UPPSALA UNIVERSITET

MSc BIOINF 12 002

Microsatellites in the flycatcher genome

Sanea Sheikh

Degree project in bioinformatics, 2012
Examensarbete i bioinformatik 30 hp till masterexamen, 2012
Biology Education Centre and Department of Ecology and Genetics, EBC, Uppsala University
Supervisor: Hans Ellegren

Abstract

The collared flycatcher (*Ficedula albicollis*) and the pied flycatcher (*F. hypoleuca*) represent a sister species model which has been studied in terms of evolutionary ecology at Uppsala University for several decades. Their tendency to hybridize where they co-occur makes them an interesting model system for the investigation of speciation and hybridization. Very little is known about the differential expression between species and information about this can lead to better understanding about the speciation process and the species differences. For the two flycatcher species, genome and transcriptome data have recently been acquired using Illumina sequencing. The genome sequence of a collared flycatcher is currently being assembled.

In this study I identified the microsatellites, their degree of polymorphism and genetic differentiation in the collared and the pied flycatchers. I tested different software that can be used to identify microsatellites and identified SciRoKo to be the most appropriate one for identifying microsatellites in the genome assembly, coding regions and untranslated regions, in particular. I used several other bioinformatics tools such as Novoalign, SAMTools and BEDTools to identify the degree of polymorphism in terms of expected heterozygosity in the flycatchers. I used the allele frequency data to estimate the genetic differentiation in terms of F_{st} in the pied and the collared flycatchers. I also compared the microsatellites in the flycatcher genome to the microsatellites in the zebra finch genome.

I found that there are more than 7 million microsatellites in the flycatcher genome. The number of microsatellites decreases with an increase in the number of repeat units and the number of nucleotides making up the microsatellite motif. As compared to the 7 million microsatellites in the whole genome, there are only 65,000 microsatellites in the coding regions including the untranslated regions. Due to different techniques used for sequencing the zebra finch genome and the flycatcher genome, there was a huge difference in the total number of microsatellites in the zebra finch genome as compared to the flycatcher genome. Occurrence of polymorphism in the flycatcher genome cannot be determined in a “yes” and “no” manner, depending on the number of reads mapping on to two different alleles. However, an estimation of expected heterozygosity can help determine the degree of polymorphism. The allele frequency data for the microsatellite loci that were common between the two species were used to estimate the genetic differentiation in terms of F_{st}. However, when expected heterozygosity, deviation from the expected heterozygosity and F_{st} estimations are mapped onto the chromosomes, a clear pattern cannot be observed like the one observed with single nucleotide polymorphism data by other members of the group. This might be due to a high frequency of noise in the microsatellite data.

Microsatellites in the flycatcher genome

Popular Science Summary

Sanea Sheikh

Collared flycatcher (*Ficedula albicollis*) and pied flycatcher (*Ficedula hypoleuca*) are two species residing in Sweden that diverged only about one million year ago. They are similar to an extent that they intermix and produce hybrids. However, the fitness of offspring produced by mixed couples is much lower than that of the pure species and the hybrid females are even sterile. Questions, such as why the fertility of mixed couples is reduced and what makes the individuals within the species and between species different from each other, have attracted considerable interest among evolutionary biologists. In order to find answers to these questions, the genome of the two species has recently been sequenced in the supervisor's laboratory.

I used the flycatcher genome sequences to find information about so called microsatellites and how the number and length of microsatellites differ within species as well as between the two species. Microsatellites are arrays of short, tandemly repeated DNA motifs found throughout the genome of eukaryotes. Studies of the evolutionary dynamics of microsatellites can be useful for understanding the pattern of molecular evolution. Comparative studies of microsatellites in the pied and collared flycatcher can help us understand the pattern of species divergence. I also used the results to compare the total number of microsatellites in flycatcher and the zebra finch which diverged about 40 million years ago.

I found more than 7 million microsatellites with at least five repeat units in the collared flycatcher genome assembly. The number of microsatellites gradually decreases with an increasing number of repeat units for a particular microsatellite motif. Zebra finch has more than 9 million microsatellites. This might not necessarily be due to a biological reason but due to the fact that it is harder to assemble large repeats with short reads generated through the sequencing technologies used in case of zebra finch. Of all these 7 million microsatellites, about 65,000 are found within coding regions including the untranslated regions of the flycatcher genome. The degree of polymorphism and the genetic distance between the collared and the pied flycatcher was also measured. These were then mapped on the chromosomes. The plots of this mapping show no clear pattern indicating that there is a lot of noise in the microsatellite data.

With this study we have nevertheless come a lot closer to identifying how the two species are different on the basis of microsatellites. We have also found out the difference in the microsatellites that occur within the species. Together with studies on other genetic differences between the species we will hopefully very soon have a conclusive picture on this question.

Degree Project, Master Program in Bioinformatics (30 hp), Spring 2012, Uppsala University.
Department of Evolutionary Biology, EBC, Uppsala University.
Supervisor: Hans Ellegren

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Microsatellites

Microsatellites are arrays of short, tandemly repeated DNA motifs (1-6bp) found throughout the genomes of eukaryotes (Buschiazzo et al., 2006). A repeat motif of two bases is referred to as a dinucleotide repeat, whereas the terms, tri-, tetra- and pentanucleotides refer to repeat motifs consisting of three, four and five repeat units, respectively (Brohede, 2003). Microsatellites can further be divided into two different categories: perfect microsatellite and imperfect or interrupted microsatellite. When a repeat tract contains a continuous stretch of one motif, for example in the sequence CACACA, the microsatellite is called a perfect microsatellite. However, if one or more base pair substitutions occur in a pure repeat tract, for example in case of CAGACA, it is called an imperfect or interrupted microsatellite.

1.1.1 Microsatellite distribution

Microsatellites may represent a significant part of the genome, for example about 3% of the human genome. Dinucleotide repeats comprise of 0.5% of the genome, making them the most common class of microsatellites in humans (Consortium IHGS, 2001; Brohede, 2003; Leclerq et al., 2007). A closer examination of the dinucleotide microsatellites reveals that 50% of them are CA repeats, 35% are AT repeats and 15% are AG repeats whereas GC repeats are only 0.1% of the dinucleotides (Brohede, 2003; Ellegren, 2004). Microsatellites seem to be equally common in intergenic regions and introns (Toth et al., 2000, Ellegren, 2004). Microsatellite density is affected by base composition, however, there is a regional variation in microsatellite density which cannot be attributed to base composition (Bachtrog, et al., 1999, Ellegren, 2004). For example, in case of the human and mouse genomes, there is almost a twofold increase in microsatellite density near the ends of chromosome arms (Mouse Genome Sequencing Consortium, 2002, Ellegren, 2004). Microsatellites are also found in the 5' UTR, but are generally rare within protein-coding regions, suggesting that at least some of these microsatellites have regulatory properties (Morgante et al., 2002; Brohede, 2003).

