Bionformatics Project I: Extending ANGSD Summary Statistics

Advisor:Dr. Anders Albrechstsen
Adham Khaled (wbf326)

2018-09-04

1 Introduction

ANGSD is a C++ open source program that was developed in 2014 for multi-threaded analysis of next generation sequence data [**?**]. ANGSD can work on a single sequence, or a set of sequences concurrently. In its current iteration, ANGSD iterates through and parses chunks of aligned reads at a time extracted from a larger file (such as BAM) or a set of files at once. The ANGSD implementation at the beginning of the project already covered the input pipeline for parsing a variety of formats, providing forward and backward read positions, as well as depths, base counts and base quality scores at each aligned read positions for a particular chunk. In this project, summary statistics generated by ANGSD were expanded to include base quality scores by depth, base composition by raw count and by ratio per depth, base quality scores in forward and backward read position, and finally, base quality score distribution of all bases found.

The motivation for generating each of the summary statistics listed above was to provide a measure of quality assurance of NGS data in an efficient and parallel manner. Sequencing depth is defined as the number of mapped reads per position. Examining the distribution of base quality scores per level of depth can be used to evaluate the quality of NGS data outside the cover statistic, which is given as a single average. This could potentially have utility where high coverage data is unavailable but also where high coverage is infact available but high base quality scores are not necessarily and uniformly guaranteed. Examining total base composition per level of depth can reveal discrepancies in over-represented bases on a level by level basis. Providing read quality by both forward and backward positions of paired end reads could potentially reveal errors in sequencing such as deamination, which is a significant issue in ancient DNA sequencing [**?**].

The task for this project was completed adding and incrementing a simple count in one of the subroutines of ANGSD for the desired statistic as each round of analysis went through a chunk of data. For full testing, summary statistics were generated for a random low coverage (10X) genome was obtained from the 1000 Genomes Project [**?**]. Furthermore, summary statistics were also generated for ancient human DNA comprising of eight 8,000 year old individuals with very low coverage (0.010x-2.14x) concurrently, and one 50,000 year old high coverage (52x) Neanderthal individual [**?**].

Materials and Methods

Code Implementation & Functionality

The general schema of the 'run', 'clean' and 'print' inherited class function from the abc superclass are listed below:

```
class general{
public:
    static aHead *header;//contains the header of a single bam;
    static std::map<char *,int,ltstr> *revMap; //contains information about all chromosomes/scaffolds
    int index; // inherited class number
    static int tot_index; //total number of inherited classes
    virtual void run(funkyPars *f)=0;
```

```
    virtual void print( funkyPars *f)0;
    virtual void clean(funkyPars *f)0;
    general(){index=tot_index++;};
    virtual  general(){};
};
```

These functions are inherited in abcSum from the abc superclass. the run and clean functions are threaded, insecure functions while print is a secure nonthreaded function (it was not used in generating the summary statistics for this report). The modifications made for generating summary statistics were added to the abcSum class and its associated header file. Firstly, in its previous iteration, the secure abcSum constructor is intialized with a struct (argStruct) containing a parse of the command line arguments and other useful information such as inferred number of files provided in the file list. The parameters for performing the analysis are declared in the class and initialized by calling the getArg function in the analysisfunction subroutine, which matches the input supplied by the user -as parsed in the input struct- to the parameter in question.

As a first step in modifying the program, a maxDepth parameter (int maxDepth in abcSum.h) was added to allow the user to set the maximum depth threshold, such that for all relevant measures (such as base quality scores) found at levels above that threshold are added lumped together as one aggregate depth level at the threshold. The constructor also initializes and opens relevant output files and variables discussed in detail herein.

abcSum's functions are firstly carried out in the insecure virtual 'run' function inherited from the abc superclass; it is called with a struct (funkyPars) as input from the file reader, which in turn contains data relevant to the current chunk undergoing processing. funkyPars keeps a record counts of bases found per file covering the length of the chunk, chunk length, as well as a number of structs generated per position and per file (NodeT **nd) containing forward and backward read positions, the base sequence and base quality scores for each aligned read at every position in the chunk. ANGSD's multithreaded nature allows chunks to be read and processed in a secure independant fashion by seperate subroutines in parallel.

The funkyPars struct is passed from the 'run' function to the 'getDepth' where depths are calculated on a file by file basis; all calculations are mainly performed inside an outer loop passing over the range of sites of the current chunk, and subsequently an inner loop passing over the range of files. Individual counts of aligned bases per file are extracted and added together to represent total depth at the current position. Then total depths are incremented as a counter in a pointer-to-pointers array by depth level, and of length set to maxDepth by the user. A second inner loop loops over the range of depths levels found at current position and current file to extract base quality scores, forward and backward read positions from each NodeT struct.

All extracted base quality scores are stored as counts under various pointers-to-pointers arrays of their respective summary statistic. For example, a pointers-to-pointers array of 'quality scores by depths' was declared in the abcSum header and initialized via constructor in the cpp with a size of [max depth set by user x range of quality scores (61)] while a 'quality scores by position' has a size of [max read length (500) x range of quality scores(61)]. Using array indexing, a count is incremented for the depth level/read position and base quality score found. On the other hand, a 'counts for bases by depth level' pointers-to-pointers array has a size of [max depth set by user x range of base codes (4)].

Once all the chunks have been processed, the class deconstructor is called. It prints the counts in long format

form, calculates the base ratios by position, closes and files and releases the various pointers from the free store. For illustratoin of long format, the counts per depth output for example would be two columns; first column keeps a record of the base where each base is repeated 61 times for the range of possible quality base scores. The second column keeps a record of counts for each base quality scores.

The statistics generated through counts were passed through R and ggplot2 [**?**] to generate the plots in this report.

