Lab Notebook for Davenport Lab

Author - Kobie Kirven

Penn State University

Metagenomic Sex Calling Pipeline: 10-06-2021

Introduction:

The goal of this project was to develop a method to determine host genetic sex using shotgun metagenomic data.

Methods:

Downloading and Simulating Data

The human reference genome build GRCh38 was downloaded from NCBI using the script "download_reference_genome.sh." Because I needed to be able to simulate reads from male and female reference genomes, I created artificial diploid reference genomes for both sexes using "make_simulated_chromosomes.py." Next, 10,000 random, paired-end sequences were generated using wgsim with default parameters implemented in the script, "simulate_shotgun_reads.sh." Once the simulated reads were generated, they were aligned to the GRCh38 reference genome using Bowtie2 with the "--local" flag. The aligned reads were then sorted using "samtools sort." Next, the depths were calculated for each position in each chromosome using "samtools depth." The alignment, sorting, and depth commands were implemented in the "align_with_bowtie2.sh" script.

Calculating Average Depth of Coverage

To determine the average depth for each chromosome, the total depth for each chromosome was computed by summing the depths at each position and the total depth was divided by the length of that chromosome (implemented in "calculate_coverage_depth.py"). This process was repeated for data from both male and female mock genomes. The results of the X:Autosomal coverage ratios are shown in Figure 1.

Validating WGSIM Worked As Expected

After looking at the resuts of the X:Autosomal, average coverage ratio plots, it was was apparent that something may have not gone correctly in the process because the plots were identical. Because the X:X average coverage rato was one for both the male and female simulated genome, I had the idea that the wgsim may not have worked as I expected and possibly it got confused by having an input file that contained repeated FASTA IDs. To test this theory, I creted a test FASTA file named "wgsim_test.fa" in which the IDs were the same, but the sequences were entirely different: one was only A repeats and the other was only C repeats. Next, I used this FASTA file as a reference genome and simulated 10 reads total with wgsim (implemented in "wgsim_validation.sh"). After looking at the output of the wgsim test, I learned that I was wrong and wgsim does not care about the FASTA IDs. It generates reads from each sequence in the reference genome.

Inspiration From Zonkey

I next tried to figure out how the authors of the Zonkey pipeline calculated their coverage. After looking around in their GitHub repo, I found on one page that they caluculate hits as: “Sum of SE, PE_1, and PE_2 hits. Note that supplementary alignments, duplicates, reads that failed QC, secondary alignments, and unmapped reads are ignored.” Using this definition for hits, I wanted to see if we could get more accurate X:autosome ratios. I used the sam files generated by the "align_with_bowtie2.sh" script and selected only those alignments that met the criteria of:

- Not Unmapped
- Mate not unmapped
- Not a secondary alignment
- Not a supplementary alignment

Next, I counted the number of alignments that passed these filtering criteria and grouped the counts by the chromosome the reads mapped to. Then, for each chromosome, I divided the number of counts by the length of that chromosome. Finally, I divided the normalized coverage of the X chromosome by the normalized coverages for each of the autosomes. This entire process, using the adapted version of the Zonkey method, was done for both male and female artifical genomes, and the results are plotted in Figure 2.

Sequencing Depth on X:Autosomal Ratios

I next asked the question, does sequencing depth effect the X:Autosomal ratios? To test this, I ran the same pipeline for simulating shotgun read generation from an artifical genome, and I used the sequencing depths 100, 250, 500, 1000, 2500,

5000, and 7,500. To make the analysis more streamlined, I combined the pipeline into one script titled "read_depth_pipeline.sh." The results from varying the seqencing depth are plotted in Figure 3.

The next question became, does preforming multiple runs of the read simulation and chromosome mapping give wildly different coverage results? To answer this question, I simulated 2500 reads from the male and female mock genome respectively for 100 runs and plotted the results in Figure 4. Additionally, I ran the same repeated simulations at a sequencing depth of 5000 reads, and the results are shown in Figure 4.