1.1.2 Polymorphism at microsatellite loci

The length of microsatellites may vary due to insertions and deletions of one or more repeat unit. The degree of polymorphism at a microsatellite locus is both species and locus specific (Amos et al., 1996; Harr et al., 1998; Ellegren, 2000b) but there is a general trend toward a higher polymorphism in longer microsatellites (Weber 1990). Polymorphism in a microsatellite located within the coding region can cause a change in the protein properties, whereas polymorphism in a microsatellite located within the 5' UTR can result in a change in expression of the associated protein.

Mutation and thereby polymorphism at microsatellite loci are thought to be due to slippage (Figure 1) (Goldstein et al., 1999, Ellegren, 2004, Leclerq et al., 2007). During replication the template strand and the newly synthesized strand temporarily dissociate from each other only to re-associate a fraction of second later. If this occurs when a repeat region is being replicated, a repeat unit on the nascent strand can re-associate out-of-frame to an incorrect repeat unit on the template strand, resulting in a change in length of the microsatellite in the newly synthesized strand (Brohede, 2003). This will result in the construction of a loop, which will either be excised or filled in after a single strand break on the opposite strand. As a result a new mutation will be established if the excision or filling is done on the wrong strand. A loop on the nascent strand that is filled in will result in an insertion mutation, while an excision on the template strand will result in a deletion mutation.

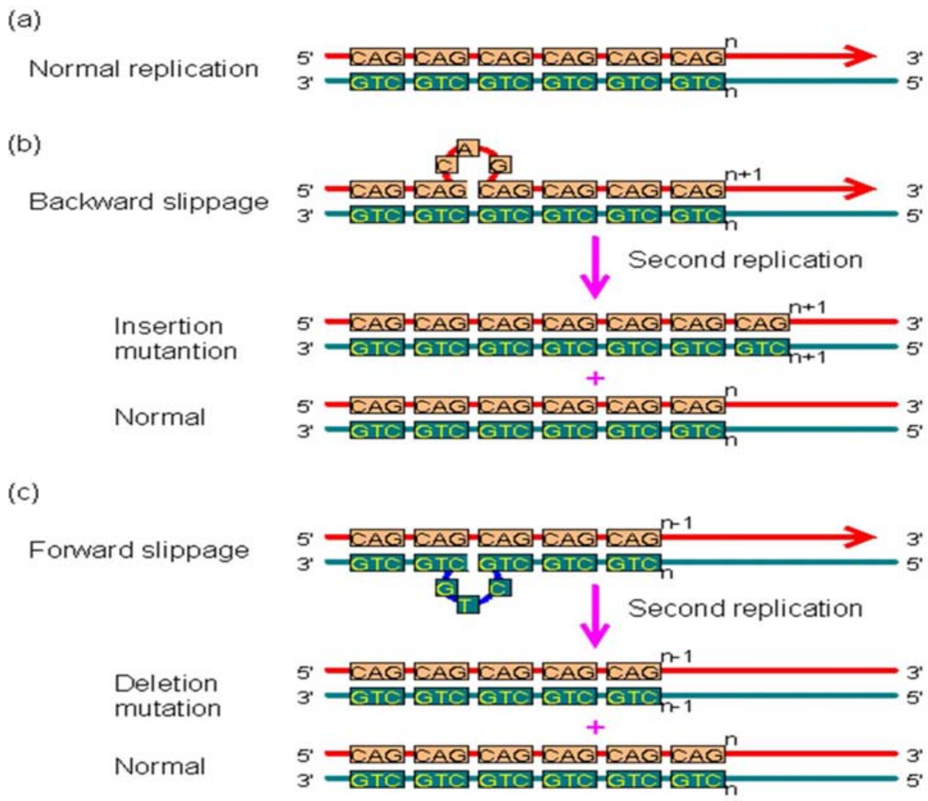

Figure 1: Polymorphism in microsatellite caused by replication slippage.

(a) Normal replication at microsatellite loci. (b) Backward slippage resulting in increase in the number of repeat units. (c) Forward slippage resulting in decrease in the number of repeat units [1].

1.1.3 Importance of microsatellites

Microsatellites have attracted great interest among biologists, mainly due to their potential role in molecular functions such as recombination (Benet, et al., 2000, Buschiazzo, et al., 2010) or regulation of transcription factors (Martin, et al., 2005, Buschiazzo, et al., 2010), in neurodegenerative disorders (Mitas M. 1997, Buschiazzo E. et al., 2010) and in some forms of cancers (Arzimanoglou, et al., 1998, Buschiazzo, et al., 2010; Goldstein, et al., 1999, Jarne, et al., 1995, Leclercq, et al., 2007). However, microsatellites have attracted the widest interest as polymorphic, neutral genetic markers for population genetics, gene mapping, forensics and paternal investigation (Schlotterer, et al., 2004, Buschiazzo, et al., 2010).

Most microsatellites are thought to be selectively neutral. A high rate of mutation results in a high rate of polymorphism within the population (Buschiazzo, et al., 2010, Leclercq, et al., 2007, Ellegren, 2004). However, in case of closely related species, most microsatellites are retained but as the evolutionary distance increases, less microsatellites are retained (Leclercq, et al., 2007, Buschiazzo, et al., 2010). Studies of the evolutionary dynamics of microsatellites can be useful in understanding the pattern of molecular evolution (Buschiazzo, et al., 2010). Also, besides addressing questions relating to microsatellite evolution, comparative studies of microsatellites in recently diverged species can help to understand the pattern of species divergence. Microsatellites have extensively been used for the analysis of population structure, both for studies of sub-populations within a single species and to determine the evolutionary relationship between species. Considering the difference in the mutation pattern seen at different loci in different species, only microsatellite loci with the same properties are suggested to be used for these studies (Landry, et al., 2002). Microsatellite data has also been used to measure the selective sweeps (Wiehe, 1998; Schlotterer, 2002) and for measuring the level of inbreeding (Coulson et al., 1998).

1.1.4 Problems with microsatellite identification

The definition of the minimum number of iterations needed for a repetitive structure to be referred to as a microsatellite can be complicated. No real consensus has been reached on whether to use a minimum number of base pairs or a minimum number of repeat units when referring to microsatellites (Ellegren, 2004). A further complication to microsatellite identification and characterization has been added by the lack of consensus on the amount of degeneracy that can be allowed in characterization of a slightly imperfect repetitive structure as a microsatellite (Ellegren, 2004).

