Preparatory project on Tragelaphine WGS data

Bioinformatics project 3 - University of Copenhagen

Thomas Bøggild

22/1 - 2024

Figure 1: Male Nyala observed in Kruger National Park in South Africa by Martyn Drabik-Hamshare

1 Introduction

This project is meant as a preparatory initiative to explore the data set that I will be using for my dissertation following this project, as well as ensuring the quality of the data. The data is from short read whole genome sequencing of a large quantity of individuals of various species from the tribe Tragelaphini from the subfamily Bovinae, or the spiral horned antelopes, of which the aim was a depth of coverage of 20-30X. The tribe includes 9 or 10 species, depending on whether the bushbuck is to be considered two separate species, of which all but one, the bongo, were sampled here. The single bushbuck sample, was discovered to be included only after analysis. These species are fairly morphologically distinct, at least among the males and have sizes ranging from 100kg to 1,200kg as well as distinct patterns of markings and horns. Part of the plan for my dissertation is the look into historical gene flow among species in the tribe and i will be starting that here while exploring the data. The tribe is distributed over large parts of sub-Saharan Africa, with distributions for each species overlapping in a lot of places and fertile cross species hybrids have been observed between multiple combinations of species within the tribe, which further opens up possibilities of historical gene flow.[1][2] There has also been found a lot of chromosomal rearrangements in relation to the cattle genome, including fusion between the Y chromosome and an autosome which could provide interesting problems to look into.[3]

2 Methods

The raw data for this project consists of 274 pairs of fastq files from short-read paired sequencing runs on a mix of in-house samples and a smaller proportion of loaned samples from various types of tissue. The initial samples consist of 87 samples of common eland, 15 of giant eland, 69 of greater kudu, 58 of lesser kudu, 5 of mountain nyala, 5 of nyala and 35 of sitatunga. The samples were named as a combination of their latin names, country of origin and a unique 4 digit ID number. Due to a mistake in the initial joining of reference tables for the samples, the nyala received the code "Scaf" instead of "Tang" due to missing information and a mixup with the buffalo, but i decided to keep the naming as i suspected that changing names of samples upstream might cause future complications for processes already started downstream. Information about samples has been collected in a master table, which can be viewed here

To map the data and thus be able to compare markers in population genetic analyses, and to ascertain the quality of the sequencing I used an in house pipeline, which with some partially handwritten input files and tweaking handles the running of a lot of different scripts and software via a combination of make and snakemake. Each species was run through the pipeline separately and a repository containing the pipeline can be found here, while my own scripts can be found here

The first step in the pipeline is the identification of sequencing adapters using AdapterRemoval, which finds a consensus adapter sequence by aligning reads under the assumption that the same

adapter was used for a single file/run and logging these for each sample, the sequence found can then be looked up and confirmed as a known sequence and specified in the input for subsequent steps to be trimmed from reads.[4]

The second and most computationally intensive step in the pipeline is the mapping of reads to a reference genome, two different ones, one closer in the taxonomy to the samples and one further away. This is done both to immediately compare the mapping effort on different references and also e.g. to have access to one reference map with a better annotation and one with closer resemblance to the chromosomal structure as the sample, later down the line. The mapping was done with bwa-mem2 in paleomix with the added options to mark bad reads instead of removing them.[5]

The next step after mapping the reads is filtering of the reads using a python script, keeping ones not marked as: PCR duplicates, bad insert sizes, bad mate orientation, mates mapping to different contigs, low mapping fraction or unmapped reads.

Next comes the main quality control of the data, the first step of which is simply using fastQC in concert with MultiQC to produce a quality report on the raw fastq read files before anything else was done. [6] [7] While this could be done before the mapping and filtering steps, the main utility is comparison with a report produced from the filtered dataset and is thus run together with the other qc steps in the pipeline. Following this, a number of different steps are run in parallel, collecting and plotting stats like depth distributions and insert sizes, and computing error rates with the "perfect individual" method implemented in ANGSD.[8] A PCA and a neighbour joining tree from identity by state distances is also made from genotypes called with some out-of-the-box default settings at this point. The program SATC is also run here, which tries to infer scaffolds linked to sex-chromosomes and assign sex to individuals.[9] Relatedness is estimated via 2d site frequency spectra computed from the possible pairs of individuals, which can then be used to calculate the king statistic between pairs of individuals. After looking over the results, a second round of the same QC analyses were run with individuals picked out based if they looked to have something wrong with the sample, primarily based on error rates, GC content and depth, to make sure the overall picture looked good after removing the possible skewing influence of these "bad" samples.

After checking the quality of samples, removing samples considered unfit for analysis and producing bam files of mapped reads to both references I went on to call genotypes for the different species together. For this project I decided to only use the Bos taurus reference mapped data for the time being, as this reference is more of a known quantity, equally distant to all tragelaphines and since the reduction in depth of coverage was only around 1X considering cattle is further removed than the bongo. To call genotypes i used another in house snakemake utilizing bcftools with a minimum cutoff on base quality of 25 and mapping quality of 30 and afterwards filtering out indels, invariable sites and sites not in a special mask file, the reason for the first two being that indels often need special assumptions to work with and invariable sites don't really contain any useful information for population genetics (except their total in some cases).[10] The mask of usable sites was derived from sites not marked as in repeat regions in the reference FASTA, found via RepeatMasker which