Results:

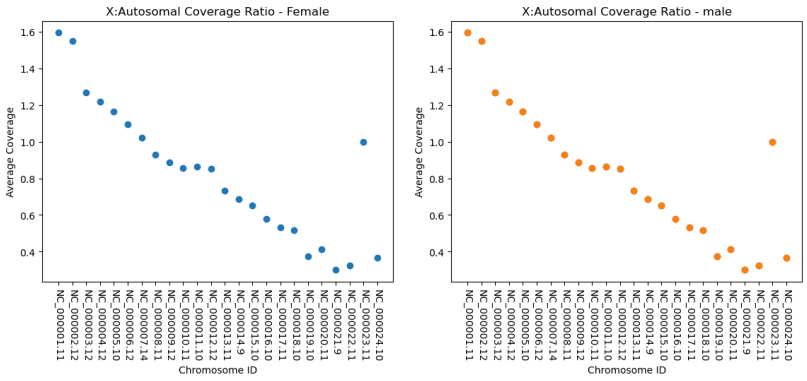


Figure 1. Plots of the average coverage of X chromosome divided by average coverage for each chromosome.

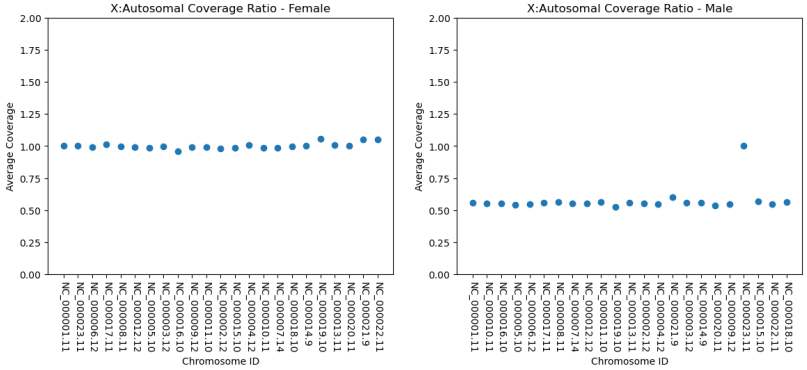
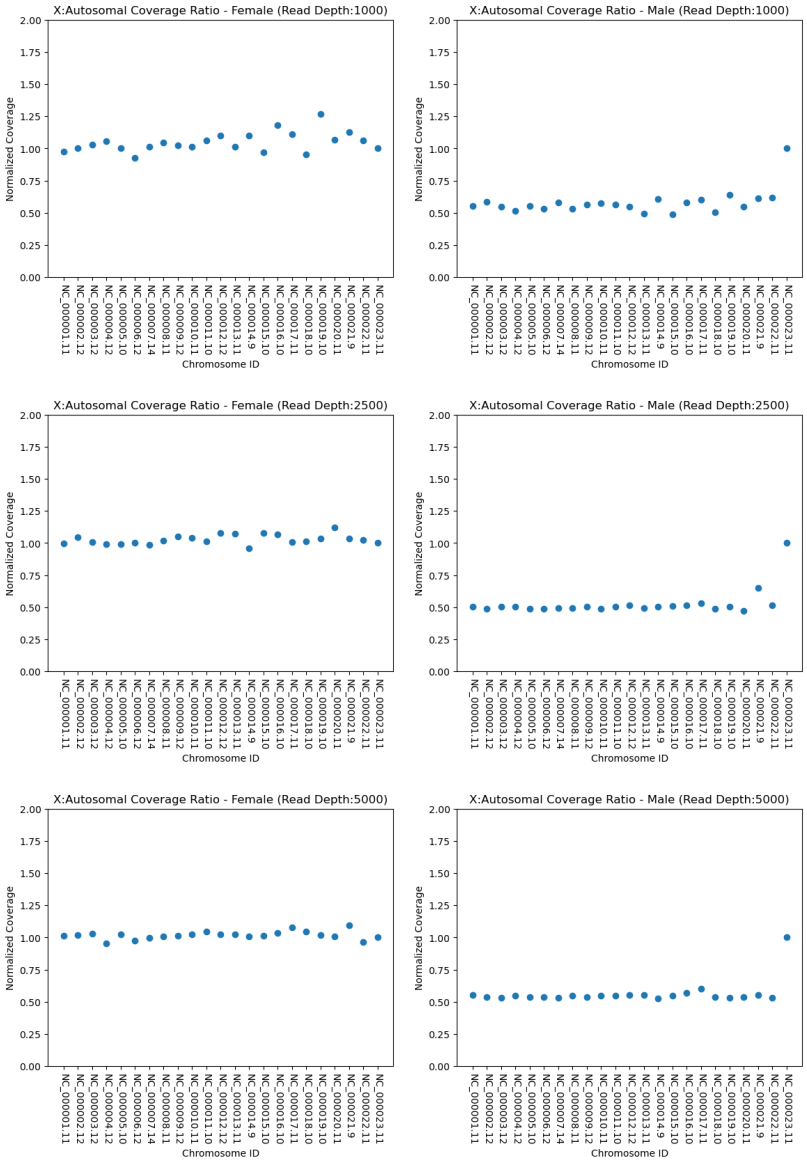


Figure 2. Plots of the normalized coverage of the X chromosome divided by normalized coverage for each chromosome.