1.1.5 Algorithms for microsatellite identification

Different algorithms have been developed over the years that use different criteria for identification and characterization of microsatellites. These criteria range from the number of base pairs/repeat units (Kruglyak, et al., 1998, Calabrese, et al., 2003, Bell, et al., 1997, Leclercq, et al., 2007) to the amount of degeneracy, motif types (Sainudiin, 2004, Leclercq, et al., 2007) and minimum distance between successive microsatellites (Bell, et al., 1997, Sainudiin, 2004, Leclercq, et al., 2007). The recent algorithms allow the user to define these criteria when characterizing microsatellites in genomic sequences, such as the number of repeat units and the type of microsatellite. Popular software systems that implement these recent algorithms for microsatellite identification include Sputnik [3], SciRoKo [4] and RepeatMasker [5], among many others.

1.2 Pied and collared flycatcher

Old world flycatchers belong to the family Muscicapidae. Collared flycatcher (*Ficedula albicollis)* and pied flycatcher (*F. hypoleuca)* form a sister species model for ecological and evolutionary research that has been carried out in Uppsala since 1980s (Alatalo, et al., 1981; Qvarnstrom, et al., 2010). The two species of flycatcher diverged about 1 million year ago (Qvarnstrom, et al., 2010) but occur in widely overlapping ranges in Central and Eastern Europe at present. Hybridization occurs regularly in these overlapping ranges, however, the hybrids produced have reduced fitness (Alatalo, et al., 1982; Svedin, et al., 2008). It is suggested that certain loci in these species result in hybrid incompatibilities (Backstrom, et al. 2010). The complete genome of the two species has recently been sequenced in the supervisor’s laboratory using the Illumina platform for next generation sequencing. The assembly was generated from sequencing of a single male collared flycatcher to a mean depth of coverage of 60x. In addition, the genome sequence for 9 other collared and 10 pied individuals has also been generated with coverage of 5x. The complete genome sequence and annotation from these species would help in finding the mating barriers that prevent these species to fully admix (Uebbing, personal communication).

This project aims at identification of microsatellites in the flycatcher genome and characterization of degree of polymorphism in the two flycatcher species followed by a comparison of heterozygosity distribution over the chromosomes for both pied and collared flycatcher.

2 Materials and Methods

2.1 Flycatcher genome sequence

The genome of the collard flycatcher was sequenced, in the supervisor's laboratory, in the form a de novo assembly where a large number of short reads, generated from the Illumina platform, were put together. All the reads were generated using paired-end and mate-pair sequencing. The depth of coverage was about 60x, which means that each base was covered by 60 reads on average.

Multiple unrelated individuals of both collared (9 individuals) and pied flycatcher (10 individuals) species were also sequenced to a much lower coverage of 5x (meaning that each base was covered by 5 reads on average). In this case, it was unlikely to have both the alleles read at the heterozygous sites as opposed to the high coverage collared individual in which it was more likely to have both the alleles at a heterozygous site.

2.2 Software for microsatellite identification and characterization

Sputnik and SciRoKo were tested to identify the most appropriate algorithm for identification and characterization of microsatellites in the flycatcher genome. Due to its speed, user friendly options and the ability to handle a large amount data (Kofler, et al., 2007), SciRoKo was selected for the identification and characterization of microsatellites. The "Perfect Repeat" model, which is used to identify perfect microsatellites, with a minimum repeat unit of 5 was selected to identify mono, di, tri and tetra nucleotide motifs. This means that all the microsatellites that are perfect and had 5 or more repeat units long would be identified using SciRoKo.

2.3 Identification of microsatellites in the flycatcher genome

2.3.1 Total number of microsatellites in the flycatcher genome

SciRoKo was used to identify all the microsatellites in the assembly. Perl scripts were used to extract all the microsatellites in terms of their type (mono-, di-, tri-, tetra- nucleotide), and in terms of motifs. The motifs of the microsatellites were grouped together based on the change in reading frame and the strand on which they are read. For example, a motif ATG was grouped with the motifs TGA, GAT, TAC, ACT and CTA.

To have an idea on how the total number of microsatellites in the flycatcher genome varies with different thresholds for the minimum number of repeat units, the threshold was changed from 5 to 8 and 10. The results were then compared to have an idea on the difference in the total number of microsatellites

2.3.2 Genomic distribution of repeat length per repeat motif in the flycatcher genome

The microsatellite data that was generated using SciRoKo was then further analyzed using different Perl scripts to identify how common it is to find a particular microsatellite motif of a particular length in the flycatcher genome, that is, to determine the genomic distribution of repeat length per repeat motif. The scripts were used to extract all microsatellites of a particular length and then group them into different categories based on the motif type. For example, all microsatellites of a length 5 were extracted by the script. It then classified the microsatellites into mono-, di-, tri- or tetra- nucleotide motifs, and the motifs were subsequently further classified into different groups based on the sequence, reading frame and the strand.

2.3.3 Microsatellites within coding sequences including untranslated regions

Since the genome of the flycatcher has recently been sequenced, there were no annotations available online. Therefore, in order to identify the microsatellites within the coding sequences including the UTRs, GFF files (Generic Feature Format files that store genomic features in a text file) were created by different group members. These files were created using MAKER which is a pipeline that identifies repeats, aligns ESTs and proteins to a genome, produces *ab initio* gene predictions and automatically synthesizes these data into gene annotations having evidence-based quality values (Cantarel, 2007). The GFF files, therefore, contained information about gene predictions in the flycatcher genome.

The sequences corresponding to coding regions and untranslated regions were extracted from the assembly based on the coordinates given in the GFF files. SciRoKo was then used to identify the microsatellites with the “Perfect Repeat” model and a threshold of 5 for the minimum number of repeat units. Perl scripts were used to generate classify the microsatellite data generated from SciRoKo into categories and types of microsatellites. The variation in the total number of microsatellite was also observed by changing the threshold for the minimum number of microsatellites in SciRoKo. The distribution of repeat length per repeat motif within these regions was estimated using the scripts that were used earlier for the estimation of the genomic distribution of repeat length per repeat motif.

2.4 Comparison between the flycatcher and zebra finch genome based on microsatellite data

Zebra finch and flycatchers are two related species that diverged about 40 million year ago (Cracraft, 2009). Both zebra finch and flycatcher are model organisms that are used for studies on birds. The genome of zebra finch has been sequenced using conventional Sanger sequencing techniques.

2.4.1 Total number of microsatellites in the zebra finch genome

It was of interest to compare how the number and length distribution of microsatellites varied between the two species. The same procedure that was used to identify the total number of microsatellites in the flycatcher genome was repeated for zebra finch genome with the goal of identification of degree of microsatellite conservation between flycatcher and zebra finch and to have an idea about how the numbers vary between the two species.

2.4.2 Genomic distribution of repeat length per repeat motif in zebra finch genome

The genomic distribution of repeat lengths per repeat motif was also estimated using the same procedure that was used for estimating the genomic distribution of repeat lengths per repeat motif for the flycatcher genome. The goal of this step was to identify the extent to which microsatellite length is conserved between the flycatcher and the zebra finch.

