MASTER'S THESIS

An exploration of linkage disequilibrium measurement and interpretation of decay using next generation sequencing data

EMMA FOX
Computational Methods in Ecology and Evolution
Imperial College London (Silwood Park)
(e.fox16@imperial.ac.uk)

Supervised by Dr. Matteo Fumagalli
Committee: Dr. Jason Hodgson and Professor Alfried Vogler

A thesis submitted in partial fulfillment of the requirements for
the degree of Master of Science/Research at Imperial College London
Formatted in the journal style of *Bioinformatics*
Submitted for the MRes in Computational Methods in Ecology and Evolution
31 August 2017

I certify that all work submitted as part of this thesis is entirely my own original effort unless referenced accordingly in the text and list of sources that follows. Simulated data was conducted by myself. The turkey and duck samples were downloaded from the NCBI website in FASTQ format from a project previously uploaded by Alison Wright (University of Sheffield). Guppy data was provided by Vicencio Oostra (University College London) in BAM format. All subsequent filtering and processing was my own work. The ngsLD package was primary developed by Filipe Vieira (University of Copenhagen). I contributed the script relating to the Fit_Exp utility as well as providing a tutorial for the package.

Emma Antonia Fox (2017)

1 Abstract

This study explored several aspects of the measurement of linkage disequilibrium (LD) and the inferences it is possible to draw about populations from these measurements, with an emphasis on the importance of accuracy in these initial calculations for downstream analysis. LD refers to observed non-random assortment that can occur between loci in a genome and is indicative of processes such as genetic drift, population size changes, and selection. Traditional methods of calculating the strength of LD between single nucleotide polymorphism (SNP) pairs rely on using called genotypes. However, this requires a threshold of certainty be designated for the genotype at a loci to be called. In the case of data with a low read depth, it becomes increasingly difficult for genotypes to exceed the threshold of certainty and a large portion of data is excluded from consideration. In an effort to expand the possibilities of LD research for non-model species, ngsLD was developed. This package uses two probabilistic methods to calculate LD without the need to call genotypes. The first part of this study involved quantifying the increase in accuracy these methods represent. Further methods were then developed to calculate decay over distance from the LD measures. These methods were applied to two case studies, the first of which compared the population history of a domestic population of turkeys and ducks. The second compared the history of populations of Trinidad guppies from 4 separate rivers, each of which has a subpopulation subjected to high predation downstream and a colonizing subpopulation subjected to low predation upstream. The comparisons of these species exemplify the applications LD measures can have for the study of many different species with this new method of calculation.

2 Introduction

Linkage disequilibrium (LD) refers to a deviation in the inheritance rate of a particular haplotype, or combination of two specific alleles on the same chromosome, between generations from the proportion expected if the loci the alleles occupy are independently assorted (Slatkin, 2008). Haplotypes can be made up of alleles, or single nucleotide polymorphisms (SNPs), from more than two loci, and the LD between them can be calculated, but this study focused on pairs of just 2 SNPs at biallelic loci unless otherwise mentioned. If loci are randomly associated, the proportion of individuals with a specific haplotype will be the product of the respective proportions of each SNP in the population (Slatkin, 2008). Loci in LD are detected as haplotype occurrence frequencies different to those expected.

Alleles can become linked in multiple ways. One of the primary mechanisms is genetic drift. Genetic drift refers to the random loss of haplotype combination from one generation of a population to the next, causing the other haplotypes to occur in greater proportion (Slatkin, 2008). Population bottlenecks refer to a sudden decline in population size, comparable to a large amount of drift, and therefore can also elevate LD across the whole genome (Pritchard and Przeworski, 2001; Slatkin, 2008). Migration between or intermixing of different populations can create LD patterns as well. As the new haplotypes mix with the existing population, haplotypes that were previously locally fixed will now be recognizable against the new haplotypes added (Pritchard and Przeworski, 2001; Slatkin, 2008). While the aforementioned mechanisms will cause changes throughout the genome natural selection will create LD within a particular region. Strong positive or negative selection will elevate LD at a particular region of a genome by either selecting an allele towards fixation in a short period of time (with the surrounding loci being carried to fixation as well) or by selecting against haplotypes including deleterious alleles being passed on to the next generation, respectively Slatkin (2008); Reich et al. (2001).

However, even strongly linked SNPs will eventually become 'un-linked' over time because of crossing-over during meiosis, which reintroduce an element of random assortment (Hartl et al., 1997; Slatkin, 2008). While each species, and often each region of the genome will have their own recombination rate, SNPs that have a greater distance between them are at a greater chance of being separated by these recombination events. Therefore the strength of LD from a whole genome set of SNPs tends to decay exponentially with distance between SNPs (Park, 2012). In addition, mutation can also weaken LD as the addition of new variants breaks up existing haplotypes (Slatkin, 2008). The size of a population affects many of these processes including the chance of genetic drift as well as the number of recombination events that occur and mutations that arise each generation (Hartl et al., 1997; Reich et al., 2001). Consequently, the strength of LD along a genome is reasonably sensitive to significant population size changes in recent history.

LD studies have a variety of uses in human biology and beyond. Particular interest in LD has been generated by genome wide association studies which use LD measures to determine the area of interest around marker SNPs (Reich et al., 2001). It is possible to determine the range around marker alleles where the 'causal allele' is likely to occur by matching how often the marker and the trait of interest occur together to a threshold of LD strength associated with a particular distance (McRae et al., 2005). LD has also been used in studies to determine past population sizes in both humans (Park, 2012) and wild animal populations (Hernandez et al., 2007). LD decay, in particular, has been investigated as an indicator of events in a population's past. As previously mentioned, a population experiencing a recent bottleneck will show depressed levels of LD decay (Pritchard and Przeworski, 2001; Slatkin, 2008). A population that is expanding will show lower levels of LD and a steeper decay because the greater number of gametes being produced increases the chance of crossing-over events or mutations weakening the strength of LD throughout the genome (Park, 2012). It therefore possible to detect these events by comparing the rates of decay across related populations (Hernandez et al., 2007; Park, 2012).