does this by matching known repetitive elements from the specified group of organism against the input sequence, in this case the reference. I only used the scaffolds marked as autosomes in the cattle reference, filtering out allosomes and unplaced scaffolds, while some of these could potentially be interesting in relation to exploring chromosomal rearrangements and possible sex linkage, they could also be a risk to include for some more general analyses. I also used an existing python script to first plot the distribution of depths across all sites for the 2hd set, which revealed some spikes at both ends of the spectrum, both very high and very low depth. The low end sites are likely to be in regions where it is exceedingly difficult to map, which could mean different potentially problematic things such as the reference conflicting with the sampled species (and besides these sites wont contain much useful information anyway due to the low depth) whereas the very high depth sites could be in a repeat region not caught by RepeatMasker causing a lot of reads to map in the same place and thus a cut off at both ends was set to filter out these sites. When calling genotypes i designated two different sets of individuals with their own VCF file, one smaller set of just the two highest depth unproblematic individuals from each species which i called 2hd and a main set consisting of all individuals not marked as too problematic. This was done to have a small and "safe" set to use as a test for tweaking parameters when running analyses before using the larger (and therefore often slower) main set.

While the produced vcf files already had some filters applied, I wrote an R script to compute and plot the 2 dimensional site frequency spectra between species to see if i could identify any other issues.[11] This led to the implementation, also in R, of another filter removing all sites which were polymorphic in more than 3 species, with the reasoning being that this would be unlikely as a true signal, since the site would likely have gone to fixation after being shared since a split basal enough to include 4 or more species and would be unlikely to be shared due to interspecies geneflow. The hope was that this would remove a problem spotted in the 2d sfs's of spikes of frequency observed in the middle of the spectra, meaning sites with half of derived and ancestral alleles for both species, which not only break the expected curve of the sfs, but also are likely to be sites where two different parts of a species genome overlap or map to the same point in the reference, causing all individuals to look heterozygous in that site.

From the called and filtered genotypes i used plink with the –distance flag to compute an identity by state matrix between all individuals in the main data set.[12] I then used R package APE to convert this matrix into a neighbour joining tree which could give an easily interpretable first estimate of taxonomic relationships between sampled species and more importantly single out individuals which did not cluster with their expected group. [13]

To confirm suspicions of wrongly marked samples from the 2D sfs's and the IBS tree, i wrote a small shell script extracting the consensus sequence for each individual of the reference mitochondrial DNA in fasta format using ANGSD, and used nBlast to find the closest matching species. The mtDNA was used since it is more conserved and because the database is likely to include a good reference sequence to match against for all relevant species.

To get an idea of how similarity between species varies across different parts of the genome, whether due to gene flow or ILS, i tried generating separate trees from random homolog sequences, initially using 25 kilobase windows and then upping the size to 100kb after seeing the level of what could potentially be noise in the first run. I wrote another shell script for this task, first using bedtools to extract a random segment of the genome within the previously mentioned filter or mask of usable sites of a set length.[14] This segment then had a consensus sequence extracted for a single highest depth individual from each species sampled via ANGSD and since these FASTA sequences were mapped to the same reference, they constitute a multiple sequence alignment. This alignment was then passed into RAxML to estimate a tree under the GTR+GAMMA model and the entire process was repeated 1000 times, resulting in a file containing 1000 different trees in newick format.[15] The first use of these trees was then to use Astral to estimate a consensus tree since it is made for this type of task of finding a consensus between incongruent trees. It does this with a heuristic approach, where it takes all possible quartet trees from each input tree and searches for the tree that matches the most possible quartets from the differing input trees. [16] To get a better understanding of how the different trees disagreed and where the inferred topology was weakest i also used the program DiscoVista, specifically the relative frequency analysis mode, which produces a labeled tree based on a specified input consensus tree and corresponding distributions of frequencies of possible topology around focal internal branches again based on these quartets.[17] Each quartet in each input tree can only agree with one of the configurations around a focal branch and thus these can be used to calculated a frequency of how often the topology appears.

To get a better understanding of where the proposed topology of the taxonomic tree of the clade could have trouble explaining the observed data, i used Dsuite to compute D-statistics for various possible trios of trees with the waterbuck as an outgroup.[18] To ease the computational load of the D-statistic calculations i did some further filtering on the VCF file used. This filtering is easily implemented and takes advantage of the fact that a lot of sites are not informative when counting ABBA and BABA sites, namely the ones where the alternative allele is only observed in one species, since those sites can't be either ABBA, BABA or BBAA. This filter removed more than 5/6 of the sites and thus improved the runtime a good deal. After computing the trios i used the output with the fbranch subprogram to generate a plot showing f_b values in relation to the topology given as input, in this case two alternative topologies found to be most likely earlier.