2.5 Identification of polymorphism at the microsatellite loci

2.5.1 Polymorphism in the genome assembly

The next step was to identify whether a particular microsatellite locus, identified in the genome assembly, was polymorphic or not, that is, whether or not the individual used for genome sequencing was heterozygous with respect to the number of repeat units. The loci with mononucleotide motifs were excluded from all the steps relating to identification of polymorphism at the microsatellite loci for the sake of clarity. “All microsatellite loci” from now on means all microsatellite loci except the ones having mono-nucleotide motifs.

Individual reads were mapped back to the assembly to identify whether two length variants for a particular microsatellite were present or not. Novoalign was used for this purpose due to its accuracy in short read alignment using fast *k*-mer index searching with dynamic programming [2]. Although Novoalign is slower than other alignment tools such as Burrows-Wheeler Aligner, it is more accurate and sensitive because it uses full dynamic programming to find the best alignment of a short read to the genome sequence [2]. The read file format provided to Novoalign was fastq with Illumina coding of quality values (ILMFQ). Random strategy was selected for reporting the repeats. SAM was the selected report format.

After the completion of alignment by Novoalign, SAMtools view option was used to extract reads aligning to all the microsatellite loci. The read IDs and sequences were extracted for the reads that mapped onto a microsatellite with unique sequence at both ends. This ensured that only the long complete reads were considered in this study and not short reads that might have only aligned to a part of the microsatellite on the assembly.

SciRoKo was used to identify microsatellites in these reads to find all the microsatellite variants that aligned to the microsatellites in the assembly. The resulting microsatellites in the reads that aligned to the assembly were compared to the ones in the assembly. If the number of reads for two different microsatellite length variants (alleles) were equal, the locus was identified as polymorphic. However, if all the alleles had the same length variant for a particular microsatellite, the microsatellite was identified as non-polymorphism. Alleles with more than two length variants were discarded from the study.

2.5.2 Polymorphism at the microsatellite loci in unrelated individuals of pied and collared flycatcher

The same procedure that had been used for identification of polymorphism in the genome assembly was used on the low coverage sequences of 9 collared individuals and 10 pied individuals. Since the coverage was only 5x, it was unlikely to find both the alleles at the heterozygous sites. In this case, the allele having the highest read count was selected. In cases where two alleles had the same read count, one of the alleles was randomly selected. The procedure was done for all the 19 low coverage individuals. The resulting data for the 9 collared individuals was then combined with the genome assembly data in order to have 10 individuals in each species. The mean length, total number of alleles, total number of individuals that had data for a particular microsatellite locus and expected heterozygosity of all the individuals in each species were then calculated.

2.5.3 Relationship between heterozygosity and mean length

The dataset was divided into categories based on the type of the microsatellites motif (di-, tri-, tetra-nucleotide). Mean of all the expected heterozygosities for a particular mean length of a microsatellite was calculated. The mean expected heterozygosity was plotted against the mean length for each category. This was used to infer the relationship between expected heterozygosity and mean length as well as how heterozygosity differs amongst the different types of microsatellites.

There was a need to have a model that best fits the heterozygosity curves. Since it was out of scope for this project to implement complex mathematical models, a logistic function was used to find the best fit for the heterozygosity plots through trial and error methodology. The following equation was used for this:

$$LOG(C_1\text{-}f(t))=C_3x + C_3C_4$$

where C_1 is the spread on the y-axis and was chosen randomly so that a linear plot can be obtained for the first half of the equation, f(t) were the data points, heterozygosity in this case. C_3 and C_4 are the slope of the linear regression line and the x-axis translation, respectively. x is the intercept of the linear regression line and the heterozygosity plot.

Different lines were obtained by changing the values of C_1, however, the C_1 coordinates of the linear line that seemed most appropriate were chosen for the next steps. The value of C_3x was the intercept of the linear regression line and the value of C_4 was obtained by dividing the intercept by C_3. R scripts were used to compute all these values and plot the final best fit curve for heterozygosity curves that were obtained for each data point against the mean length for each of the three types of microsatellites (di-, tri-, tetra-nucleotide microsatellites).

The equation of each line was used to compute the values for each locus. These values were subtracted from the expected heterozygosity to compute the deviation of each locus. These values of deviation were mapped on to the chromosomes using the same procedure that was used to compute the expected heterozygosity over the chromosomes.

2.5.4 Variability distribution per chromosome

After the variability estimates of microsatellites were obtained in terms of expected heterozygosity and deviation, they were mapped on to the chromosomes using 200kb windows. Two files that had been generated earlier by group members were used for this step. One of the files contained information about the scaffolds that map on to each chromosome, the length of the chromosome and the direction of the scaffold on the chromosome. The other file contained information about the division of all the scaffolds making up the assembly into 200kb windows.

Perl scripts were used to extract all the scaffolds that mapped onto the chromosomes, using the first reference file, together with the information about the start and stop position of a microsatellite, expected heterozygosity and deviation for that particular microsatellite locus.

Once this information was generated, the second reference file was used to divide the extracted scaffolds in to 200kb windows and to estimate the mean expected heterozygosity and mean deviation for all the microsatellites in a particular window. IntersectBed, a package of BEDtools, was used for this purpose. IntersectBed finds the overlap between two sets of genomic features. Two files were provided to intersectBed: one having the scaffold ID and the start and stop position of the microsatellites in the assembly, the other having scaffold ID and the start and stop position of the 200kB window. The output file contained information about the scaffold ID, the start and stop position of the microsatellites as well as the 200kB window. The scaffold ID and start and stop positions of the microsatellites were then used to extract the information about expected heterozygosity and deviation from the files generated earlier.

The final file that had to be used for mapping variability onto the chromosomes had information about the scaffold ID, start and stop position of the microsatellite and the 200 kb window, expected heterozygosity and the deviation for each locus. MySQL server was used to create a database from the final file generated from the last step. SQL queries were used to find the mean heterozygosity and deviation for all the loci that were included in a particular window.

R scripts, generated by other group members, were used to plot the mean heterozygosity and mean deviation for each 200kb window against the chromosomes for all the pied and collared individuals based on the direction of the scaffold on the chromosome deduced from the first reference file.

2.5.5 Autosomal variability vs. sex chromosome variability

A mean of expected heterozygosity was calculated for all the autosomes and was compared to that of the sex chromosome.

2.5.6 Density of microsatellite per chromosome

The total number of microsatellites in all the scaffolds mapping on to a particular chromosome was calculated. This was then divided by the total length of the chromosome to estimate the density of microsatellites per chromosome.