The accuracy with which LD can be measured is particularly important because it directly determines the precision and accuracy of the inferences made from downstream analysis. The usual method of calculating LD uses SNPs from called genotypes to calculate various measures of LD strength (Slatkin, 2008). However, especially in the case of poor quality or low read depth data, using only SNPs whose determined genotypes exceed a specified confidence level may leave very little data or homozygous skewed data to work with (Nielsen et al., 2011). This method also requires a suitable confidence level be identified before calculations can be done but this may not always be known for non-model species. ngsLD (in development by Vieira *(*et al.) was developed to address the issues with the traditional method of LD measurement. ngsLD uses both expected genotypes and genotype posterior probabilities to measure LD without needing to call genotypes. This utilizes of the full range of quality and prior probability information available at these sites to greatly improving the accuracy of LD calculation for non-model species (Nielsen et al., 2011; Fumagalli et al., 2013).

The first purpose of this study was to compare the accuracy of LD calculations using the methods employed by ngsLD to calculations done using the traditional method of genotype calling. The second purpose was to exemplify the usefulness of these measurements for the comparison of populations using LD decay through two case studies. The first case study examined the whole genome LD decay curves of a population of domestic turkeys (*Meleagris gallopavo*) and mallard ducks (*Anas platyrhynchos*) to compare the recent population history of each (Harrison et al., 2015). The second case study was run on a data set of Trinidad guppy (*Poecilia reticulata*) populations from 4 rivers which included samples from a highly- and less-predated subpopulation in each river.

3 Methods

3.1 Measuring LD

There are several different metrics that quantify the strength of LD. The fundamental indicator of LD occurring, though, is when two alleles at different loci appear together in a greater proportion than that which is expected from the allele frequencies behaving according to Hardy-Weinberg equilibrium (HWE) (Slatkin, 2008). This is represented by the quantity *D*

$$D = p_h - (p_{a1} * p_{a2}) \tag{1}$$

where p$_h$ is the proportion the haplotype is observed at in the sample and p$_{a1}$ and p$_{a2}$ are the proportions the alleles at the first and second locus, respectively, exist at in the sample. A *D* value of 0 indicates the alleles are inherited according to HWE or, in other words, associate perfectly randomly. *D* measures the strength of linkage for a specific combination of 2 alleles at 2 different loci therefore each combination of alleles between the two loci will have their own *D*. If the loci are both biallelic, *D* will be have the same absolute value for all 4 combinations with 2 of the combinations having a negative value while the others have a positive value so it is only necessary to report a single value for the set of loci (Slatkin, 2008).

The frequency each allele occurs at constrains the possible values of *D*, though, making it an unsuitable measure for comparing between SNP pairs. The quantity *D'* scales *D* by dividing *D* by the maximum possible value as determined by the allele frequencies (Lewontin, 1964). Both of the following equations are evaluated and the smaller denominator (which will consequently yield the larger *D'* value is chosen as the maximum *D*.

$$D' = D/(p_{a1} * (1 - p_{a2})) \tag{2}$$

or

$$D' = D/(p_{a2} * (1 - p_{a1})) \tag{3}$$

The primary way LD was evaluated in this study was the r^2 metric. This measure further modifies the *D* quantity to measure the correlation between the presence of one allele and the presence of another at a separate loci (Slatkin, 2008). It is defined by the following formula:

$$r^2 = D^2/((p_{a1} * (1 - p_{a1})) * (p_{a2} * (1 - p_{a2}))) \tag{4}$$

LD is most easily calculated using phased data. Phased data refers to sets where the chromosome each allele of a genotype is on has been identified, allowing for a haplotype to be established across loci. This becomes an issue in the case of individuals that are heterozygous at both loci because it is unknown whether the individual's haplotypes are a pair of both loci's minor alleles and a pair of the major or if each haplotype is the major allele at one locus and the minor at the other (Clayton and Leung, 2007). If the genotype data is unphased, it is possible to mathematically interpolate the haplotype using data from the wider population (Marchini et al., 2007), using an expectation-maximization (EM) algorithm to resolve haplotypes (Slatkin et al., 1996), or solving for a cubic equation for the proportion a particular haplotype occurs at which is derived from a series of equations involving both SNP and genotype frequencies (Clayton and Leung, 2007). Once the haplotypes are estimated, the measures of LD can then be calculated using the formulas above (Clayton and Leung, 2007) or using likelihood ratio tests (Slatkin et al., 1996).

3.2 Measuring LD with ngsLD

Rather than calling genotypes, ngsLD works with expected genotypes and genotype posterior probabilities. It calculates these by using genotype likelihood data files (.glf format). The likelihood of a genotype at a particular position is calculated using the likelihoods of the constituent alleles from each read. The quality score (usually phred) of the base is interpreted along with information about the reads' cycle and other prior information to give the likelihood of each allele which can be used to calculate a measure of certainty that the particular locus is each possible genotype (Nielsen et al., 2012; Li et al., 2009). Genotype posterior probabilities apply prior probability data such as genotype frequencies under Hardy-Weinberg equilibrium (HWE) to the genotype likelihoods to give the probability of a locus being a particular genotype (Fumagalli et al., 2013). Therefore, genotype posterior probabilities incorporate both error information from alignment and base calling with available prior probability data from that population including the allele frequencies from the wider population or sample. This results in a measure of how likely that locus is a specific genotype in comparison to the other two possible genotypes with the probabilities of all three possible genotypes summing to one (the three possible genotypes being a pair of alleles consisting of 0, 1, or 2 derived alleles) (Nielsen et al., 2011; Fumagalli et al., 2013).

To get the expected genotype of a locus, each possible genotype is defined by the number of derived alleles, weighted by its probability of occurring, and the resultant product is treated as the genotype (Ross, 2014). ngsLD then calculates an r^2 using the expected genotypes for each locus. D, D' and a second measure of r^2 are calculated as well, this time using an expectation maximization algorithm (based on bcftools v0.1.18 (Li et al., 2009)) which determines the relative frequency haplotypes occur at. In essence, this method uses all the genotype posterior probabilities to calculate the probability and strength of linkage using each possible genotype for each locus and weighting the measures of linkage by the genotype likelihoods (Li et al., 2009). While calculations using the expected genotype still rely on using 1 summary value of the locus, albeit with more background information incorporated than a called genotype, using genotype posterior probabilities allows for each possible genotype outcome to be evaluated and its LD results weighted to come up with the final measure.