3 Results

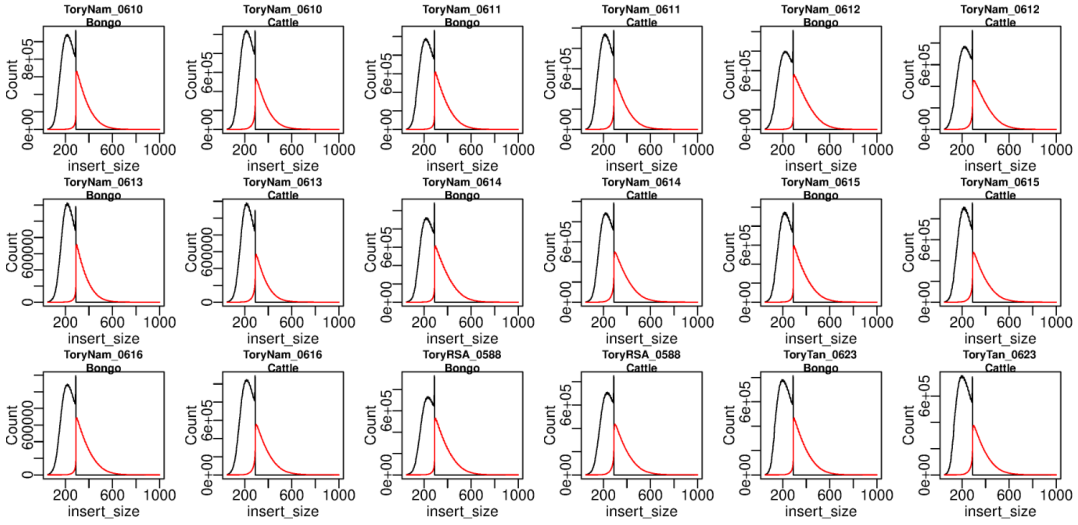

Figure 2: Example of insert size distributions, with merged mate pairs in black and paired ones in red

Some of the first results i looked at from the QC were the plots of the depth distributions, to see if we actually got the targeted depth in sequencing. This was not quite the case and while the distributions looked fine, the mean depths were mostly in the range of 12-14 (the exact mean depths for the data mapped to the bongo reference can be found in the master table), with values around 1x lower for the cow reference across the board. The likely culprit for this lower depth than the target could be found in the plots of insert size distributions, of which a representative example can be seen in figure 2. This pattern of most of the distribution lying below the 300 mark means that insert sizes were so small in most cases that reads had to be merged and thus we lose depth in the overlapping regions, as the mates contain the same information in the overlapping part. The pattern was the same or worse across all species and samples.

Most error rates looked fine at around 1e-4 to 5e-4 like the example shown for eland in figure 3, and while some were much higher than the norm, these corresponded with samples with other problems like high GC content, as shown in figure 5, or very low depth in general, which can be assumed to be the result of bad tissue samples, contamination or as in the case of 5 greater kudu samples, fecal matter samples which had ended up being sequenced. This was with the exception of a couple of cases like TstrZam_0843, where the only red flag was a high error rate and weird placement in PCA and IBS tree. These cases were resolved later when the mtDNA of the samples were blasted and found to most likely be a mixup of what species the sample was marked as at some point.

As for relatedness, some pairs of sampled had king statistics around or above 0.25 and these pairs were marked as such in the master table so that one in each pair can be picked out later in analyses where this matters, but the information was not used otherwise in this project.