X:Autosomal Coverage Ratio - Female (Read Depth:100)
X:Autosomal Coverage Ratio - Male (Read Depth:100)
X:Autosomal Coverage Ratio - Female (Read Depth:250)
X:Autosomal Coverage Ratio - Male (Read Depth:250)
X:Autosomal Coverage Ratio - Female (Read Depth:500)
X:Autosomal Coverage Ratio - Male (Read Depth:500)
Normalized Coverage
Chromosome ID


X:Autosomal Coverage Ratio - Female (Read Depth:1000)
X:Autosomal Coverage Ratio - Male (Read Depth:1000)
X:Autosomal Coverage Ratio - Female (Read Depth:2500)
X:Autosomal Coverage Ratio - Male (Read Depth:2500)
X:Autosomal Coverage Ratio - Female (Read Depth:5000)
X:Autosomal Coverage Ratio - Male (Read Depth:5000)
Normalized Coverage
Chromosome ID

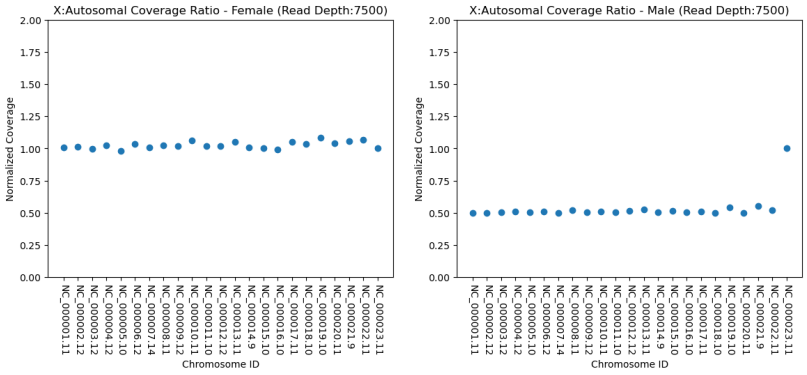


Figure 2. Plots of the normalized coverage of the X chromosome divided by normalized coverage for each chromosome with varying sequencing depth.

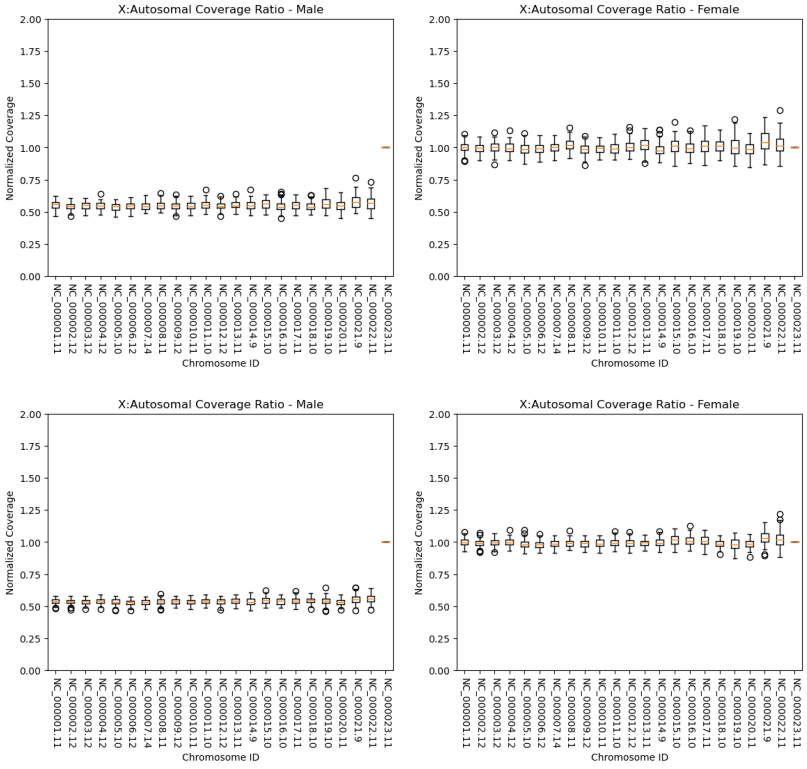


Figure 3. Box and whisker plots of normalized coverage for each autosome divided by the normalized coverage for X for 100 runs at a sequenceing depth of (top) 2500 and (bottom) 5000.