The second phase of this study focused on developing a set of scripts to measure whole-genome LD decay as part of the ngsLD package. The first of these was a python (v2.7.13) script that uses the LmFit python package (Newville et al., 2016) to fit an exponential decay curve to a scatter plot of LD strength versus distance between pairs. This decay curve can be represented by the equation

$$r^2 = r_0^2 * e^{\lambda x} \tag{5}$$

This equation describes the average LD strength between two SNPs separated by distance x to be the curve's intercept times *e* raised to the product of the decay coefficient λ and distance x. The script uses the minimize function of LmFit to vary both the parameters λ and r_0^2 to find the exponential curve that gives the best fit of the data using non-linear least-squares fitting method (Newville et al., 2016). In other words, it tests out different combinations of parameter coefficients until it finds one with the least amount of residual error.

Once the coefficients are determined for each data set, they are written to a csv file which can be passed to the R (v3.4.1) script that was also developed for this project. The R script takes the coefficient file and will plot each curve individually with the scatter plot data in the background, each curve on a separate blank plot, and/or all of the curves on a single plot for comparison using ggplot2 (Wickham, 2016).

3.3 Quantifying Accuracy

The first objective of this study was to quantify the increase in accuracy ngsLD provides in calculating LD relative to the traditional method of using called genotypes. To test this, the ms package (Hudson, 2002) was used to generate a set of 10e7 variable sites for 50 individuals from a hypothetical diploid population of 10,000 individuals with a historically constant growth rate. Additional parameters included a mutation rate of 1.5e-8 and an error rate of 0.01. The resulting file was then processed using ANGSD v0.919 (Korneliussen et al., 2014). The msToGlf function created 5 files of genotype likelihoods based on the hypothetical sequencing of the ms output population at read depths of 2, 5, 10, 20, and 50. Subsequently, the rest of the sequences around the variable sites were filled in, also using ANSGD with a minimum minor-allele frequency threshold of 0.02 with the exception of the 2 read data set which was subjected to a p-value threshold of 0.001 to yield a file of similar size to the rest of the sets. These sequence files were run through ngsLD (cite?) twice to determine the strength of LD between each pair of SNPs using called genotypes, expected genotypes, and genotype posterior probabilities.

The data set created using 50 reads and called genotypes was chosen to represent the 'true' data. This data set was chosen because it had the highest read coverage and used the commonly accepted method to calculate LD (Nielsen et al., 2011) so the increase in accuracy ngsLD provides would be easiest to visualize in comparison. All of the files were then filtered using AWK to keep only data from pairs of SNPs that appeared in all 10 files. Using data points shared across all files was necessary to compare measures of LD for specific SNP pairs between methods. The files were compared by calculating the root mean squared deviation (RMSD) and standard bias of SNP pair's LD strength in comparison to the 'true' data set. RMSD is determined by taking the square root of the sum of the squared differences between the observed and expected or ‘true‘ data divided by total number of observations. The standard bias is calculated by summing the differences between the observed and expected and dividing by the total number of comparisons.

3.4 Case Study 1: Turkeys and Ducks

mRNA data for the turkeys and ducks was downloaded in paired-end FASTQ form from SRA PRJNA271731 (Harrison et al., 2015) on the NCBI website using the SRA Toolkit v2.8.2-1 fastq-dump function (https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software). This data was sequenced from spleen and gonad samples of 14 turkeys and 20 ducks from captive populations (Harrison et al., 2015). The mRNA data quality was initially analysed using FastQC v0.11.5 (http://www.bioinformatics.babraham.ac.uk). Adapters and low-quality sequence data was eliminated using Trimmomatic v0.36 (Bolger et al., 2014) with the leading and trailing quality threshold set at 3, the average quality of a 4 base window set to 15, and minimum sequence length set to 50. HISAT2 (Kim et al., 2015) was used to map the sequences onto the reference genome or reference scaffolds available through NCBI for the turkey and ducks, respectively. In the case of the ducks, only scaffolds over 1000 bases in length were considered. HISAT2 options were used to exclude the reporting of discordant or mixed matches and unaligned reads. The resulting files were converted from SAM to BAM format, sorted, and indexed using the corresponding SAMtools (Li et al., 2009) functions.

The bam files were analyzed as a single set of turkeys and single set of ducks by ANGSD (Korneliussen et al., 2014) for a final quality check and trimming. The two sets of bam files were then run through ANGSD once more to calculate the genotype likelihoods. The maximum depth for ANGSD was set to the 95th percentile of the global depth and other options included using only reads that mapped to a single location, excluding unpaired reads, a minimum base quality of 20, a minimum read quality of 20, only including loci with data available from at least half of the samples, assigning the major and minor alleles using the genotype likelihoods, excluding triallelic loci, using a minimum allele frequency filter (equal to 1/(number of individuals), using allele frequencies for prior probabilities, and output posterior probabilities. ngsLD calculated LD between SNP pairs up to 1 mega-base pairs apart using these resulting genotype likelihoods.

The decay equation for each population was determined using the Fit_Exp utility. A bootstrapping of the data was also conducted to test for the significance of the difference between the two curves. The rows of the ngsLD output files were divided into 100 groups. The groups were then sampled with replacement 100 times and Fit_Exp was run on these data sets to determine the 5th and 95th percentiles. This information was used to plot the shaded confidence intervals around each curve.

3.4.1 Case Study 2: Guppies

The DNA used for this case study was collected from 8 populations of guppies in the Northern Range Mountains of Trinidad (Wright et al., 2017). These rivers are characterized by high levels of predation upon guppies in the downstream populations but low predation in the upstream population. Convergent evolution has been previously shown to occur between populations introduced into a in similar predation environments but different rivers (Fraser et al., 2015). 5 males from both the upstream and downstream region of the Aripo, Marianne, Quare, and Yarra rivers were sampled. The only exception to this was the downstream population of the Yarra river of which only 4 samples were collected.

The trimmed and mapped sequence data was provided in bam format by Vicencio Oostra. The full details of the processing can be found in Wright et al. (2017). These files were divided up into groups by river and predation level. They were analysed with ANGSD to determine the 95th percentile global depth threshold for the maximum depth filtering. Genotype likelihoods were calculated with ANGSD with similar options and specifications as those described above. These likelihoods were then passed to ngsLD for LD calculation and decay curve fitting on a data set of SNPs separated by up to 1 mega-base pairs and a set of SNPs separated by up to 200 kilo-base pairs.