Program Runs

All file were processed with the default command: ./angsd -out file.output -doSum 1 -b bam.filelist -doCounts 1 -maxDepth 5000

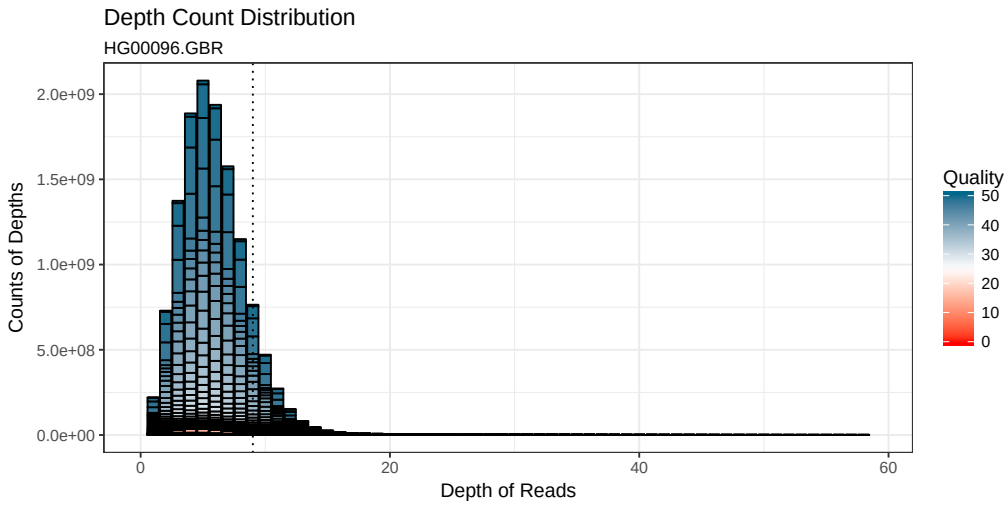
Whole Genome Sequences

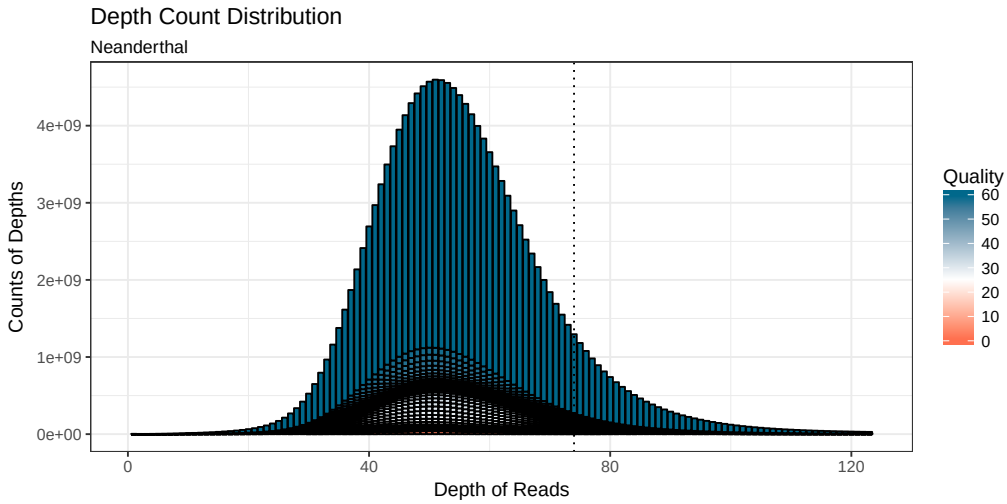
Whole genome sequences of 8 individuals (Motala) of very low coverage (0.010x-2.14x) were obtained from the Online Ancient Genome Repository (OAGR) that was originally part of a 2014 study by Harvard University's Broad Institute in association with the Max-Planck Institute[**?**] (European Nucleotide Archive Aceession:PRJEB6272). The random modern day low coverage (10x) human genome was obtained from the 1000 Genomes online repository (HG00096;British), and the 50,000 year old high coverage (50x) Neanderthal genome was obtained from the Max-Planck Institute Repository:
(http://cdna.eva.mpg.de/neandertal/altai/AltaiNeandertal/bam/)

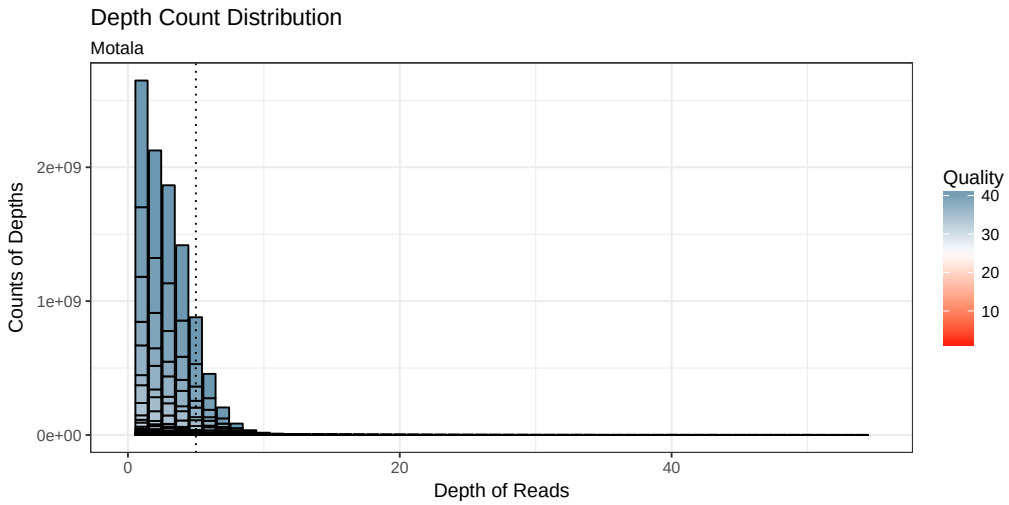
Results

Depth Distributions: All

For the purpose of illustration and in determining the significant range of depths, depth distributions were plotted as stacked bar charts and color coded by quality scores. The dashed vertical lines is a reference set to 90 percent mark for the cumulative sum of counts, while the color legend guides the distribution of base quality scores. Each stacked bar represents relative count for depths found with a base quality score given by its color shade from the lowest(red) to the highest(blue); base quality scores range from 0 to 60 in the Neanderthal chart, 0 to 40 in Motala, and 0 to 50 in the modern day human genome. It is readily noticable that the majority of depths exhibited a relatively hight base quality score, and that at most, a depth of 190 in the high coverage Neanderthal sample, is enough to capture 90 percent of all depth counts.