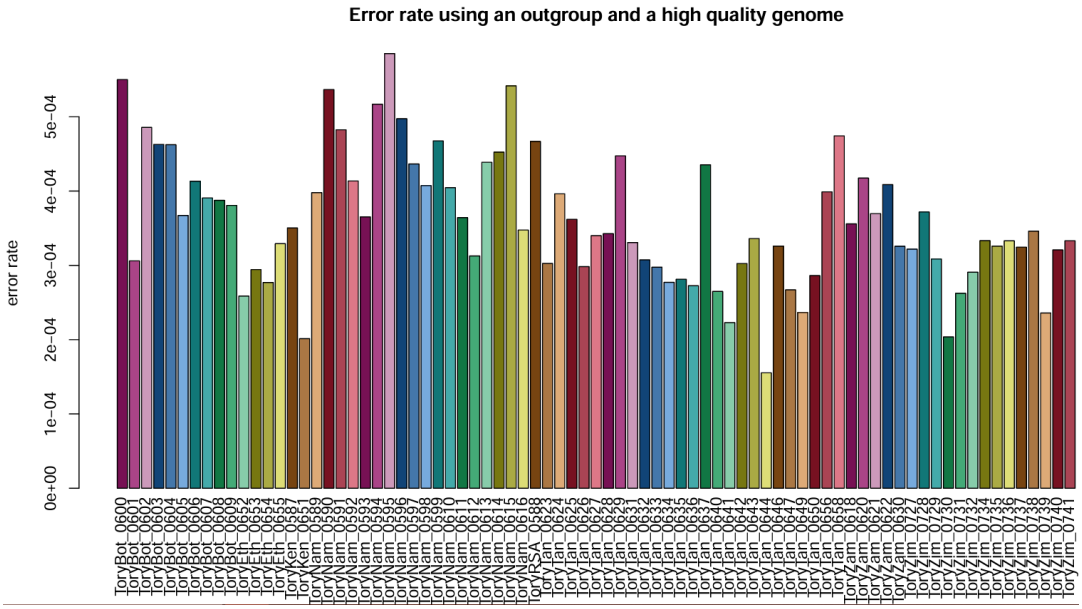

Figure 3: Error rates for second round of QC for eland samples

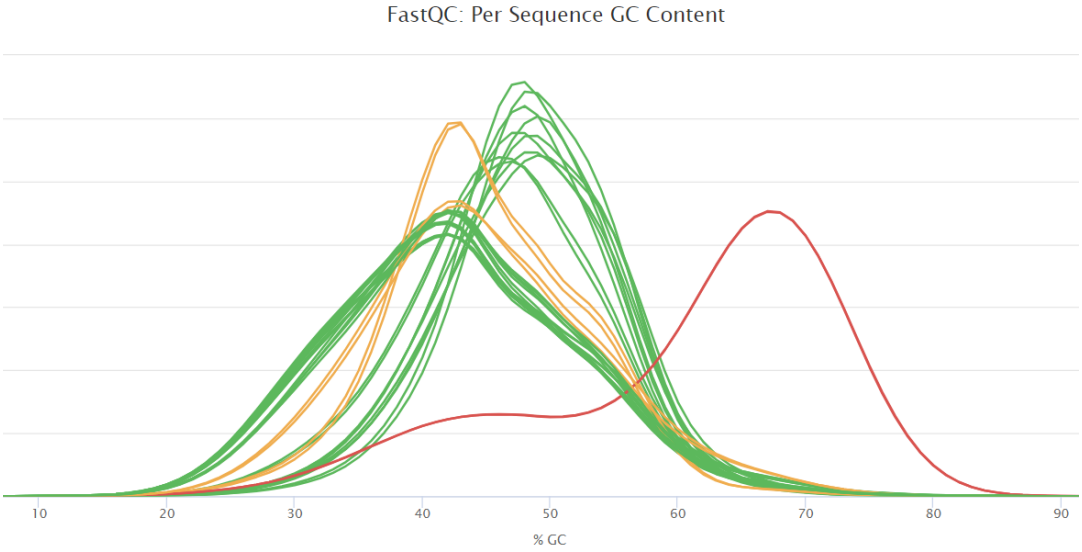

Figure 4: GC content distributions for giant eland, first round of QC

While GC content was mentioned earlier, in general after picking out bad samples for the second round of QC, the reports looked good except for some slight downward skew in GC content distributions in some individuals of lesser kudu and sitatunga or the detection of poly A sequences in a low number of sequences in all species, both phenomena which i havent yet found a good explanation for. In the second round of QC the intraspecies PCA and IBS trees both looked as expected with adherence to sampling country in most cases except for those samples which were later found to most likely be marked as the wrong species. SATC didn't work for the lesser kudu, nyala and mountain nyala batches, for the rest of the species on the normalized depth of the largest scaffolds, the chromosomes, the picture was the same with the largest scaffold having a split of

either 1 or 0.5 for each individuals which would point to it as the X chromosome, and another at a little under half the size with a normalized depth ranging from 0 to a little above a half which appears to be the Y chromosome. Checking one of the output tsv files confirms these as the ones marked as allosomes in the cattle reference. Further down in in size we see more scaffold with splits in normalized depth that don't really seem to conform to what we would expect to see in terms of homo/heterogametic chromosomes, which might indicate some partial sex-chromosome linkage which would be interesting to investigate further.

The rest of the QC results can be found here (or here for the results before picking out bad samples) with the relevant plots found in the SATC, errorate, IBS, multiqc, pca and plots folders.

The genotype calling resulted in 201,974,955 SNP's after filtering out sites not in the mask and indels, and removing sites with missing data only reduced this by 16.2%. An example of how the 2d SFS's looked before and after the filter for multiple polymorphism and correcting wrongly labeled individuals clearly shows the improvement, both along the upper margins and in the middle. The filter for multiple polymorphism only removed 0.73% of sites.

The IBS tree, its topology shown in figure 6 and shown fully in the supplementary, displays a mostly consistent topology of species, though with some exceptions in individuals discussed below. The internal branches between species though are fairly short, suggesting either short subsequent splits or something the tree is not able model like gene flow. Here the individuals within species also seem to largely group together with their own countries of origin, though some of the intraspecies internal branches are very short on this larger scale.

The mtDNA blast results have been logged as a txt file and can be found here and the general picture is that all samples investigated that did not cluster with their assigned species, had matches for the species they clustered with in the IBS tree and these matches were all much better than the second closest match. The exception to this were 3 individuals, TspeGha_0887 which matched the harnessed bushbuck, TstrZam_0843 which matched the waterbuck (a convenient outgroup to tragelaphines) and TstrKen_0763 which placed a bit outside the rest of the clade of greater kudu and didn't get as good a match to any sequence in the database as the others, warranting further investigation. After removing bad samples and relabeling based on blast results, there are now 85 eland individuals, 7 giant eland, 58 greater kudu, 53 lesser kudu, 4 mountain nyala, 4 nyala, 34 sitatunga, 1 bushbuck and 1 waterbuck in the main data set.

In figure 7 i have shown the output of DiscoVista for the 1000 100kb windows, with the label tree topology from Astral on these same windows. It shows pretty consistent patterns around branches 0, 2 and 4, and the two alternate topologies are very similar in frequency, suggesting they be caused by ILS. Branch 6 has a somewhat frequent second configuration, which is one where nyala and waterbuck have switched places. If we still place the root on the waterbuck, this would mean that this tree would have the lesser kudu branching off first separately before the nyala. Branch 1 has a second configuration that is almost at the same frequency as the first and it has branch 9 and 3 switching places, having the bushbuck splitting off before the clade containing greater kudu and