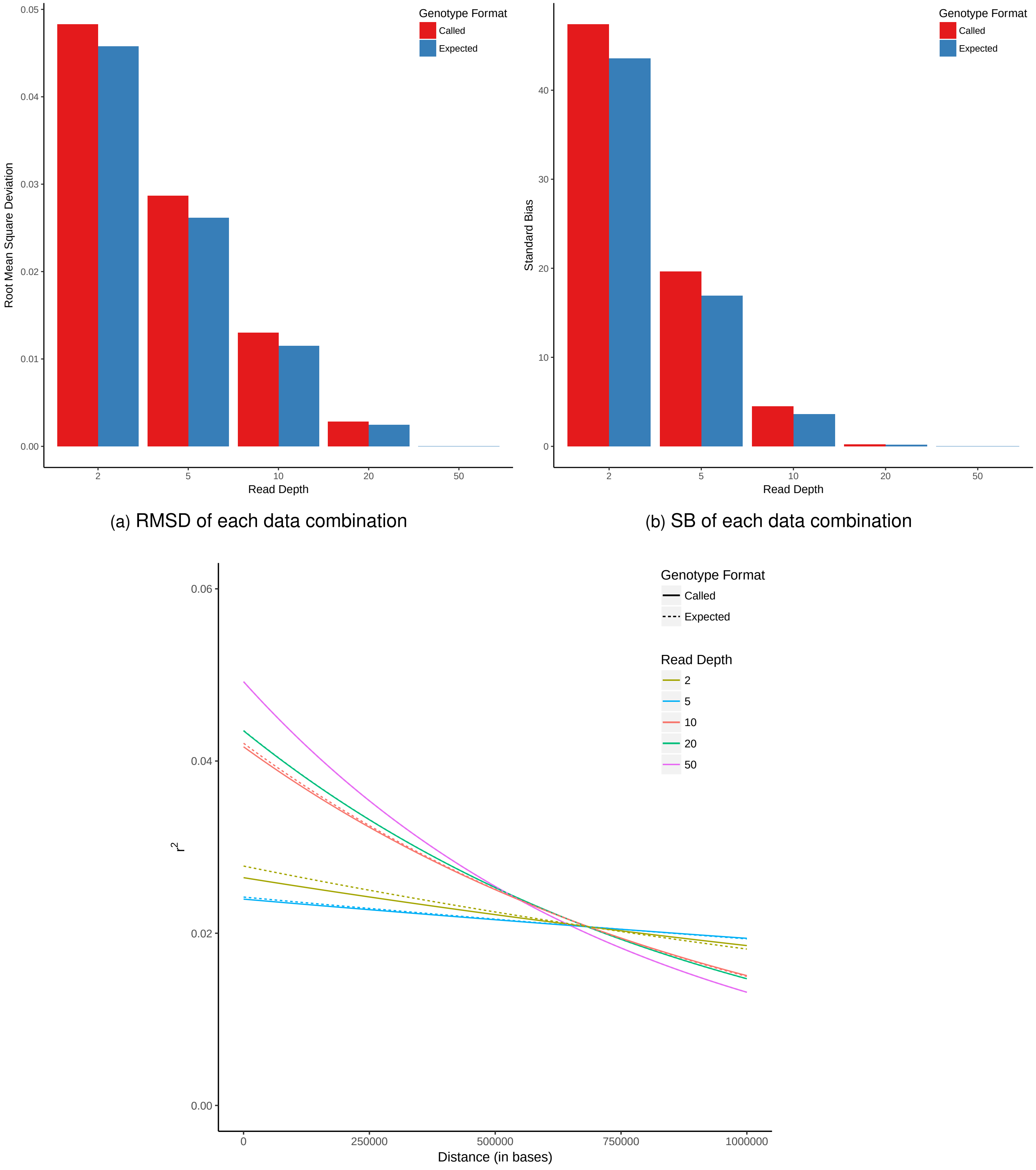
4 Results

4.1 Quantifying Accuracy

At each hypothetical sequencing depth, using the ngsLD expected genotype method to calculate LD gave measurements that were closer to the 'true' data than the LD measurements calculated using called genotypes, particularly at low depths (Figure 1). For both expected genotypes and genotype posterior probabilities in terms of both RMSD and SB, the LD measures estimated from the genotype likelihoods with a read depth of 50 were nearly identical to the 'true' LD measures derived from the called genotypes sequenced at a depth of 50 reads with the difference in LD measured increasing with read depth (Figures 1a&b and 2a&b).

The exponential decay curves for the data set using expected genotypes and genotype posterior probabilities with read depths of 50, 20, and 10 showed a reasonably similar pattern to the 'true' data set of called genotypes from 50 reads (Figures 1c&2c). However, using a read depth of 2 or 5 gives LD data which noticeably underestimates both the y intercept and slope of the 'true' curve. The effect is actually more dramatically off for the 5 read data sets than for the 2 read data sets in the case of expected genotypes, though.

Using genotype posterior probabilities tends to give higher estimates of LD as opposed to using called genotypes with the same r^2 calculation algorithm (Vieira et al., 2015). This is evidenced by the positive SB for the genotype posterior probabilities and negative SB for the called genotypes in Figure 2b. It can also be observed in the offset of the curves in Figure 2c. Even though this bias leads to a bigger difference between the called genotypes and posterior genotype probability measures, the overall bias when using the EM algorithm is significantly lower.