2.5.7 Degree of genetic differentiation between pied and collared flycatcher

F_{st} (fixation index) is used to measure the population differentiation, genetic distance, based on genetic polymorphism data. F_{st} is a special case of F-statistics. The microsatellite loci which were common between the individuals of each species were considered for the calculation of F_{st} through R scripts generated by one of the group members. The microsatellite loci and the alleles present in 10 individuals from each species were given as input together with a specific window size (200kb). The program used this data to calculate the F_{st} for each locus in a particular window. The resulting data for each scaffold was combined on the basis of the chromosomes and plotted using the R scripts that had been used earlier for plotting mean expected heterozygosity. Since the results obtained when all the loci were included had a lot of 0 values for the loci which had only one allele, there was a chance that these values brought the lines on the plot down. To avoid this, all the loci which had different alleles were considered for the computation of F_{st} (Figure 10).

Figure 2 shows a summary of the procedure used for the manipulation of microsatellite data throughout this project.

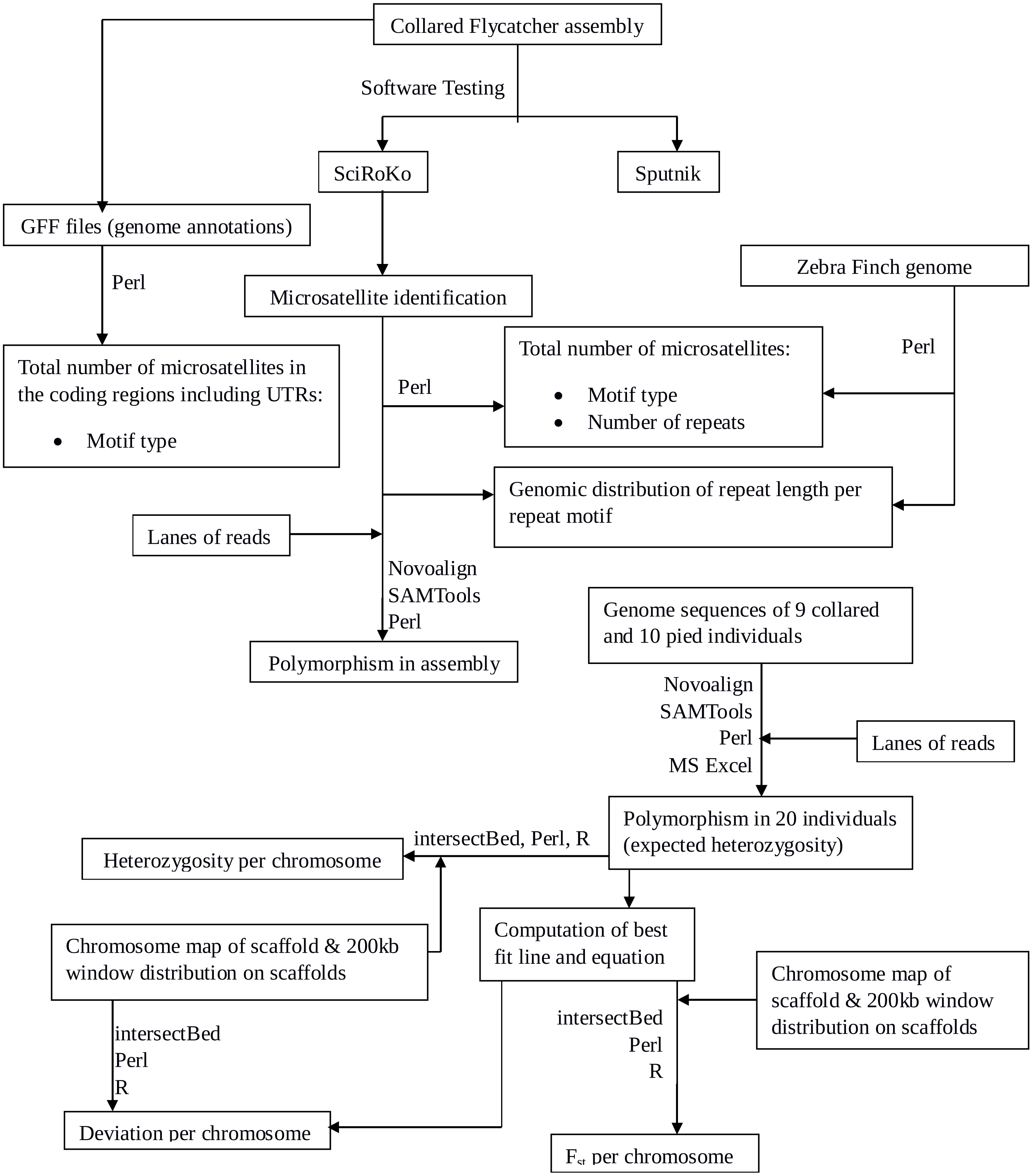
Collared Flycatcher assembly
Software Testing
SciRoKo
Sputnik
GFF files (genome annotations)
Perl
Microsatellite identification
Zebra Finch genome
Perl
Total number of microsatellites in the coding regions including UTRs:
Motif type
Perl
Total number of microsatellites:
Motif type
Number of repeats
Genomic distribution of repeat length per repeat motif
Lanes of reads
Novoalign
SAMTools
Perl
Polymorphism in assembly
Genome sequences of 9 collared and 10 pied individuals
Novoalign
SAMTools
Perl
MS Excel
Lanes of reads
Polymorphism in 20 individuals (expected heterozygosity)
intersectBed, Perl, R
Heterozygosity per chromosome
Chromosome map of scaffold & 200kb window distribution on scaffolds
Computation of best fit line and equation
Chromosome map of scaffold & 200kb window distribution on scaffolds
intersectBed
Perl
R
intersectBed
Perl
R
Deviation per chromosome
F_{st} per chromosome

Figure 2: Flowchart showing the procedure used throughout the project to manipulate the microsatellite data.

The main languages used were Perl and R.

3 Results and Discussion

Microsatellites have been used to study the parentage of pied flycatcher which is known to have extra-pair paternity in some populations (Leifjeld, et al., 1991, Gelter, et al., 1992, Ellegren, 1995, Craig, et al., 1996). The identification of microsatellites and characterization of polymorphism at the microsatellite loci in the flycatcher genome can be of great importance in detection of extra-pair paternity and identification of true biological parents. Studies of the evolutionary dynamics of microsatellites can be useful in understanding the pattern of molecular evolution (Buschiazzo, et al., 2010). Also, besides addressing questions relating to microsatellite evolution, comparative studies of microsatellites in recently diverged species can help to understand the pattern of species divergence. Microsatellites have extensively been used for the analysis of population structure, both for studies of sub-populations within a single species and to determine the evolutionary relationship between species.