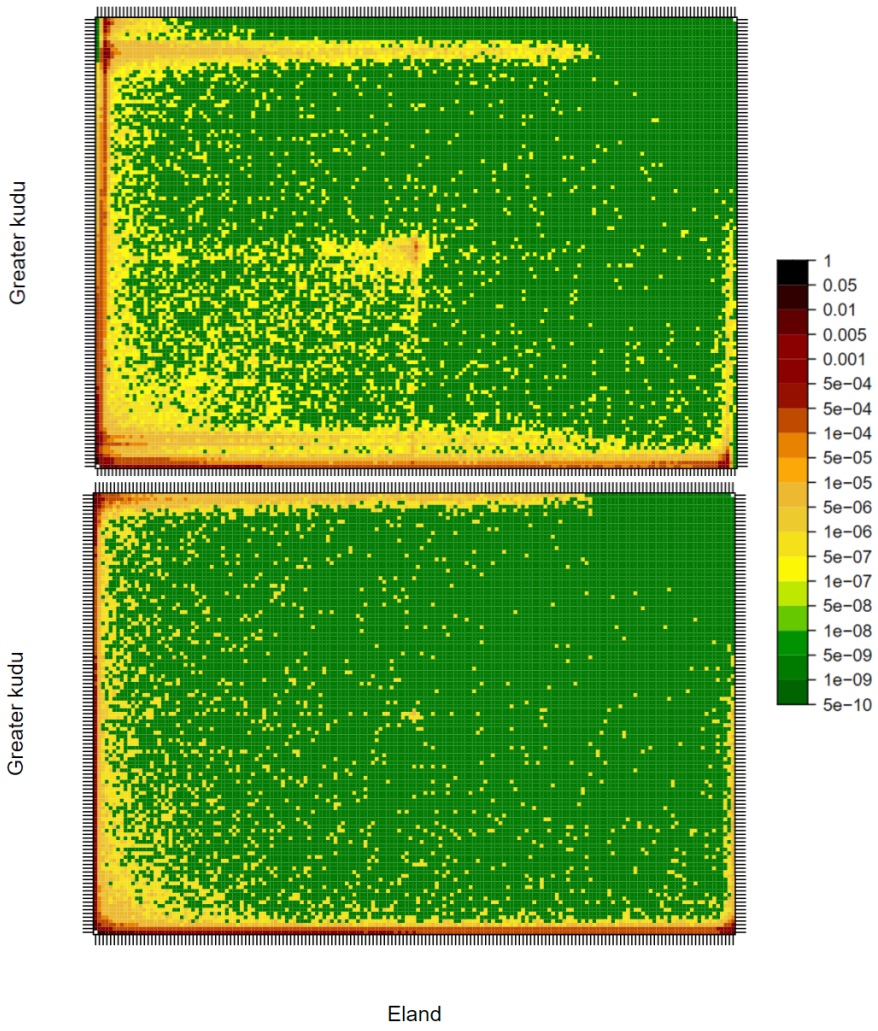

Figure 5: Comparison of two 2D SFS's before (above) and after (below) filtering and correction of labels

the elands. Branch 3 is a lot like 1, except that here the swap places the greater kudu as the first split off, just after lesser kudu and nyala. Both branch 1 and 3 have a second frequently observed configuration that are in line with the alternative topology that Astral found when the window sizes were smaller.

In figure 8 we can see the results of the Dsuite Fbranch run with the topology found by Astral for the 100kb windows. There is a clear indication that the placement of the nyala is not modeled sufficiently by this tree, which is also the case in the alternate topology run shown in the supplementary. The estimated gene flow is especially strong between nyala and bushbuck. The same is the case for the eland group with the rest on the other side of greater kudu, here the signal is very strongly with sitatunga, while not being present in the alternate topology. In the alternate topology the problem seems to stem from buschbuck being moved out to split off earlier.

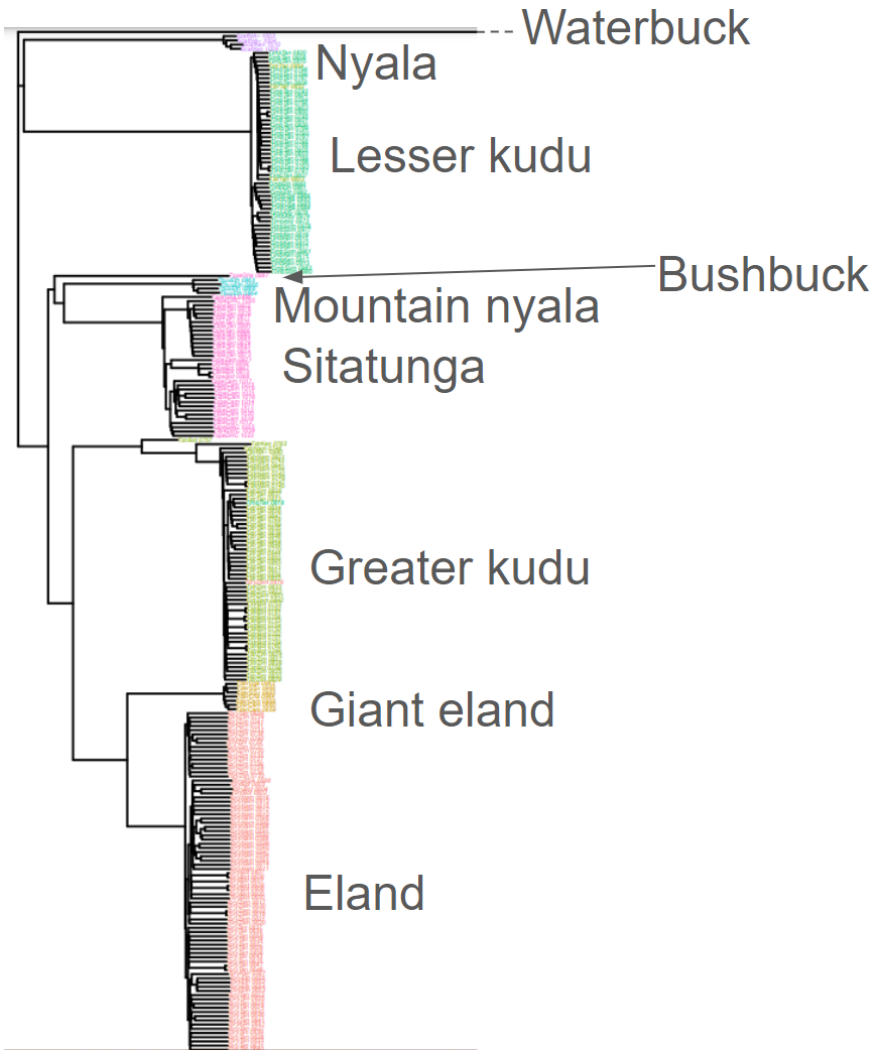

Figure 6: NJ tree made from IBS distances, can be seen in greater detail in supplement

