Huang et al. “Corrigendum: Donkey genome and insight into the imprinting of fast karyotype evolution”. In: *Sci Rep* 5 (Dec. 2015), p. 17124.

[50] G. Renaud et al. “Improved de novo genomic assembly for the domestic donkey”. In: *Sci Adv* 4.4 (Apr. 2018), eaaq0392.

[51] L. Orlando et al. “Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse”. In: *Nature* 499.7456 (July 2013), pp. 74–78.

[52] N. H. Putnam et al. “Chromosome-scale shotgun assembly using an in vitro method for long-range linkage”. In: *Genome Res.* 26.3 (Mar. 2016), pp. 342–350.

[53] S. G. Nergadze et al. “Birth, evolution, and transmission of satellite-free mammalian centromeric domains”. In: *Genome Res.* 28.6 (June 2018), pp. 789–799.

[54] F. Yang et al. “Refined genome-wide comparative map of the domestic horse, donkey and human based on cross-species chromosome painting: insight into the occasional fertility of mules”. In: *Chromosome Res.* 12.1 (2004), pp. 65–76.

[55] P. Musilova et al. “Subchromosomal karyotype evolution in Equidae”. In: *Chromosome Res.* 21.2 (Apr. 2013), pp. 175–187.

[56] S. F. Altschul et al. “Basic local alignment search tool”. In: *J. Mol. Biol.* 215.3 (Oct. 1990), pp. 403–410.

[57] W. J. Kent. “BLAT–the BLAST-like alignment tool”. In: *Genome Res.* 12.4 (Apr. 2002), pp. 656–664.

[58] J. T. Robinson et al. “Integrative genomics viewer”. In: *Nat. Biotechnol.* 29.1 (Jan. 2011), pp. 24–26.

[59] H. Thorvaldsdottir, J. T. Robinson, and J. P. Mesirov. “Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration”. In: *Brief. Bioinformatics* 14.2 (Mar. 2013), pp. 178–192.

[60] A. R. Quinlan and I. M. Hall. “BEDTools: a flexible suite of utilities for comparing genomic features”. In: *Bioinformatics* 26.6 (Mar. 2010), pp. 841–842.

[61] Andrews S. *FastQC: a quality control tool for high throughput sequence data.* 2010. URL: `http://www.bioinformatics.babraham.ac.uk/projects/fastqc`.

[62] P. Ewels et al. “MultiQC: summarize analysis results for multiple tools and samples in a single report”. In: *Bioinformatics* 32.19 (Oct. 2016), pp. 3047–3048.

[63] Ben Langmead et al. “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome”. In: *Genome Biology* 10.3 (Mar. 4, 2009), R25. ISSN: 1474-760X. DOI: `10.1186/gb-2009-10-3-r25`.

[64] H. Li and R. Durbin. “Fast and accurate short read alignment with Burrows-Wheeler transform”. In: *Bioinformatics* 25.14 (July 2009), pp. 1754–1760.

[65] H. Li et al. “The Sequence Alignment/Map format and SAMtools”. In: *Bioinformatics* 25.16 (Aug. 2009), pp. 2078–2079.

[66] X. Wang et al. “CNVcaller: highly efficient and widely applicable software for detecting copy number variations in large populations”. In: *Gigascience* 6.12 (Dec. 2017), pp. 1–12.

[67] S. Rossel et al. “Domestication of the donkey: timing, processes, and indicators”. In: *Proc. Natl. Acad. Sci. U.S.A.* 105.10 (Mar. 2008), pp. 3715–3720.