This project focuses on the identification of microsatellites and polymorphism at these microsatellites loci in the pied and collared flycatchers using bioinformatics methods. The degree of polymorphism has been measured using expected heterozygosity calculations which are then used to find the distribution of heterozygosity and F_{st} for the pied and collared individuals per chromosome. The density of microsatellites over each chromosome and a comparison of the expected heterozygosity between the autosomes and the sex chromosome have also been estimated during this project.

3.1 Identification of microsatellites in the flycatcher genome

3.1.1 Total number of microsatellites in the flycatcher genome

There were more than 7.5 million microsatellites with 5 or more repeat units in the flycatcher genome. As the threshold for the minimum number of repeat units was changed from 5 to 8 and 10, a change in the total number of microsatellites was observed from more than 6.6 million to more than 2.7 million. Therefore, as the number of repeat units increased, the total number of microsatellites decreased with a sudden decrease of about 4 million microsatellites when the threshold for minimum number of repeat units was changed from 8 to 10. Table 1 shows a detailed distribution of the number and type of microsatellite, and microsatellite motif, found when the threshold for minimum number of repeat units was 5, 8 and 10. As mentioned earlier, it has been inferred from different studies, that among the dinucleotide microsatellite motifs, (GC) motifs are the least commonly occurring ones (Ellegren, 2004). This can also be observed from the results in Table 1 where the total number of $(GC)_n$ (microsatellites with a GC motif) is significantly less than the rest of the trinucleotide microsatellites. This can be due to the fact that the bonding between GC residues is very strong and the chances of a slippage or a mutation occurring in a GC rich region are very low (Ellegren, personal communication). Also, as GC rich regions are the coding regions, chances of a microsatellite occurring in a coding region are also less since polymorphism at these loci can greatly affect the protein regulation and expression (Ellegren; Uebbing, personal communication). The table also shows that (TG) motifs are the most common ones followed by (TA). These are mostly the non-coding/ intergenic regions where a mutation is less likely to effect the protein regulation and expression and there are higher chances of mutations. The bonds between (TA) motifs are also 2 which further allow mutations through slippage to occur more easily.

Table 1: Total number of microsatellites in the flycatcher genome.

Repeat Type	Microsatellite Motif	Repeat Unit >=5	Repeat Unit >=8	Repeat Unit >=10
Mono	(A)(T)	6286881	616954	258514
Mono	(G)(C)	1241099	32494	12014
Di	(AT)(TA)	24854	5082	2220
Di	(GC)(CG)	198	5	0
Di	(TC)(GA)(CT)(AG)	21506	1856	764
Di	(TG)(CA)(GT)(AC)	36604	4897	2580
Tri	(CAT)(ATG)(ATC)(GAT)(TCA)(TGA)	1238	235	93
Tri	(CAA)(TTG)(AAC)(GTT)(ACA=(TGT)	2598	270	68
Tri	(AAT)(ATT)(ATA)(TAT)(TAA)(TTA)	3388	474	153
Tri	(AAG)(CTT)(AGA)(TCT)(GAA)(TTC)	787	92	27
Tri	(GCT)(AGC)(CTG)(CAG)(TGC)(GCA)	2606	236	52
Tri	(TCC)(GGA)(CCT)(AGG)(CTC)(GAG)	3524	590	170
Tri	(GCG)(CGC)(CGG)(CCG)(GGC)(GCC)	468	15	0
Tri	(CTA)(TAG)(TAC)(GTA)(ACT)(AGT)	502	129	50
Tri	(CAC)(GTG)(ACC)(GGT)(CCA)(TGG)	528	22	3
Tri	(GTC)(GAC)(TCG)(CGA)(CGT)(ACG)	12	1	0
Tetra	(TTGT)(ACAA)(TGTT)(AACA)(GTTT)(AAAC)(TTTG)(CAAA)	2171	179	10
Tetra	(AATG)(CATT)(ATGA)(TCAT)(TGAA)(TTCA)(GAAT)(ATTC)	104	10	1
Tetra	(AATA)(TATT)(ATAA)(TTAT)(TAAA)(TTTA)(AAAT)(ATTT)	1133	102	9
Tetra	(AAGA)(TCTT)(AGAA)(TTCT)(GAAA)(TTTC)(AAAG)(CTTT)	449	13	1
Tetra	(TCAC)(GTGA)(CACT)(AGTG)(ACTC)(GAGT)(CTCA)(TGAG)	36	8	0
Tetra	(GAAG)(CTTC)(AAGG)(CCTT)(AGGA)(TCCT)(GGAA)(TTCC)	918	118	7
Tetra	(TAAT)(ATTA)(AATT)(TTAA)	141	10	0
Tetra	(GAGG)(CCTC)(AGGG)(CCCT)(GGGA)(TCCC)(GGAG)(CTCC)	475	11	0
Tetra	(AACC)(GGTT)(ACCA)(TGGT)(CCAA)(TTGG)(CAAC)(GTTG)	380	49	5
Tetra	(AGAC)(GTCT)(GACA)(TGTC)(ACAG)(CTGT)(CAGA)(TCTG)	463	36	2
Tetra	(CTAA)(TTAG)(TAAC)(GTTA)(AACT)(AGTT)(ACTA)(TAGT)	39	3	0
Tetra	(GATA)(TATC)(ATAG)(CTAT)(TAGA)(TCTA)(AGAT)(ATCT)	368	55	10
Tetra	(CATC)(GATC)(ATCC)(GGAT)(TCCA)(TGGA)(CCAT)(ATGG)	4955	278	12
Tetra	(TACA)(TGTA)(ACAT)(ATGT)(CATA)(TATG)(ATAC)(GTAT)	204	26	1
Tetra	(AGCA)(TGCT)(GCAA)(TTGC)(CAAG)(CTTG)(AAGC)(GCTT)	32	4	0
Tetra	(CTGG)(CCAG)(TGGC)(GCCA)(GGCT)(AGCC)(GCTG)(CAGC)	13	3	0
Tetra	(TGAT)(ATCA)(GATT)(AATC)(ATTG)(CAAT)(TTGA)(TCAA)	175	53	2
Tetra	(TGAC)(GTCA)(GACT)(AGTC)(ACTG)(CAGT)(CTGA)(TCAG)	36	1	0
Tetra	(CTGC)(GCAG)(TGCC)(GGCA)(GCCT)(AGGC)(CCTG)(CAGG)	48	5	0
Tetra	(TGCG)(CGCA)(GCGT)(ACGC)(CGTG)(CACG)(GTGC)(GCAC)	4	1	0
Tetra	(GTCC)(GGAC)(TCCG)(CGGA)(CCGT)(ACGG)(CGTC)(GACG)	94	9	1
Tetra	(AGGT)(ACCT)(GGTA)(TACC)(GTAG)(CTAC)(TAGG)(CCTA)	27	4	0
Tetra	(GGGT)(ACCC)(GGTG)(CACC)(GTGG)(CCAC)(TGGG)(CCCA)	10	0	0
Tetra	(TAAG)(CTTA)(AAGT)(ACTT)(AGTA)(TACT)(GTAA)(TTAC)	32	4	0
Tetra	(ATGC)(GCAT)(TGCA)(GCTA)(TAGC)(CATG)	10	1	0
Tetra	(GCTC)(CAGC)(CTCG)(CGAG)(TCGC)(GCGA)(CGCT)(AGCG)	2	0	0
Tetra	(GACC)(GGTC)(ACCG)(CGGT)(CCGA)(TCGG)(CGAC)(GTCG)	2	1	0
Tetra	(GCGG)(CCGC)(CGGG)(CCCG)(GGGC)(GCCC)(GGCG)(CGCC)	3	0	0
Tetra	(GAGC)(CGTC)(AGCG)(CGCT)(GCGA)(TCGC)(CGAG)(CTCG)	1	0	0
	Total Number	**7639118**	**664336**	**276769**