4 Discussion

As for the primary goal of getting assessing and getting the data ready for my dissertation i would say the project was a success. While the targeted goal of high depth sequencing wasn't really met and it certainly is a waste to have it be because of too small insert sizes, the result is definitely workable, though uncertainty of base calls might have to be factored in depending on what is to be done with the data. We end up keeping most samples and were for the most part lucky with respect to which species the loss of samples affect, for example only losing 1 sample from each nyala, the two least sampled species, or while the giant eland lost half of its samples, we only lost coverage of one country.

In terms of how the sampled species relate to each other, I haven't found a single clear resolved topology, but this wasn't really the expectation either. The result from the Astral consensus tree for the 100kb windows and the IBS tree agrees with previous findings.[19] While the alternate topology found with the smaller windows and somewhat supported by the results of DiscoVista for the larger

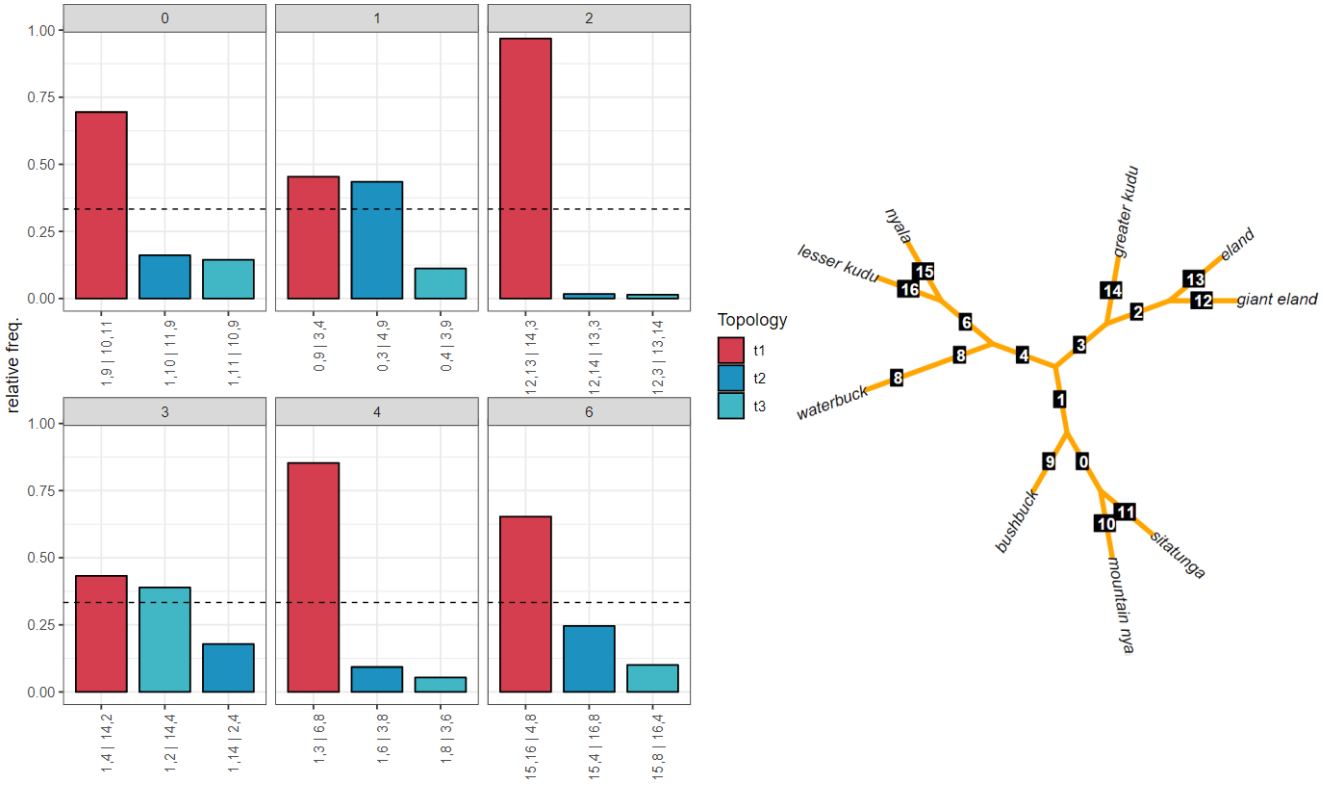

Figure 7: Ouput from the relative frequency analysis of DiscoVista for the 100kb windows

windows as well, makes it clear that a bifurcating tree is not enough to explain the observed data. If we were looking to place the gene flow events in the future to help account for the unexplained allele sharing in the tree i think we should use the main topology as a template, but with the possibility of compensating for the allele sharing with either gene flow events between before the eland split and somewhere in or before the (bushbuck,(mountain nyala, sitatunga)) clade, or possible by moving the greater kudu or bushbuck split offs further back. As for the nyala and lesser kudu branch, the short branch before this one and the outgroup in the IBS tree suggests low certainty around this part of the topology. The f-branch suggests closer relation between nyala and the rest than lesser kudu (though also stronger with bushbuck especially) and the frequency analysis seems to support this somewhat with around a 1/4 of quartets supporting a topology where the lesser kudu splits off before nyala instead of with it. This would of course also be easier to hypothesize about if we had reliable estimates for the splits between species, though such estimates would perhaps be hard to get if there has in fact been much gene flow post speciations.