Discussion:

References:

1. Schubert, Mikkel, et al. "Zonkey: A simple, accurate and sensitive pipeline to genetically identify equine F1-hybrids in archaeological assemblages." Journal of Archaeological Science 78 (2017): 147-157.

Testing SCIMS on Human Data

Introduction:

The next question became, will scims actually work on real human data? To test this, I ran the pipeline on fecal metagenomic data from the TwinsUK study.

Materials and Methods:

Data Analysis

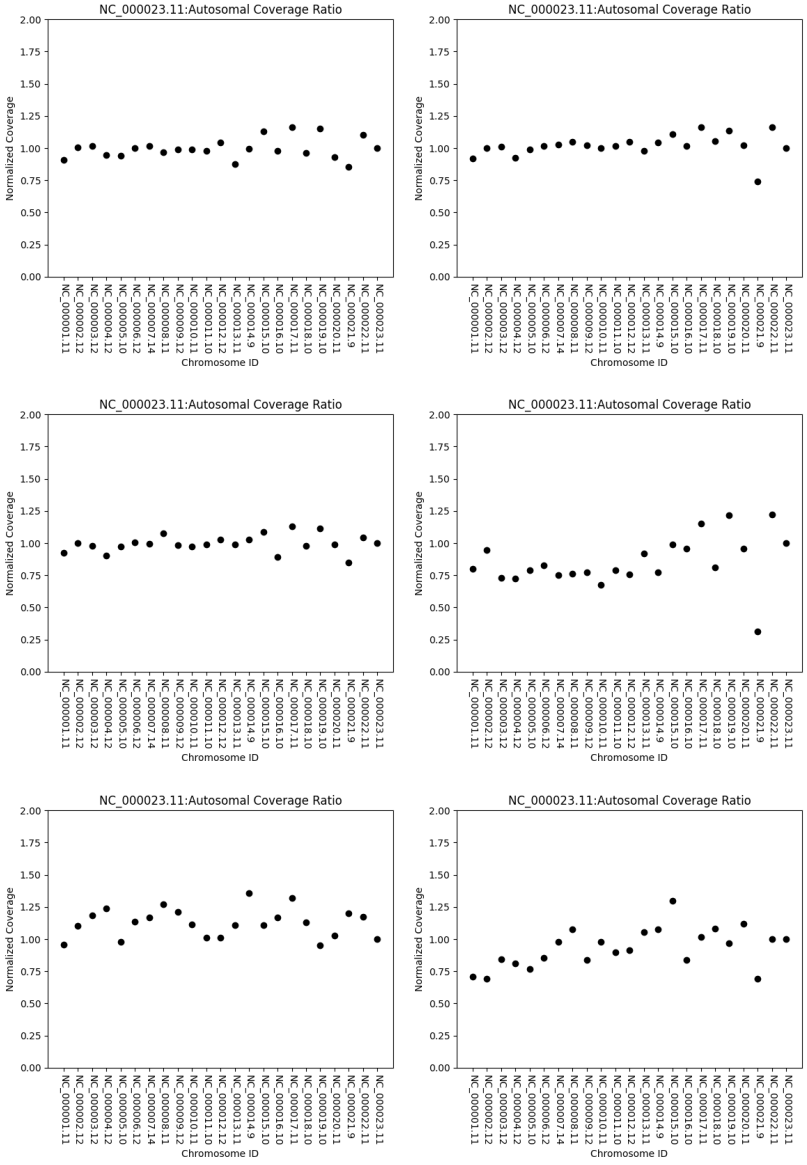
Data was transfered by Emily Davenport to the "/gpfs/group/exd44/default/data/TwinsUK/BGImetagenomes" directory. Next, the data were quality filtered usign trimmomatic with a sliding window of 4 with a minimum average quality of 30. After quality filtering, the paired reads that passed the filtering step were used as input for SCIMS. The process was implemented in bin/twinsuk/twinsuk.pbs, and the results are plotted in Figure 4.