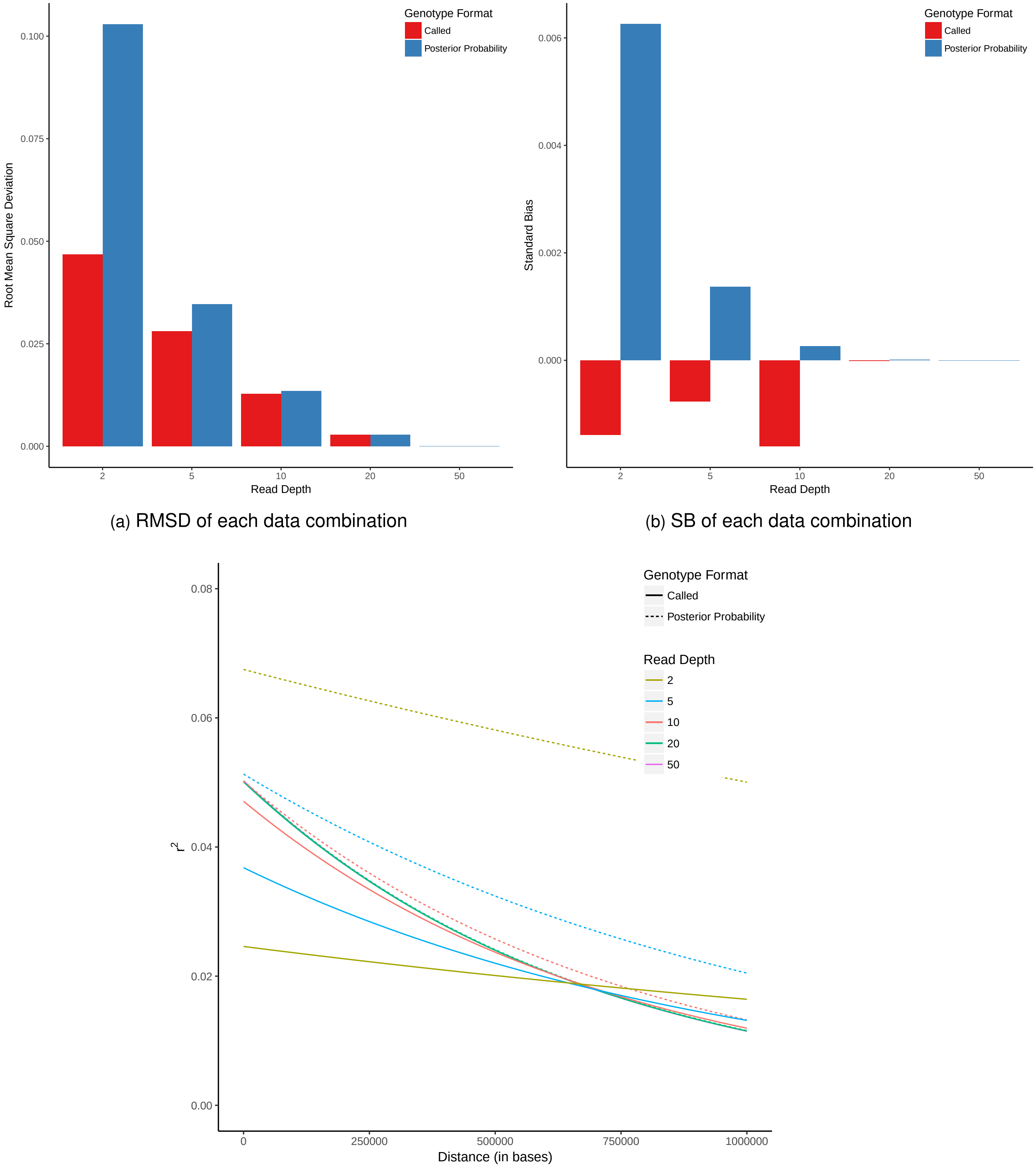


(a) RMSD of each data combination

(b) SB of each data combination

(c) Fit LD decay curves for each data combination

Figure 1: Results comparing the accuracy of LD measurement for the simulated population at read depths of 2, 5, 10, 20, and 50 using both called genotypes and expected genotypes. RMSD and SB values are in comparison to r^2 values for the same pair of SNPs in the called genotype data set from 50 reads.



(a) RMSD of each data combination

(b) SB of each data combination

(c) Fit LD decay curves for each data combination

Figure 2: Results comparing the accuracy of LD measurement for the simulated population at read depths of 2, 5, 10, 20, and 50 using both called genotypes and genotype posterior probabilities. RMSD and SB values are in comparison to r^2 values for the same pair of SNPs in the called genotype data set from 50 reads.

4.2 Case Study 1: Turkeys and Ducks

The LD exponential decay curve of the turkey had both a higher intercept and steeper slope as compared to the duck decay curve, with the 90% confidence intervals only overlapping as the curves begin to approach 0 at exceptionally far distances (Figure 3).