Depth Count Distribution
HG00096.GBR
Counts of Depths
2.0e+09
1.5e+09
1.0e+09
5.0e+08
0.0e+00
0
20
40
60
Depth of Reads
Quality
50
40
30
20
10
0


Depth Count Distribution
Neanderthal
Counts of Depths
4e+09
3e+09
2e+09
1e+09
0e+00
0
40
80
120
Depth of Reads
Quality
60
50
40
30
20
10
0



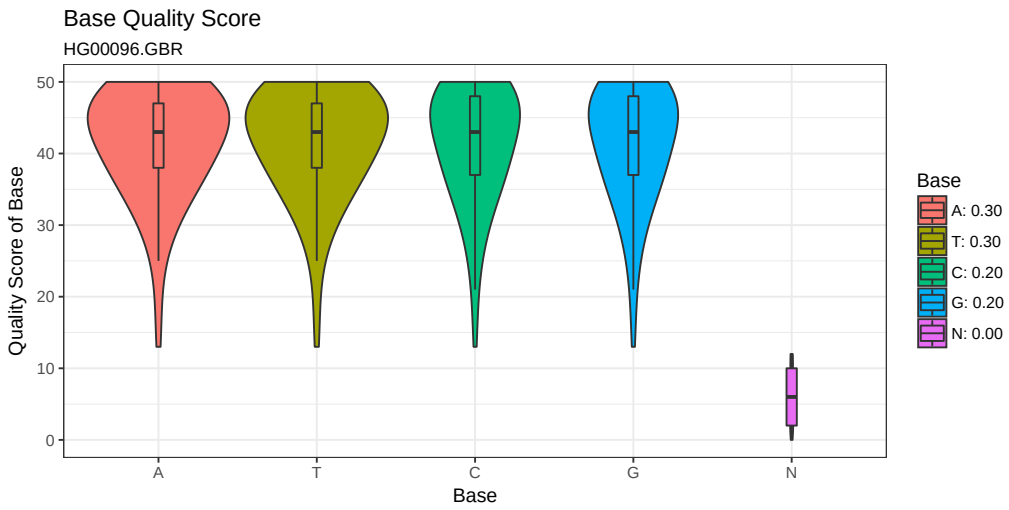
Base on the plots above, all subsequent plots involving depths were limited to a range up to 200 depths.

Human DNA: Singular Plots

Base Quality Score Distributions

Total base quality distributions are presented as violin plots allthroughout this report. Low quality base calls (N) were also included as no filtering step was performed. The plot below shows symmetric distributions between A/T and C/G bases.

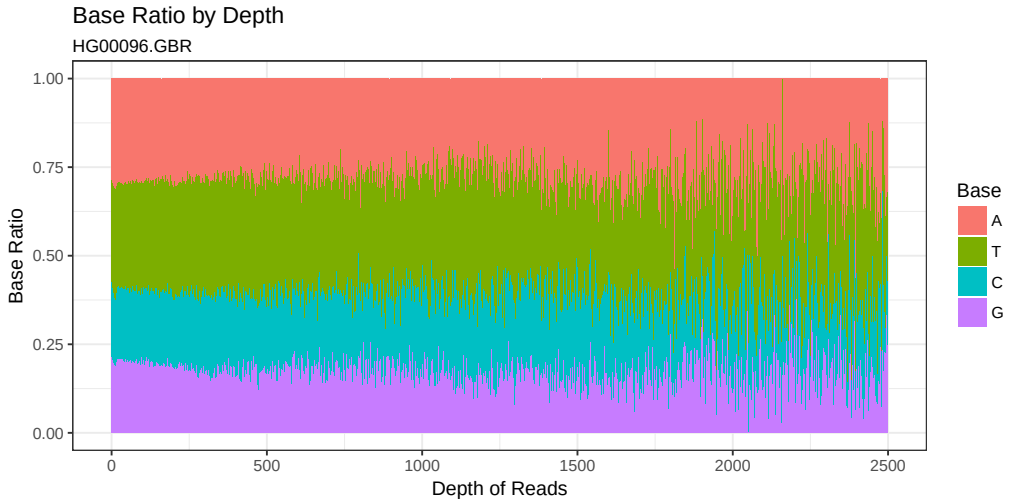
British



Base Distribution by Depth (Ratio)

Base ratios per depth were plotted as individual stacked barplots across the range of depths up to 2500 (half the maximum supplied with the input). The distributions appear to stablilize after an initial phase going up to the depth mark throughout, until roughly the 1000th depth mark, possibly due to the low base counts at depths higher than 1000.