There are still some unexplained observations which probably should be looked into, like the slight downward skewing of GC content in lesser kudu and things like error rates should probably be logged all together into the master table to be more easily cross referenced and accounted for, since there still are outliers present after the second round of QC. There is also the case of two of samples behaving strangely, TstrBot_0757 and TstrKen_0763, with the latter not even getting a good match in the mtDNA blast. We would need to find at least a reasonable explanation as to

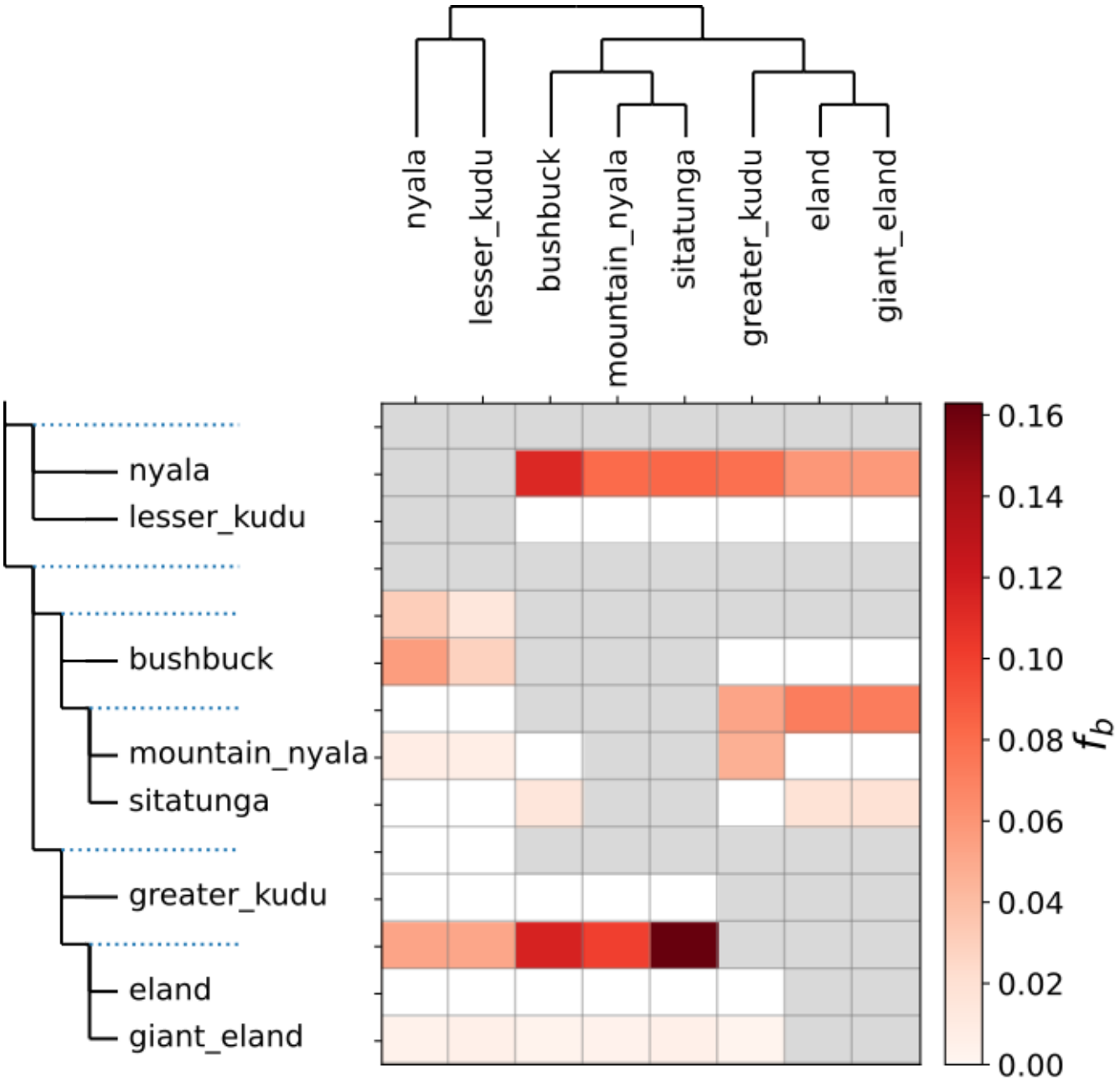

Figure 8: f-branch values for topology found by Astral for 100kb windows and IBS

why they place as they do in the IBS tree or have to exclude them.

Since generation time is well known to roughly correlate with body size in animals, it would be interesting to look at the differing rates of substitution or by proxy generation time within the tribe to see if any patterns could be found, such as a correlation with the total branch length from root to terminus in the IBS tree or differences in f-branch values between the same branch. [20] If not empirical generation times, then at least estimation of them modeled on other more easily measurable traits.

As mentioned previously there are probably also interesting things to look at in terms of differ-

ences between the reference cattle genome and the sampled species. Even though we see roughly the same depth for the mapping to the two references, it would be good to know how much of this is mapped to unused scaffolds downstream in the cattle mapped data compared to data mapped to the bongo genome which is more closely related. There are also probably more to gleam from the SATC results in terms of weird sex link splits for scaffolds and results for the species that failed to run properly.