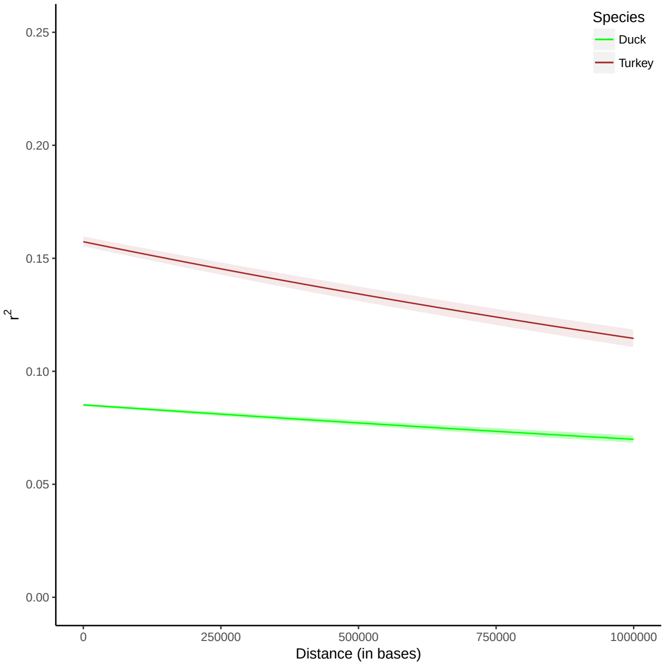
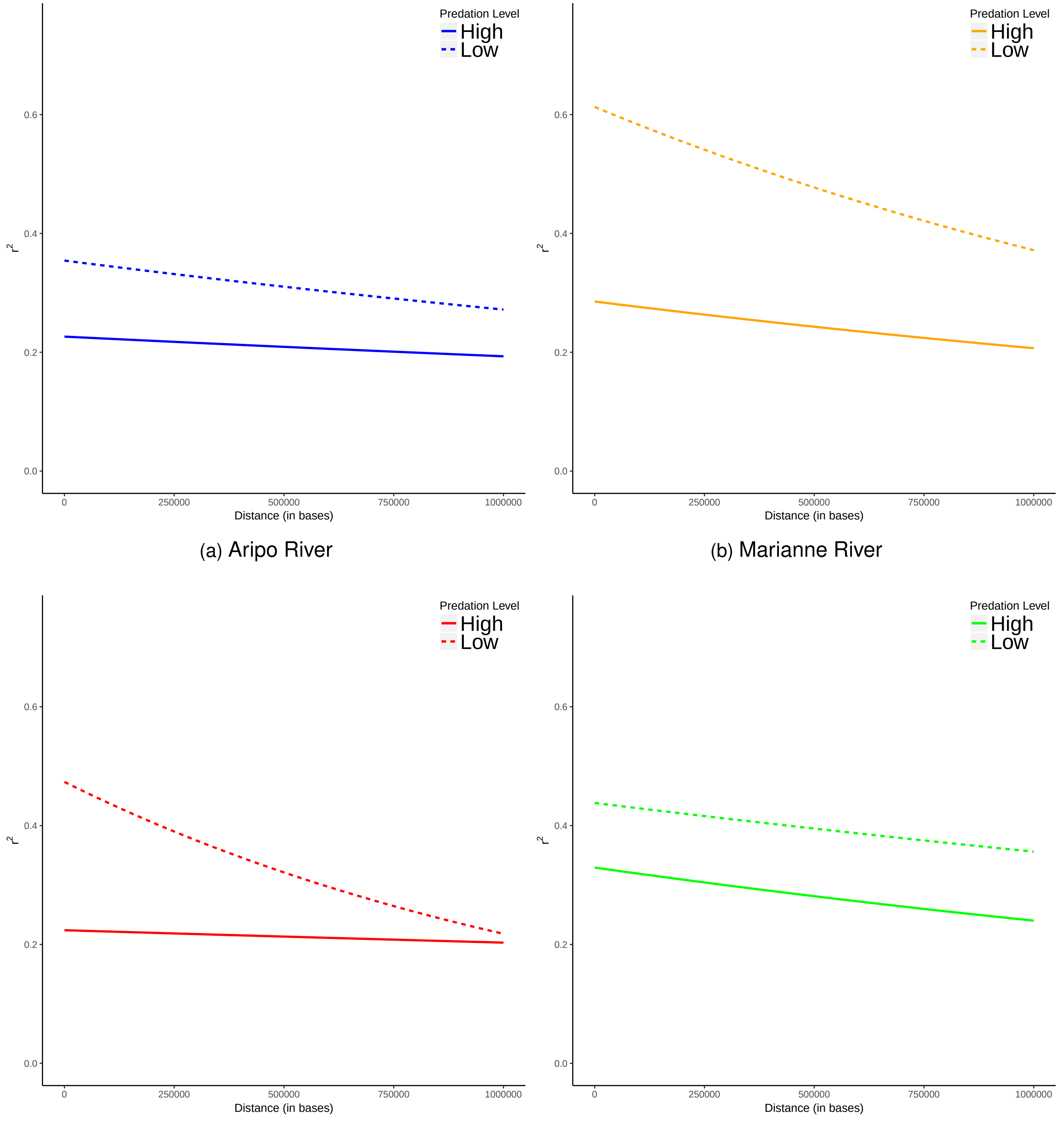


Figure 3: Comparison of exponential decay of LD between duck and turkey populations. Shaded region represents 90th percentile of curves for data when bootstrap was performed. Equation of the duck curve: r^2 = 0.068 * $e^{-2.301e-7*Distance}$. Equation of the turkey curve: r^2 = 0.116 * $e^{-2.631e-7*Distance}$

4.3 Case Study 2: Guppies

In each river and for both test distances of up to 1 megabase pair (Figure 4) and 200 kilobase pairs (Figure 5), the less predated upstream population exhibited higher levels of LD than the other population in the same river. The slope of curves fit to data from the less predated subpopulations, with SNPs up to 1 megabase apart considered, were noticeably steeper than the highly predated subpopulations for the same river with the exception of the Yarra subpopulations. However, the data from SNP pairs up to 200 kilobases apart did not show a similar trend with only the less predated subpopulations in the Aripo and Quare rivers having a higher slope than their neighbor subpopulation. In the Marianne and Yarra river, the highly predated subpopulations had a steeper slope.