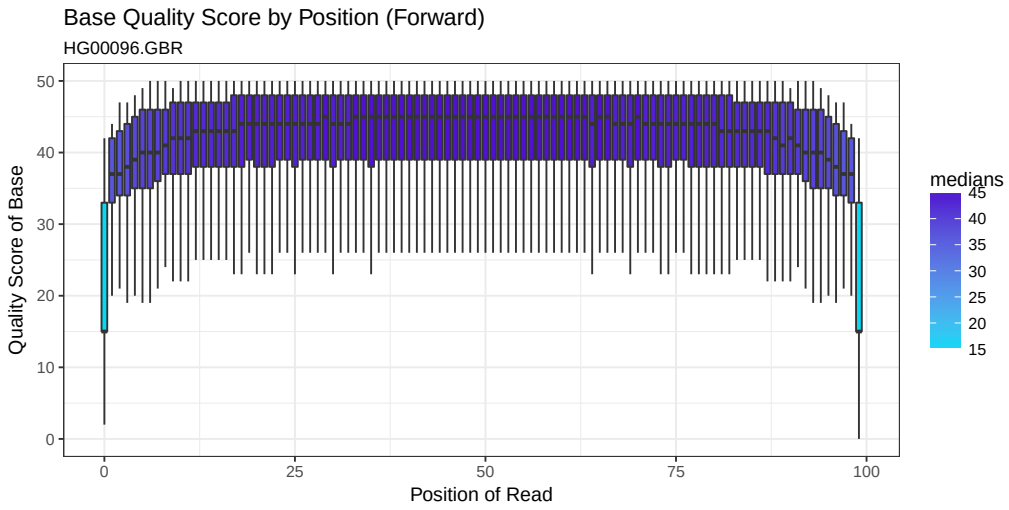
British

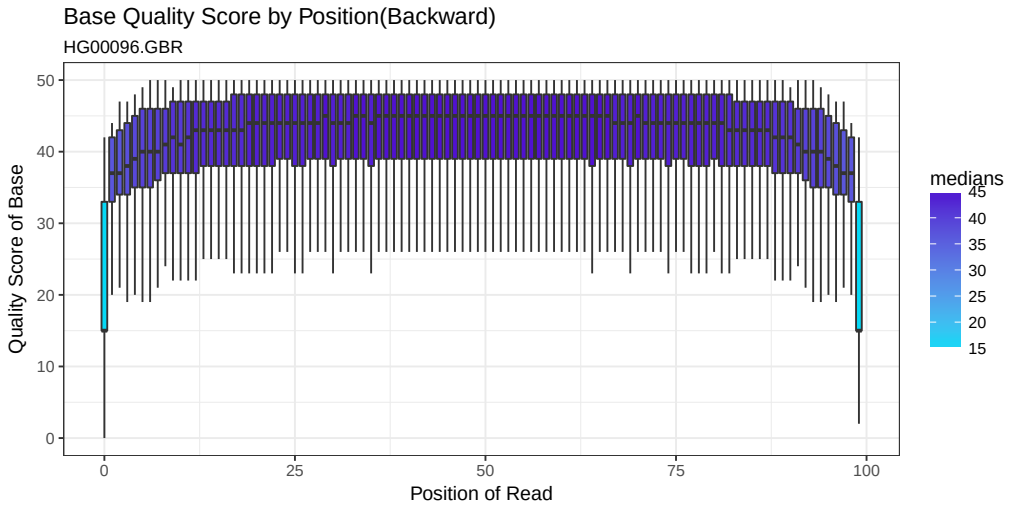


Base Quality Score by Position: British Individual

Base quality scores are displayed as boxplots ranging across the read lengths found in the bam file. It is noteworthy to note that ANGSD was modified to account for a maximum read lengths of 500, so it seems however that the maximum read depths found was 150. The forward and backward median read quality scores have a high mark of 40 for most of the read range, and dips towards the ends to the low 20s. There appeared to be no discernible differences in base quality scores between forward and backward reads.

British

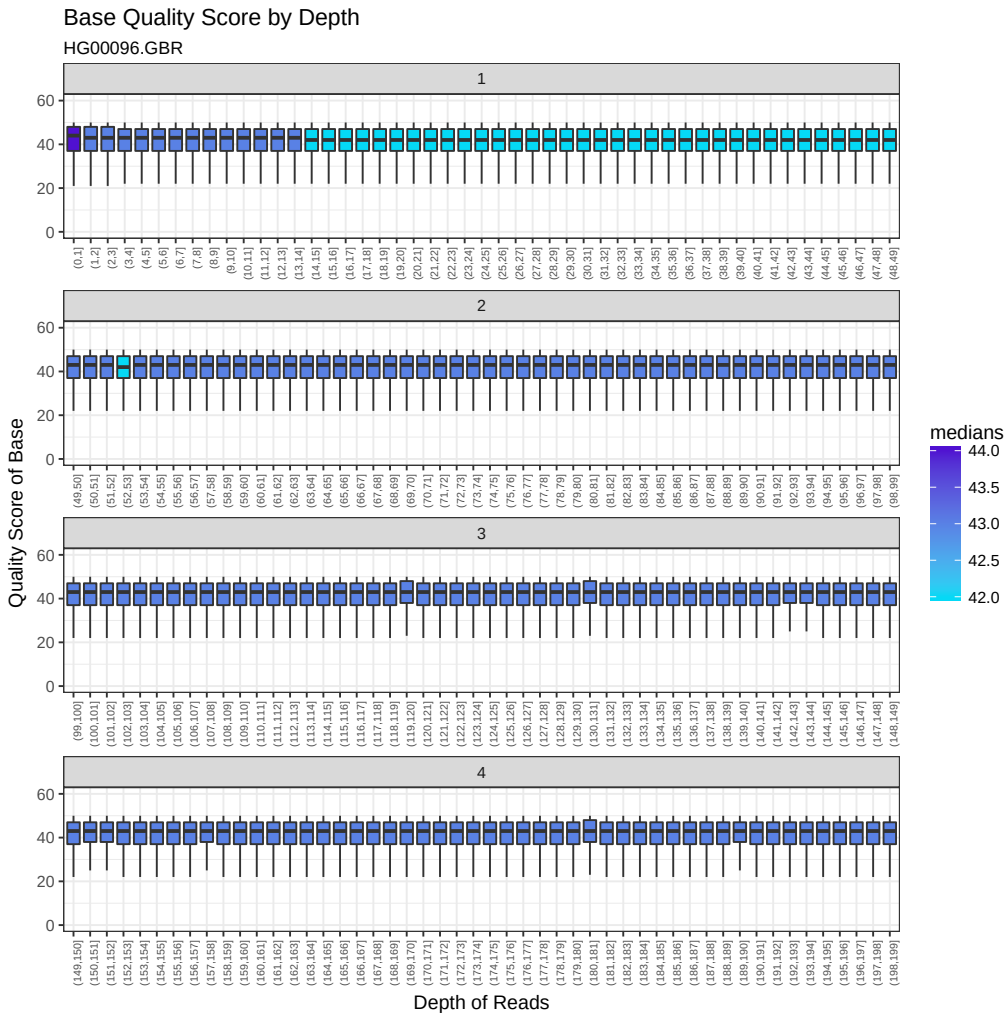




Base Quality Score by Depth: British Individual

There is some variation in the medians between the 14th and 50th mark with medians being relatively lower, but the plot shows consistenly high medians otherwise.