The Twins UK samples from batch 2 were ran throught the same pipeline.

Chromosome Specific Coverages

In order to see if specific chromosomes had X:Autosomal ratios that differed from 1, the X:Autosomal ratio for each chromosomes was calculated accross all Twins UK samples, and the results were plotted in figure 5.

Results:


NC_000023.11:Autosomal Coverage Ratio
Normalized Coverage
Chromosome ID

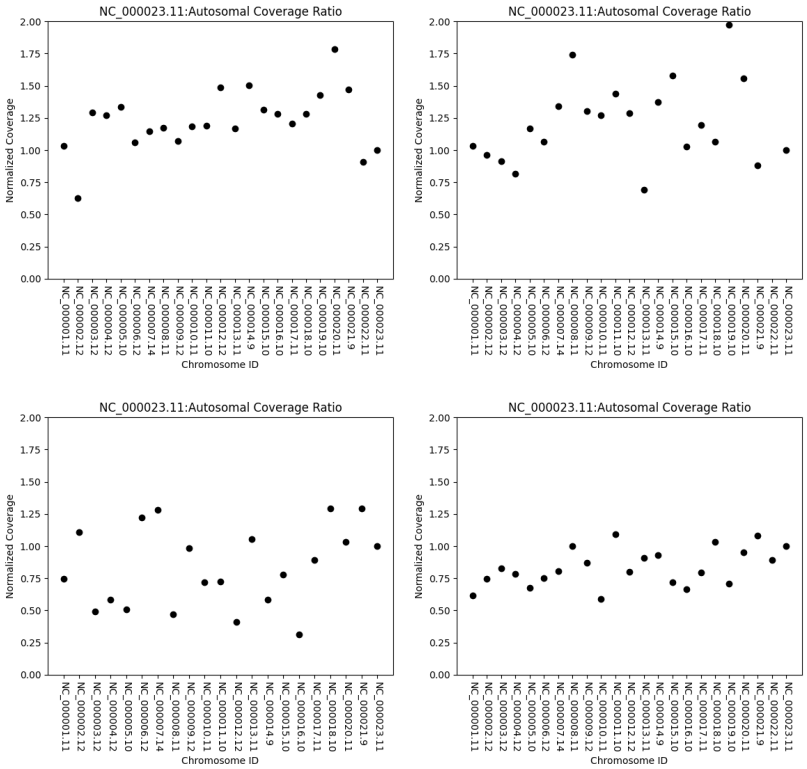
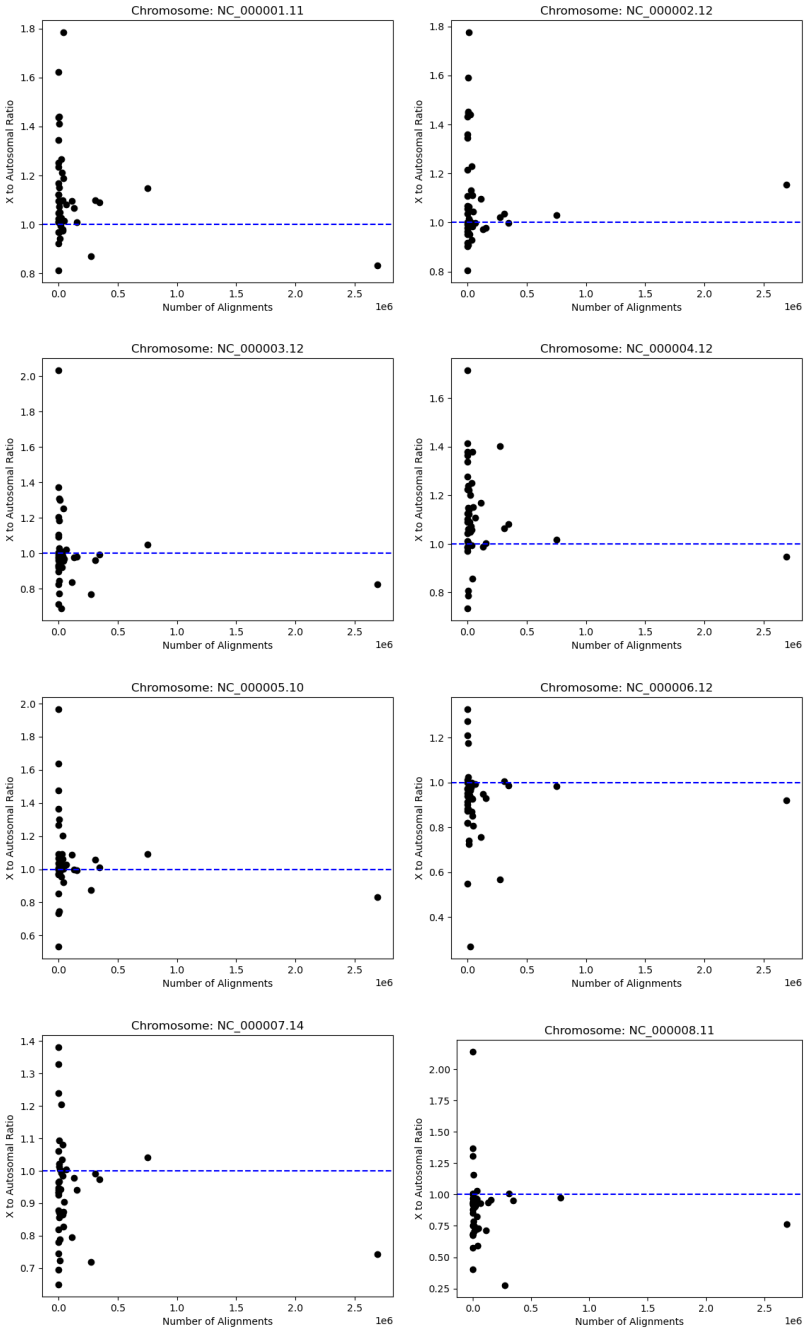
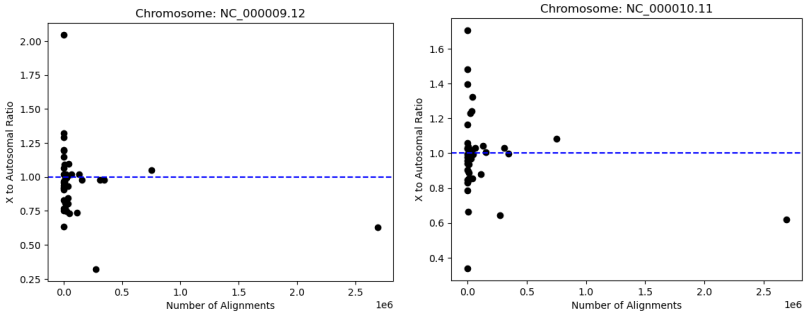