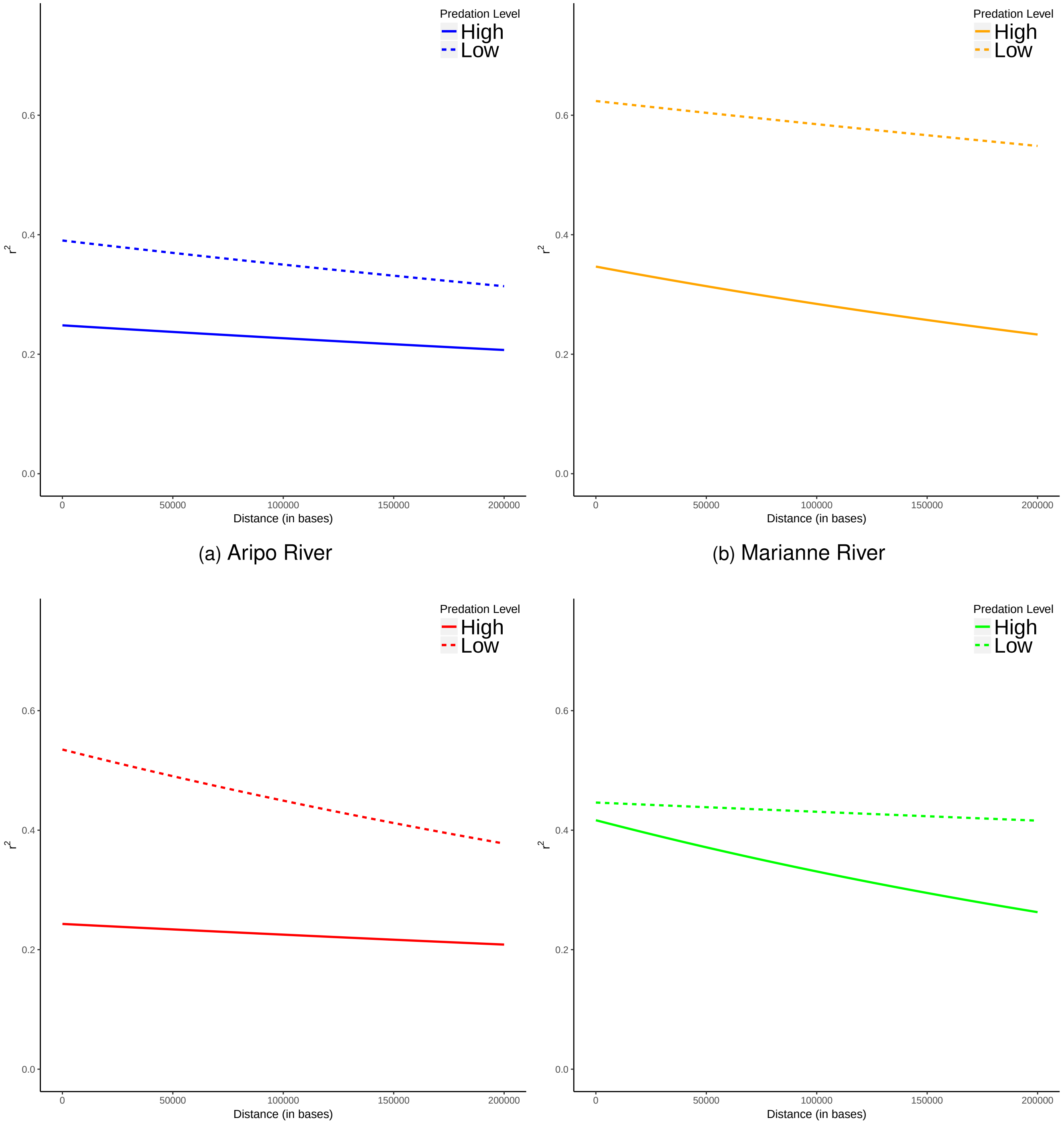
(a) Aripo River

(b) Marianne River

(c) Quare River

(d) Yarra River

Figure 4: Plots show a comparison of LD decay curves for the high-predation/downstream (solid line) and low-predation/upstream (dashed line) populations for SNPs up to 1 megabase pairs apart



(a) Aripo River

(b) Marianne River

(c) Quare River

(d) Yarra River

Figure 5: Plots show a comparison of LD decay curves for the high-predation/downstream (solid line) and low-predation/upstream (dashed line) populations for SNPs up to 200 kilobase pairs apart

5 Discussion

5.1 Quantifying Accuracy

By using the various quality and probability data available, ngsLD presents a way to calculate LD that is flexible enough to be used on any read data without needing to establish a threshold to call a genotype in that sample. Low read depth increases the chance of error in base, SNP, and genotype calling which can lead to inaccurate measures of LD (Nielsen et al., 2011). While the read depth had a greater impact on LD accuracy than which type of genotype data was used, it is important to note that the calculation method used in ngsLD are still an improvement over traditional methods and that this improvement is more pronounced as read depth decreases. So, even though ngsLD cannot completely surmount the difficulties associated with genotype determination at low read depth, it does increase the sophistication with which LD can be calculated on all data sets.

The results of the LD decay curves continued the pattern of error between the observed and ’true’ data increasing with decreasing error rate. There was a notable exception in that the curves from the 2 reads data sets were closer to the ’true’ curve than the curves from the 5 reads sets when using expected versus called genotypes. One of the reasons accuracy decreases with read depth is that, at particularly low depths, the chance of detecting the allele from both chromosomes in a diploid is greatly reduced (Nielsen et al., 2011; Vieira et al., 2015). This leads to an inflated estimate of homozygosity in the population. This could depress the slope of the LD decay curve because, for example, if an individual is homozygous at 2 loci, recombination won’t make a detectable change in haplotype frequency. Therefore, an over-estimation of homozygosity could dampen the effect of distance. Having just 2 reads, however, may yield a lower overall homozygosity because it becomes more difficult to determine if a heterozygote is actually a heterozygote or simply the result of a sequencing error. In this way, an inflation of heterozygosity at some loci could slightly cancel out the over-estimation of homozygosity at others, making the curves generated from data sets of 2 reads closer to the ’true’ curves than those from the 5 reads data sets. Vieira et al. (2015) similarly found the lower depth to be less biased than the middle depth they tested. However, genotype posterior probabilities do a better job of dealing with these errors and therefore the curves show the trend of decreasing quality with decreasing read depth.

5.2 Case Study 1: Turkeys and Ducks

The curve from the turkey population having a higher strength of LD up to 10MB and steeper slope than the duck population curve is consistent with the results found in Wright et al. (2015). In that study, the effective population size (N_e) of duck population was estimated to be 3 times larger than that of the turkey population studied. Smaller populations are more greatly affected by genetic drift which increases overall LD in the genome (Slatkin, 2008). This could explain why the turkey population showed higher levels of LD. The slope of the turkey curve is steeper than the slope of the duck curve, though. A steeper curve could simply be from the turkeys having a naturally higher rate of recombination than the ducks. In this case, the small population size would explain the level of LD and the recombination rate would explain the decay rate. However, the steeper slope could also be indicative of a larger population because larger populations have a greater total number of recombination events per generation (Park, 2012). This higher chance of recombination makes the strength of linkage weaker which is compounded by the distance between SNPs, leading to a steeper slope. An alternate hypothesis, therefore, is perhaps the turkeys experienced a bottleneck in the past (most likely when the domestic population was established) which increases the overall level of LD, but have recently been experiencing an expansion which would lead to the strength of LD decaying quicker as distance increases. Unfortunately, there is not much data available on the history of these captive populations to verify either of these hypotheses (Wright et al., 2015).

5.3 Case Study 2: Guppies

The upstream guppy subpopulations in these rivers are started by a few individuals that manage to reach upstream pools from downstream pools(Ghalambor et al., 2004). The consistently higher LD in the upstream/less-predated subpopulations can be explained by this founders effect which acts similarly to a population bottleneck (Slatkin, 2008). Work done by Vicencio Oostra on the same data set (Wright et al., 2017), showed a similar pattern of the less-predated populations having consistently higher levels of LD when SNPs up to 10 megabase pairs and 100 kilobase pairs are considered (personal communication, 1 August 2017). Previous work by Fraser et al. (2015) found that guppy subpopulations in less-predated areas had shown elevated population expansion since they were founded in comparison their source populations. The noticeably steeper curves found for these populations could be due to a similar method as suggested for the turkeys above. Although the founder effect creates a population with skewed haplotype proportions, the lack of predators allows the population allows the population to rapidly expand. This expansion, in turn, begins to break up linked loci quicker than a population at a constant size (Park, 2012).