<u>British</u>



Ancient Human DNA: Singular & Combined

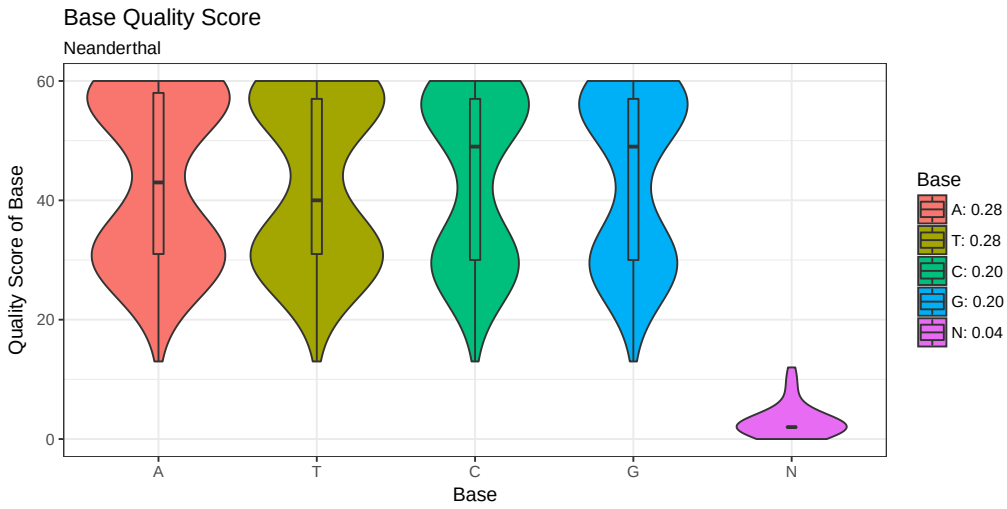
Same for the modern human plots above, the plots for Molata and Neanderthal follow the same format but show different results

Base Quality Score Distributions

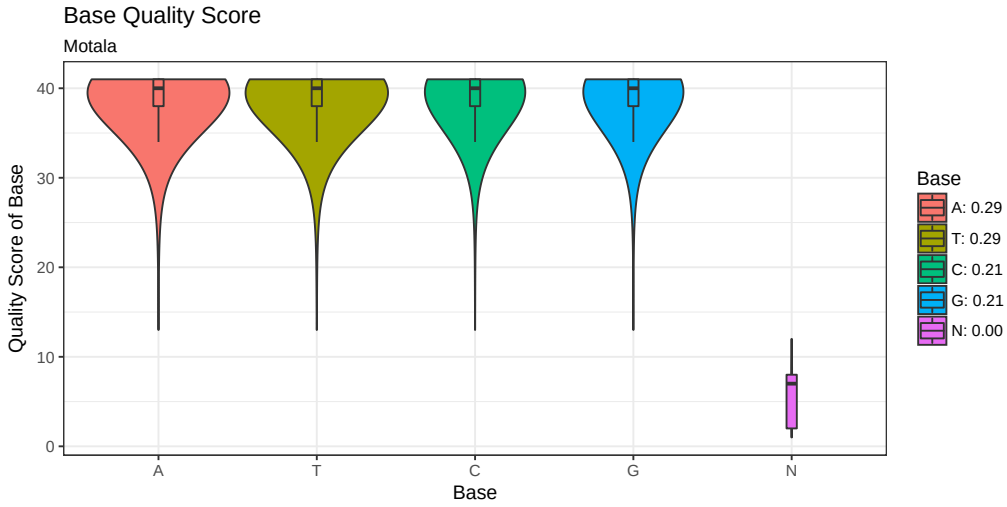
Quality score distributions for Neanderthals show a near twin peak around the 60 and 30 quality base score, unlike Molata which has a singular peak around the highest possible score indicating where the singular majority of base pair qualities fall. With the Neanderthal sample however, medians for the boxplots embedded

within the violin plots are relatively lower in A/T than C/G. Both plots shows symmetry between A/T and C/G pairs, indicating correctness of counts.

Neanderthal



Motala

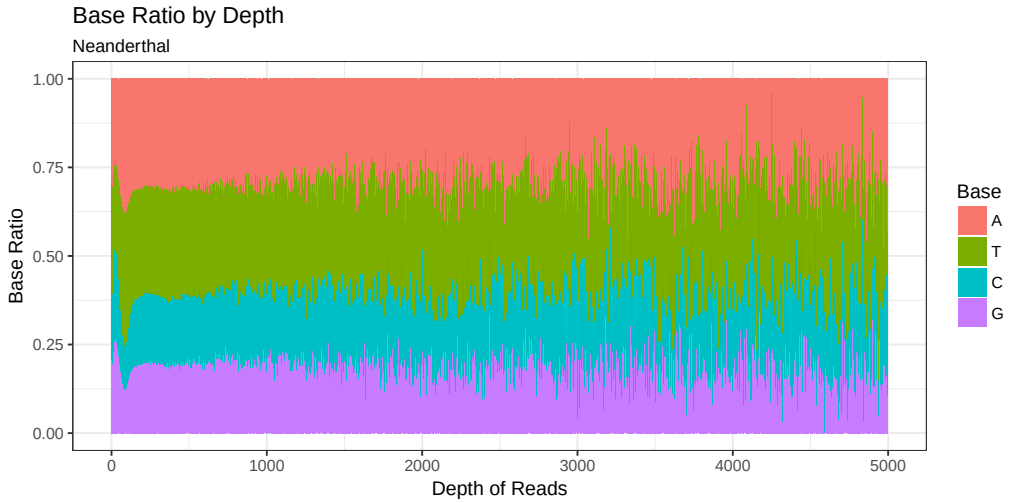


Base Distribution by Depth (Ratio)