Figure 4. From left to right: BGI36049, BGI36050, BGI36702, BGI36703, BGI37181, BGI37182, BGI37201, BGI37202, BGI37268, BGI37269.

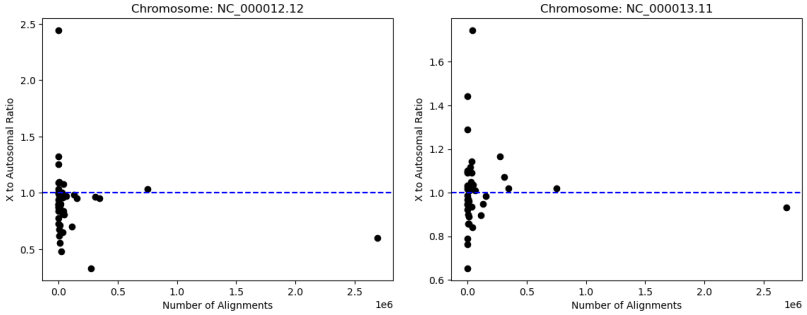
Table 1.

Sample Name	True Genetic Sex	Total # of Reads Mapping	Average X:Autosomal
36049	F	38285	0.998
36050	F	345092	1.021
36702	F	66208	0.997
36703	F	1352	0.847
37181	F	4905	1.133
37182	F	4084	0.932
37201	F	3632	1.237
37202	F	2290	1.324
37268	F	262	0.824
37269	F	2194	0.829