The first column shows the type of the microsatellite, the second column shows the microsatellite motif and the third, fourth and fifth columns show how the total number of microsatellites varies with a change in the threshold of minimum number of repeat units from 5 to 8 and 10.

3.1.2 Genomic distribution of repeat length per repeat motif in the flycatcher genome

Analysis of the commonly occurring repeat lengths of a particular repeat motif in the assembly showed that as the type of a microsatellite motif varies from mononucleotide to dinucleotide, trinucleotide and tetranucleotide, it is less common to find longer microsatellites. This means that, for example in case of trinucleotide microsatellite motifs, microsatellites that are longer than 40 repeat units are found 800 times less than the trinucleotide microsatellite motif with a length of 5 repeat units (Table 2).

The results in Table 2 show that there are no tri- and tetranucleotide microsatellites longer than 10 repeat units. Whereas, in case of dinucleotide motifs, no microsatellites longer than 20 repeat units were detected using SciRoKo. The last column in the table shows that there are no microsatellites longer than 45 repeat units.

The results in Table 2 therefore confirm that microsatellites with a larger motif are less likely to reach greater length as compared to microsatellites with a smaller motif.

There is a decrease in the number of microsatellites with an increase in the threshold for the minimum number of repeats. This shows that longer microsatellites are less frequent than the smaller microsatellites. Also, in case of a change of the threshold from 8 to 10, a significant decrease is observed in the total number of tetranucleotide microsatellites, which means that the chances of finding tetranucleotide microsatellites longer than or equal to 10 repeat units are very little as compared to finding tetranucleotide motifs with a length greater than or equal to 8 repeat units.

In case of tetranucleotide microsatellites, a significant increase in the total number is observed at a threshold of 5 in case of (CATC) motif, and a significant decrease in GC rich tetranucleotide motifs such as (GACC) and (GCGG).

The evolution of microsatellites is a dynamic process. The repeat might shrink or expand over evolutionary timescales. Replication slippage might remove the microsatellite interruptions which might result in a transition instead of decay of microsatellite during the evolution of microsatellites. This factor emphasizes that point mutations normally destroys perfect repeats (Ellegren, 2004). Different studies suggest that longer alleles have a mutation bias towards a reduction in the number of repeat units (Primmer et al., 1998; Xu et al., 2000; Harr et al., 2000). Each point mutation in a microsatellite reduces the number of uninterrupted repeats; thus a higher base substitution rate (relative to the slippage rate) leads to shorter microsatellites (Harr et al., 2000)

Since there is a high number of mononucleotide microsatellites these were excluded from the later steps that involve polymorphism studies in order to have a smaller and more precise dataset.

Table 2: Genomic distribution of repeat length per repeat motif in the flycatcher genome.