5 Bibliography

References

[1] Desiré L Dalton et al. "Interspecific hybridization between greater kudu and nyala". In: *Genetica* 142.3 (June 2014), pp. 265–271. ISSN: 0016-6707. DOI: 10.1007/s10709-014-9772-7. URL: https://repository.up.ac.za/bitstream/2263/50500/1/Dalton_Interspecific_2014.pdf.

[2] L Koulischer, J Tijskens, and J Mortelmans. "Chromosome studies of a fertile mammalian hybrid: the offspring of the cross bongo x sitatunga (Bovoidea)". en. In: *Chromosoma* 41.3 (1973), pp. 265–270.

[3] Jiri Rubes et al. "Phylogenomic study of spiral-horned antelope by cross-species chromosome painting". In: *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology* 16 (Sept. 2008), pp. 935–47. DOI: 10.1007/s10577-008-1250-6.

[4] Mikkel Schubert, Stinus Lindgreen, and Ludovic Orlando. "AdapterRemoval v2: rapid adapter trimming, identification, and read merging". en. In: *BMC Res. Notes* 9.1 (Feb. 2016), p. 88.

[5] Mikkel Schubert et al. "Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX". In: *Nature protocols* 9.5 (May 2014), pp. 1056–1082. ISSN: 1754-2189. DOI: 10.1038/nprot.2014.063. URL: https://doi.org/10.1038/nprot.2014.063.

[6] Simon Andrews et al. *FastQC*. Babraham Institute. Babraham, UK, Jan. 2012.

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

[8] Thorfinn S. Korneliussen, Anders Albrechtsen, and Rasmus Nielsen. "ANGSD: Analysis of Next Generation Sequencing Data". en. In: *BMC Bioinformatics* 15.1 (Nov. 2014), p. 356. ISSN: 1471-2105. DOI: 10.1186/s12859-014-0356-4. URL: http://www.biomedcentral.com/1471-2105/15/356/abstract (visited on 11/26/2014).

[9] Casia Nursyifa et al. "Joint identification of sex and sex-linked scaffolds in non-model organisms using low depth sequencing data". In: *Molecular Ecology Resources* 22.2 (2022), pp. 458–467. DOI: https://doi.org/10.1111/1755-0998.13491. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.13491. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13491.

[10] Heng Li. "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data". In: *Bioinformatics* 27.21 (Sept. 2011), pp. 2987–2993. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr509. eprint: https://academic.oup.com/bioinformatics/article-pdf/27/21/2987/48865923/bioinformatics_27_21_2987.pdf. URL: https://doi.org/10.1093/bioinformatics/btr509.

[11] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2021. URL: https://www.R-project.org/.

[12] Christopher C Chang et al. "Second-generation PLINK: rising to the challenge of larger and richer datasets". In: *GigaScience* 4.1 (Feb. 2015), s13742-015-0047–8. ISSN: 2047-217X. DOI: 10.1186/s13742-015-0047-8. eprint: https://academic.oup.com/gigascience/article-pdf/4/1/s13742-015-0047-8/25512027/13742_2015_article_47.pdf. URL: https://doi.org/10.1186/s13742-015-0047-8.

[13] Emmanuel Paradis and Klaus Schliep. "ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R". In: *Bioinformatics* 35 (2019), pp. 526–528. DOI: 10.1093/bioinformatics/bty633.

[14] Aaron R. Quinlan and Ira M. Hall. "BEDTools: a flexible suite of utilities for comparing genomic features". In: *Bioinformatics* 26.6 (Jan. 2010), pp. 841–842. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq033. eprint: https://academic.oup.com/bioinformatics/article-pdf/26/6/841/48854754/bioinformatics_26_6_841.pdf. URL: https://doi.org/10.1093/bioinformatics/btq033.

[15] Alexandros Stamatakis. "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies". In: *Bioinformatics* 30.9 (Jan. 2014), pp. 1312–1313. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu033. eprint: https://academic.oup.com/bioinformatics/article-pdf/30/9/1312/48923136/bioinformatics_30_9_1312.pdf. URL: https://doi.org/10.1093/bioinformatics/btu033.

[16] Chao Zhang et al. "ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees". en. In: *BMC Bioinformatics* 19.Suppl 6 (May 2018), p. 153.

[17] Erfan Sayyari, James B. Whitfield, and Siavash Mirarab. "DiscoVista: Interpretable visualizations of gene tree discordance". In: *Molecular Phylogenetics and Evolution* 122 (2018), pp. 110–115. ISSN: 1055-7903. DOI: https://doi.org/10.1016/j.ympev.2018.01.019. URL: https://www.sciencedirect.com/science/article/pii/S1055790317306590.

[18] Milan Malinsky, Michael Matschiner, and Hannes Svardal. "Dsuite - Fast D-statistics and related admixture evidence from VCF files". en. In: *Mol. Ecol. Resour.* 21.2 (Feb. 2021), pp. 584–595.

[19] Lei Chen et al. “Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits”. In: *Science* 364.6446 (2019), eaav6202. DOI: 10.1126/science.aav6202. eprint: https://www.science.org/doi/pdf/10.1126/science.aav6202. URL: https://www.science.org/doi/abs/10.1126/science.aav6202.

[20] A P Martin and S R Palumbi. “Body size, metabolic rate, generation time, and the molecular clock”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 90.9 (May 1993), pp. 4087–4091.