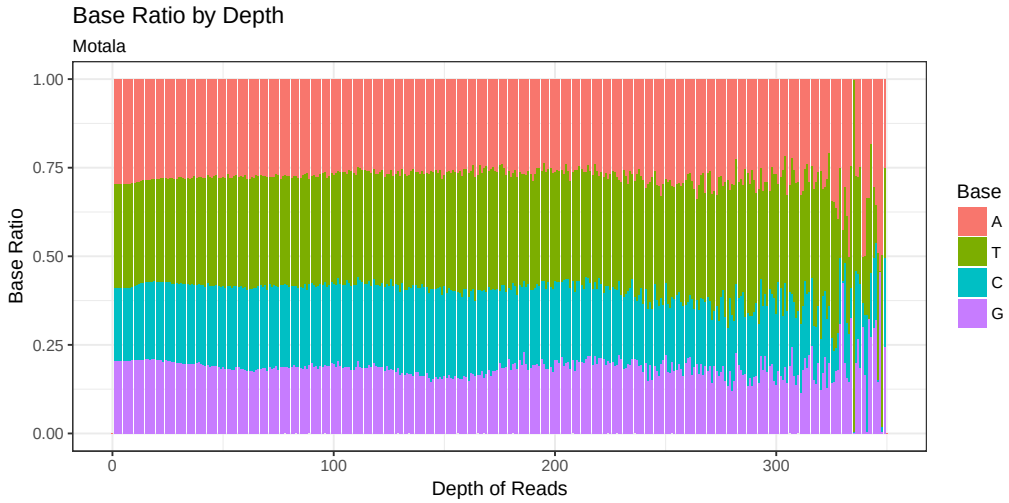
Base distributions ratios by depth for the Neanderthal sample display a stable pattern after an initial phase up to the 50th depth mark, and up until the 1000ths mark, afterwhich the ratios get gradually more inconsistent, again possibly due to low base counts after the 1000s depth mark.

The Motala plot show a similar pattern in common with all the previous plots in terms of initial phase followed by a consistent phase, afterwhich stability gradually worsens after the 200th mark.

Neanderthal



Motala

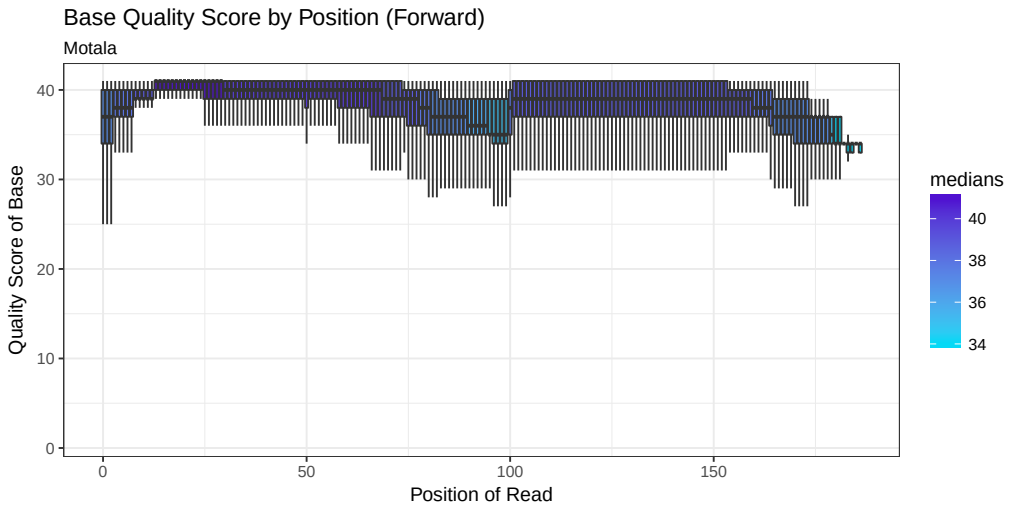


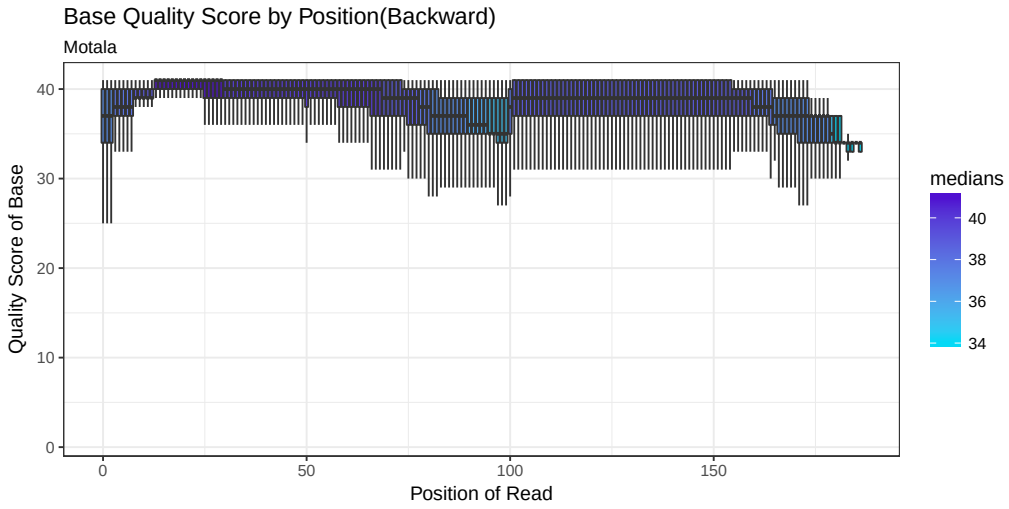
Base Quality Score by Position: Motala

The Motala base quality score by read position is much more varied than the modern day low coverage human sample, with dips in quality towards the ends and once in the middle before an abrupt rise around the 100th mark. It is noteworthy to mention that the Motala samples were sequenced on an Illumina Genome Analyzer IIx with 2x76 + 7 cycles, and then treated with UDG for deamination and further deep sequencing with ”8

HiSeq 2000 lanes of 100-bp paired-end reads" [?]. As before, there were no discernible differences between the forward and backward read positions. This could indicate the success of the UDG treatment.