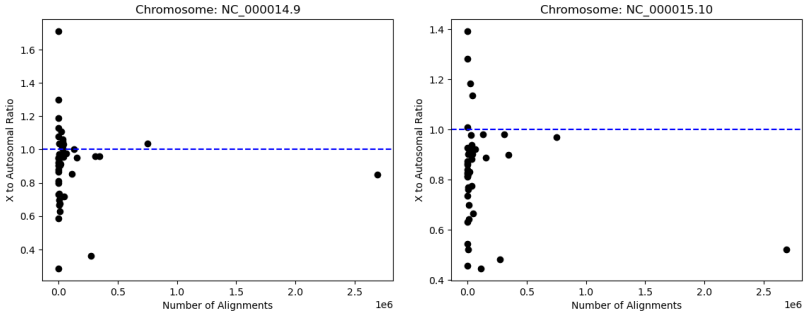

Chromosome: NC_000001.11
Chromosome: NC_000002.12
Chromosome: NC_000003.12
Chromosome: NC_000004.12
Chromosome: NC_000005.10
Chromosome: NC_000006.12
Chromosome: NC_000007.14
Chromosome: NC_000008.11
X to Autosomal Ratio
Number of Alignments
1e6


Chromosome: NC_000009.12
X to Autosomal Ratio
Number of Alignments
1e6

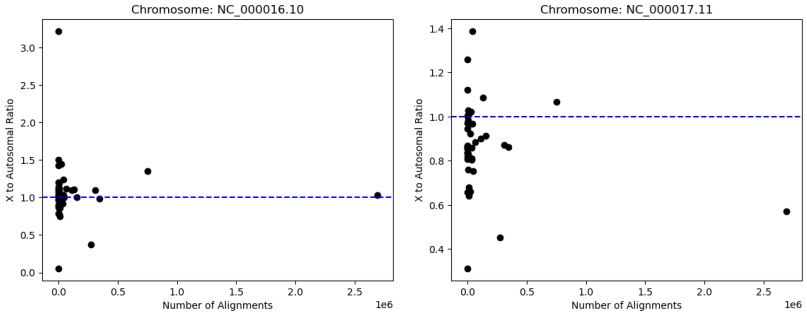
Chromosome: NC_000010.11
X to Autosomal Ratio
Number of Alignments
1e6


Chromosome: NC_000012.12
X to Autosomal Ratio
Number of Alignments
1e6

Chromosome: NC_000013.11
X to Autosomal Ratio
Number of Alignments
1e6


Chromosome: NC_000014.9
X to Autosomal Ratio
Number of Alignments
1e6

Chromosome: NC_000015.10
X to Autosomal Ratio
Number of Alignments
1e6


Chromosome: NC_000016.10
X to Autosomal Ratio
Number of Alignments
1e6

Chromosome: NC_000017.11
X to Autosomal Ratio
Number of Alignments
1e6

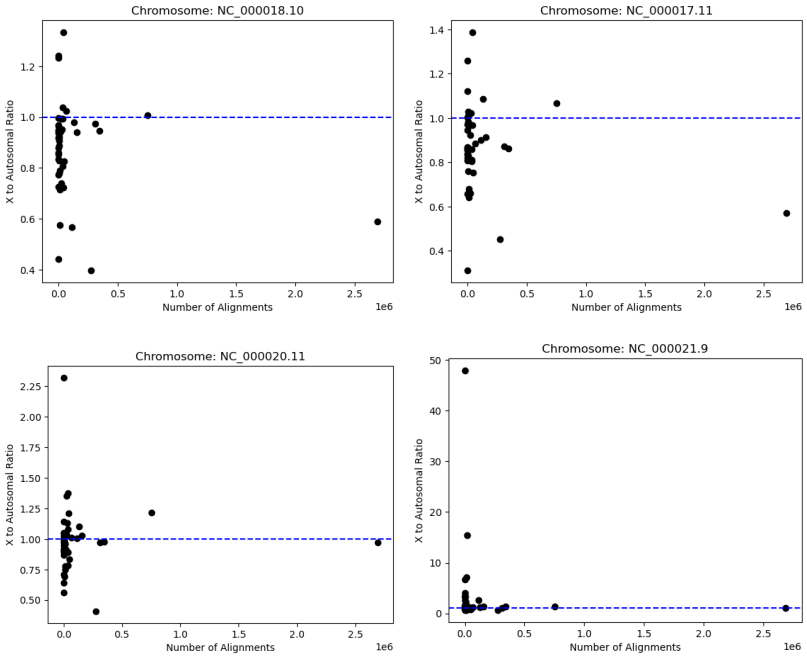


Figure 5. X:Autosomal ratio for each chromosome accross Twins UK samples. Blue line indicates and X:Autosomal ratio of 1.

SCiMS With Single-end Data