6 Conclusion

In addition to being more accurate than traditional LD measurement methods, application of ngsLD and the developed decay-curve fitting methods detected the previously proven population patterns in each case study. This preliminary analysis of population history by this study represents the beginning of the possible applications for this new method. While it cannot completely make up for a lack of read depth, ngsLD can calculate the most accurate possible measures of LD and is flexible enough in requirements to be used on a variety of systems, including non-model species. Future work can be done with this package to explore the issue of which range of SNPs to consider, as well as looking into to quantifying the correlations between decay coefficients and populations parameters.

Acknowledgements

The author would like to thank Matteo Fumagalli (Imperial College London) for supervising this research and the many others who contributed their data or expertise including Filipe G Vieira (University

of Copenhagen), Alison Wright (Sheffield University), and Vicencio Oostra (University College London).

Funding

Funding for this project was provided by the Imperial College London postgraduate Life Sciences department as part of a grant available to all students in the course.

References

A. M. Bolger, M. Lohse, and B. Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

D. Clayton and H.-T. Leung. An r package for analysis of whole-genome association studies. *Human heredity*, 64(1):45–51, 2007.

B. A. Fraser, A. Künstner, D. N. Reznick, C. Dreyer, and D. Weigel. Population genomics of natural and experimental populations of guppies (poecilia reticulata). *Molecular Ecology*, 24(2):389–408, 2015.

M. Fumagalli, F. G. Vieira, T. S. Korneliussen, T. Linderoth, E. Huerta-Sánchez, A. Albrechtsen, and R. Nielsen. Quantifying population genetic differentiation from next-generation sequencing data. *Genetics*, 195(3):979–992, 2013.

C. K. Ghalambor, D. N. Reznick, and J. A. Walker. Constraints on adaptive evolution: the functional trade-off between reproduction and fast-start swimming performance in the trinidadian guppy (poecilia reticulata). *The American Naturalist*, 164(1):38–50, 2004.

P. W. Harrison, A. E. Wright, F. Zimmer, R. Dean, S. H. Montgomery, M. A. Pointer, and J. E. Mank. Sexual selection drives evolution and rapid turnover of male gene expression. *Proceedings of the National Academy of Sciences*, 112(14):4393–4398, 2015.

D. L. Hartl, A. G. Clark, and A. G. Clark. *Principles of population genetics*, volume 116. Sinauer associates Sunderland, 1997.

R. D. Hernandez, M. J. Hubisz, D. A. Wheeler, D. G. Smith, B. Ferguson, J. Rogers, L. Nazareth, A. Indap, T. Bourquin, J. McPherson, et al. Demographic histories and patterns of linkage disequilibrium in chinese and indian rhesus macaques. *Science*, 316(5822):240–243, 2007.

R. R. Hudson. Generating samples under a wright–fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.

D. Kim, B. Langmead, and S. L. Salzberg. Hisat: a fast spliced aligner with low memory requirements. *Nature methods*, 12(4):357–360, 2015.

T. S. Korneliussen, A. Albrechtsen, and R. Nielsen. ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, 15(1):356, Nov. 2014. ISSN 1471-2105. doi: 10.1186/s12859-014-0356-4. URL `http://www.biomedcentral.com/1471-2105/15/356/abstract`.

R. Lewontin. The interaction of selection and linkage. i. general considerations; heterotic models. *Genetics*, 49(1):49, 1964.

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics*, 39(7):906, 2007.

A. F. McRae, J. M. Pemberton, and P. M. Visscher. Modeling linkage disequilibrium in natural populations: the example of the soay sheep population of st. kilda, scotland. *Genetics*, 171(1):251–258, 2005.

M. Newville, A. Nelson, T. Stensitzki, A. Ingargiola, D. Allan, Y. Ram, C. Deil, G. Pasquevich, T. Spillane, P. A. Brodtkorb, et al. Lmfit: non-linear least-square minimization and curve-fitting for python. *Astrophysics Source Code Library*, 2016.

R. Nielsen, J. S. Paul, A. Albrechtsen, and Y. S. Song. Genotype and snp calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443–451, 2011.

R. Nielsen, T. Korneliussen, A. Albrechtsen, Y. Li, and J. Wang. Snp calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PloS one*, 7(7):e37558, 2012.

L. Park. Linkage disequilibrium decay and past population history in the human genome. *PloS one*, 7 (10):e46603, 2012.

J. K. Pritchard and M. Przeworski. Linkage disequilibrium in humans: models and data. *The American Journal of Human Genetics*, 69(1):1–14, 2001.

D. E. Reich, M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward, et al. Linkage disequilibrium in the human genome. *Nature*, 411(6834):199–204, 2001.

S. M. Ross. *Introduction to probability models*. Academic press, 2014.

M. Slatkin. Linkage disequilibrium: understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477–485, 2008.

M. Slatkin, L. Excoffier, et al. Testing for linkage disequilibrium in genotypic data using the expectation-maximization algorithm. *Heredity*, 76(4):377–383, 1996.

F. G. Vieira, F. Lassalle, T. S. Korneliussen, and M. Fumagalli. Improving the estimation of genetic distances from next-generation sequencing data. *Biological journal of the Linnean Society*, 117(1): 139–149, 2015.

H. Wickham. *ggplot2: elegant graphics for data analysis*. Springer, 2016.

A. E. Wright, P. W. Harrison, F. Zimmer, S. H. Montgomery, M. A. Pointer, and J. E. Mank. Variation in promiscuity and sexual selection drives avian rate of faster-z evolution. *Molecular ecology*, 24(6): 1218–1235, 2015.

A. E. Wright, I. Darolti, N. I. Bloch, V. Oostra, B. Sandkam, S. D. Buechel, N. Kolm, F. Breden, B. Vicoso, and J. E. Mank. Convergent recombination suppression suggests role of sexual selection in guppy sex chromosome formation. *Nature Communications*, 8:14251, 2017.