Motala

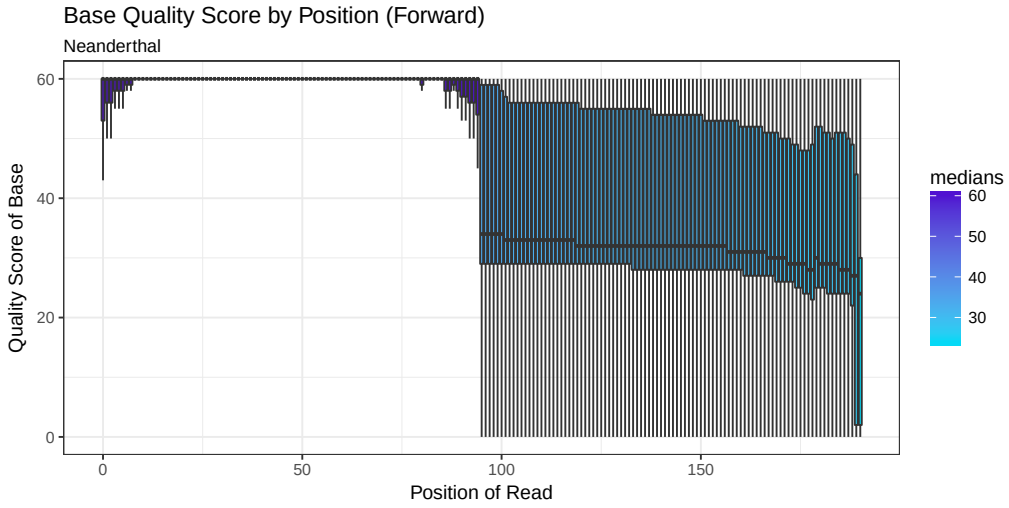


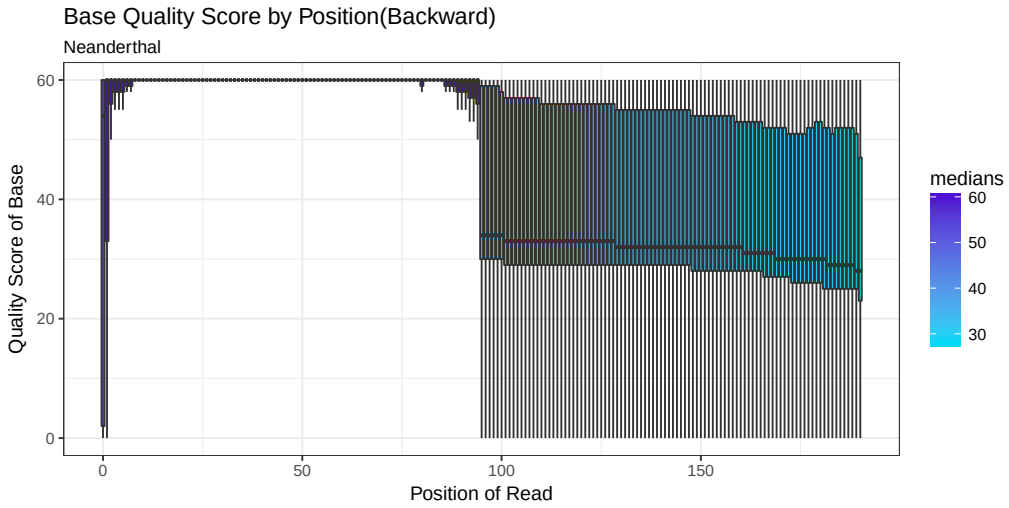


Base Quality Score by Position: Neanderthal

As with the Motala sample, quality score by read position show two dips in quality at the beginning and at the 100th mark, after which an abrupt dip in medians that gradually worsens to illegibility towards the end.

Neanderthal

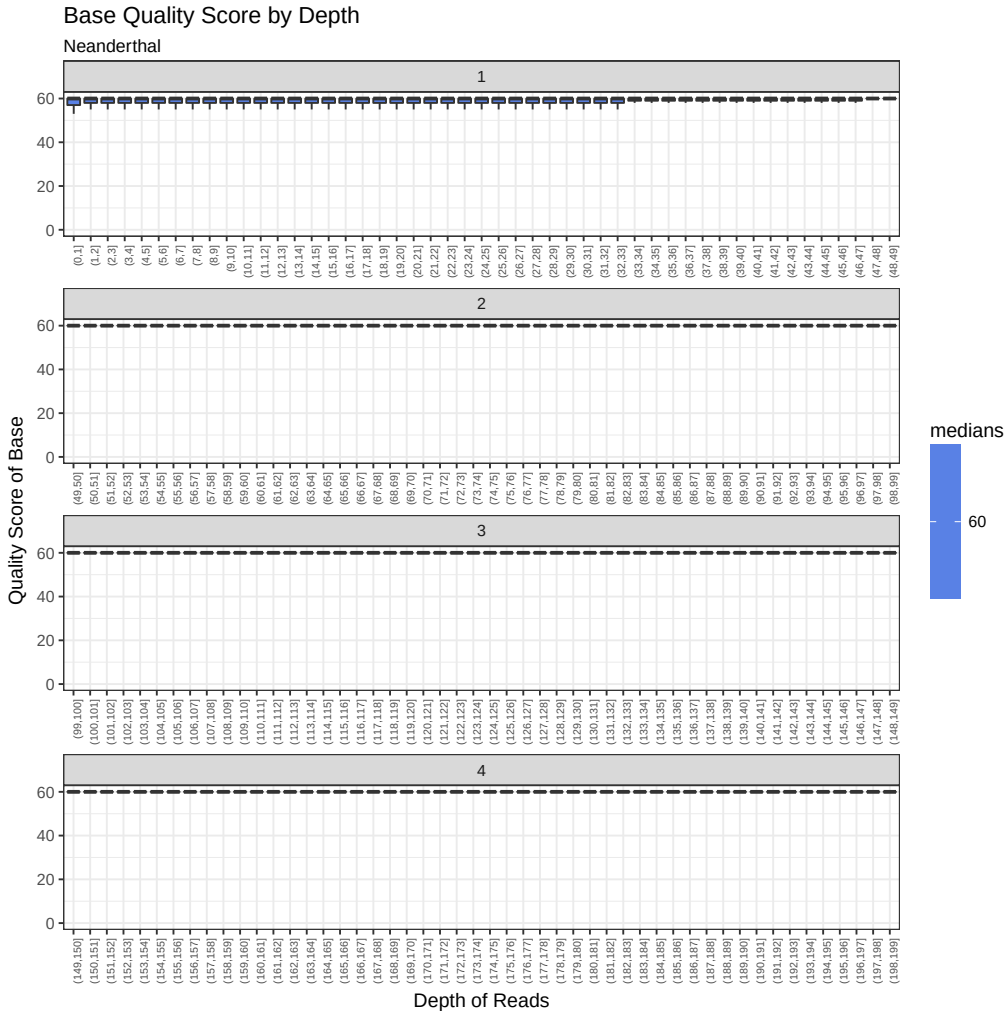




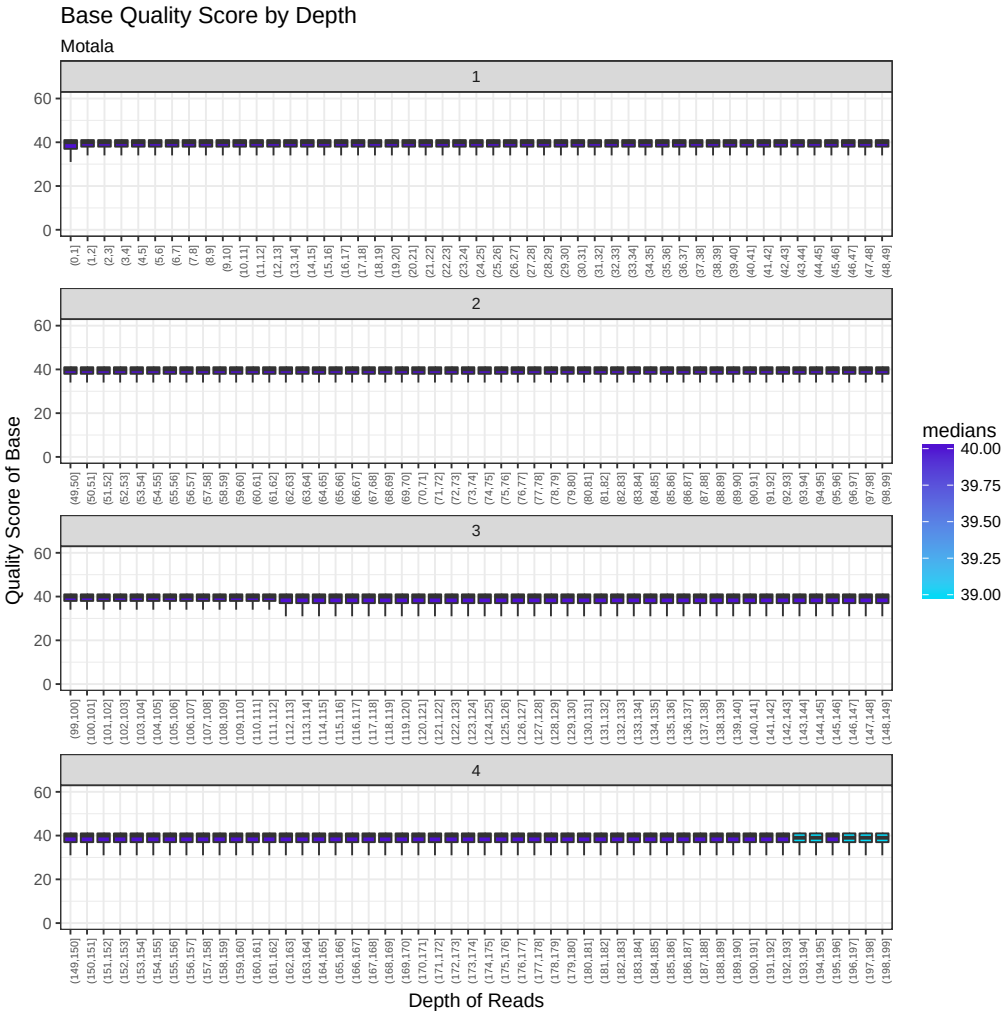
Base Quality Score by Depth: Neanderthal

Base quality scores by depth for Neanderthals show consistently short boxplots with a high median at 60 allthroughout, with boxplots getting gradually shorter past the 33 depth mark. Motala samples also show a consistently high but wider distribution of base quality scores up until the 192nd mark after which a slight dip in medians from 40 to 39 become more prevalent.

<u>Neanderthal</u>


Base Quality Score by Depth
Neanderthal
1
2
3
4
Quality Score of Base
Depth of Reads
medians
60

Motala



Discussion

ANGSD was modified to perform summary statistics on genomic sequences while perserving the paradigm of data chunking and multithreading. A peak into the distributions of bases by depth was generated and showed irregulartieis at higher depths, which is expected as the frequency of occurance of highly overlapping reads get less at higher depths. Total uniformality in base quality scores by depth was lacking in all the samples, with some unexpectedly low medians for the Neanderthal sample between depths 40 and 60, and yet more conservative variations in all others. This still goes unexplained, but could point towards specific areas in the genome where mapped reads are of poor base quality and low in coverage. The summary statistics for all samples did not show but the slightest variation in base quality scores between forward and backward reads. In ancient DNA samples specifically, this could be proof of success of the UDG deamination treatment. However,

the lengths of reads in the Motala and Neanderthal appeared to be double the expected size. Both were initially sequenced with an illumina 76bp library method, and then subjected to 101bp deeper sequencing. This could potenitally point to a fault in ANGSD, or a specific format in the construction of the bam files themselves, as both samples were sequenced at the same Institute and by the same team of researchers. It is likely that some paired end reads (all the Neanderthal libraries were paired end, while the Motala samples were mixed) are read as one continuous read of double the actual length, leading to the inconistensies in base qualtiy scores after roughly the 101bp mark. Another possible explanation is that the researchers themselves merged the libraries with purposfully abjacent read positions to differentiate between regular sequencing and the deeper sequencing reads. This actual reason with proof however remains unknown. And finally, as a minor point, the depth distributions have shown that a depth of 200 and below is enough to capture 90 percent of all aligned bases in a high coverage genome, and setting a max depth above that figure may not be necessary to qualtiy control, unless a particular interest in rarer high depths is given.