Repeat Type	Motif	Length=5	Length=10	Length=15	Length=20	Length>20	Length>=25	Length>=30	Length>=35	Length>=40	Length>=45
Mono	(A)(T)	3891629	79083	14038	2918	18572	10458	4361	1906	266	0
Mono	(G)(C)	919582	2915	350	221	4175	3330	2296	1236	216	0
Di	(AT)(TA)	11833	642	99	8	0	0	0	0	0	0
Di	(GC)(CG)	140	0	0	0	0	0	0	0	0	0
Di	(TC)(GA)(CT)(AG)	14625	225	31	3	0	0	0	0	0	0
Di	(TG)(CA)(GT)(AC)	22073	593	173	27	0	0	0	0	0	0
Tri	(CAT)(ATG)(ATC)(GAT)(TCA)(TGA)	594	29	0	0	0	0	0	0	0	0
Tri	(CAA)(TTG)(AAC)(GTT)(ACA)(TGT)	1483	29	0	0	0	0	0	0	0	0
Tri	(AAT)(ATT)(ATA)(TAT)(TAA)(TTA)	1849	55	0	0	0	0	0	0	0	0
Tri	(AAG)(CTT)(AGA)(TCT)(GAA)(TTC)	486	11	0	0	0	0	0	0	0	0
Tri	(GCT)(AGC)(CTG)(CAG)(TGC)(GCA)	1640	21	0	0	0	0	0	0	0	0
Tri	(TCC)(GGA)(CCT)(AGG)(CTC)(GAG)	1763	85	0	0	0	0	0	0	0	0
Tri	(GCG)(CGC)(CGG)(CCG)(GGC)(GCC)	321	0	0	0	0	0	0	0	0	0
Tri	(CTA)(TAG)(TAC)(GTA)(ACT)(AGT)	205	16	0	0	0	0	0	0	0	0
Tri	(CAC)(GTG)(ACC)(GGT)(CCA)(TGG)	364	1	0	0	0	0	0	0	0	0
Tri	(GTC)(GAC)(TCG)(CGA)(CGT)(ACG)	8	0	0	0	0	0	0	0	0	0
Tetra	(TTGT)(ACAA)(TGTT)(AACA)(GTTT)(AAAC)(TTTG)(CAAA)	1255	10	0	0	0	0	0	0	0	0
Tetra	(AATG)(CATT)(ATGA)(TCAT)(TGAA)(TTCA)(GAAT)(ATTC)	56	1	0	0	0	0	0	0	0	0
Tetra	(AATA)(TATT)(ATAA)(TTAT)(TAAA)(TTTA)(AAAT)(ATTT)	590	9	0	0	0	0	0	0	0	0
Tetra	(AAGA)(TCTT)(AGAA)(TTCT)(GAAA)(TTTC)(AAAG)(CTTT)	266	1	0	0	0	0	0	0	0	0
Tetra	(TCAC)(GTGA)(CACT)(AGTG)(ACTC)(GAGT)(CTCA)(TGAG)	12	0	0	0	0	0	0	0	0	0
Tetra	(GAAG)(CTTC)(AAGG)(CCTT)(AGGA)(TCCT)(GGAA)(TTCC)	270	7	0	0	0	0	0	0	0	0
Tetra	(TAAT)(ATTA)(AATT)(TTAA)	83	0	0	0	0	0	0	0	0	0
Tetra	(GAGG)(CCTC)(AGGG)(CCCT)(GGGA)(TCCC)(GGAG)(CTCC)	276	0	0	0	0	0	0	0	0	0
Tetra	(AACC)(GGTT)(ACCA)(TGGT)(CCAA)(TTGG)(CAAC)(GTTG)	200	5	0	0	0	0	0	0	0	0
Tetra	(AGAC)(GTCT)(GACA)(TGTC)(ACAG)(CTGT)(CAGA)(TCTG)	240	2	0	0	0	0	0	0	0	0
Tetra	(CTAA)(TTAG)(TAAC)(GTTA)(AACT)(AGTT)(ACTA)(TAGT)	21	0	0	0	0	0	0	0	0	0
Tetra	(GATA)(TATC)(ATAG)(CTAT)(TAGA)(TCTA)(AGAT)(ATCT)	99	10	0	0	0	0	0	0	0	0
Tetra	(CATC)(GATC)(ATCC)(GGAT)(TCCA)(TGGA)(CCAT)(ATGG)	4327	12	0	0	0	0	0	0	0	0
Tetra	(TACA)(TGTA)(ACAT)(ATGT)(CATA)(TATG)(ATAC)(GTAT)	95	1	0	0	0	0	0	0	0	0
Tetra	(AGCA)(TGCT)(GCAA)(TTGC)(CAAG)(CTTG)(AAGC)(GCTT)	19	0	0	0	0	0	0	0	0	0
Tetra	(CTGG)(CCAG)(TGGC)(GCCA)(GGCT)(AGCC)(GCTG)(CAGC)	9	0	0	0	0	0	0	0	0	0
Tetra	(TGAT)(ATCA)(GATT)(AATC)(ATTG)(CAAT)(TTGA)(TCAA)	50	2	0	0	0	0	0	0	0	0
Tetra	(TGAC)(GTCA)(GACT)(AGTC)(ACTG)(CAGT)(CTGA)(TCAG)	21	0	0	0	0	0	0	0	0	0
Tetra	(CTGC)(GCAG)(TGCC)(GGCA)(GCCT)(AGGC)(CCTG)(CAGG)	29	0	0	0	0	0	0	0	0	0
Tetra	(TGCG)(CGCA)(GCGT)(ACGC)(CGTG)(CACG)(GTGC)(GCAC)	1	0	0	0	0	0	0	0	0	0
Tetra	(GTCC)(GGAC)(TCCG)(CGGA)(CCGT)(ACGG)(CGTC)(GACG)	44	1	0	0	0	0	0	0	0	0
Tetra	(AGGT)(ACCT)(GGTA)(TACC)(GTAG)(CTAC)(TAGG)(CCTA)	12	0	0	0	0	0	0	0	0	0
Tetra	(GGGT)(ACCC)(GGTG)(CACC)(GTGG)(CCAC)(TGGG)(CCCA)	8	0	0	0	0	0	0	0	0	0
Tetra	(TAAG)(CTTA)(AAGT)(ACTT)(AGTA)(TACT)(GTAA)(TTAC)	17	0	0	0	0	0	0	0	0	0
Tetra	(ATGC)(GCAT)(TGCA)(GCTA)(TAGC)(CATG)	7	0	0	0	0	0	0	0	0	0
Tetra	(GCTC)(CAGC)(CTCG)(CGAG)(TCGC)(GCGA)(CGCT)(AGCG)	1	0	0	0	0	0	0	0	0	0
Tetra	(GACC)(GGTC)(ACCG)(CGGT)(CCGA)(TCGG)(CGAC)(GTCG)	0	0	0	0	0	0	0	0	0	0
Tetra	(GCGG)(CCGC)(CGGG)(CCCG)(GGGC)(GCCC)(GGCG)(CGCC)	2	0	0	0	0	0	0	0	0	0
Tetra	(GAGC)(CGTC)(AGCG)(CGCT)(GCGA)(TCGC)(CGAG)(CTCG)	1	0	0	0	0	0	0	0	0	0

The first two columns show the type of microsatellite and microsatellite motif, respectively. The rest of the columns show the number of microsatellites of a particular length, given in terms of repeat units (5, 10, 15, 20, >=20, >=25, >=30, >=35, >=40 and >=45)

3.1.3 Microsatellites within coding sequences including untranslated regions (UTRs)

Table 3 shows that there are approximately 65,000 microsatellites within coding regions including UTRs. The number of mono-nucleotide microsatellites is 28,000 as compared to 6.2 million in the whole genome. As can be seen from the table, the number of tetranucleotide microsatellite is somewhat low showing that tetranucleotide microsatellites motifs are not found within the coding regions and UTRs.

Table 3: Total number of microsatellites within coding regions including UTRs.

Repeat Type	Motif	Repeat units >=5
Mono	(A)(T)	28625
Mono	(G)(C)	35247
Di	(AT)(TA)	5
Di	(GC)(CG)	10
Di	(TC)(GA)(CT)(AG)	268
Di	(TG)(CA)(GT)(AC)	129
Tri	(CAT)(ATG)(ATC)(GAT)(TCA)(TGA)	62
Tri	(CAA)(TTG)(AAC)(GTT)(ACA)(TGT)	8
Tri	(AAT)(ATT)(ATA)(TAT)(TAA)(TTA)	12
Tri	(AAG)(CTT)(AGA)(TCT)(GAA)(TTC)	40
Tri	(GCT)(AGC)(CTG)(CAG)(TGC)(GCA)	197
Tri	(TCC)(GGA)(CCT)(AGG)(CTC)(GAG)	292
Tri	(GCG)(CGC)(CGG)(CCG)(GGC)(GCC)	55
Tri	(CTA)(TAG)(TAC)(GTA)(ACT)(AGT)	0
Tri	(CAC)(GTG)(ACC)(GGT)(CCA)(TGG)	51
Tri	(GTC)(GAC)(TCG)(CGA)(CGT)(ACG)	2
Tetra	(TTGT)(ACAA)(TGTT)(AACA)(GTTT)(AAAC)(TTTG)(CAAA)	0
Tetra	(AATG)(CATT)(ATGA)(TCAT)(TGAA)(TTCA)(GAAT)(ATTC)	0
Tetra	(AATA)(TATT)(ATAA)(TTAT)(TAAA)(TTTA)(AAAT)(ATTT)	0
Tetra	(AAGA)(TCTT)(AGAA)(TTCT)(GAAA)(TTTC)(AAAG)(CTTT)	3
Tetra	(TCAC)(GTGA)(CACT)(AGTG)(ACTC)(GAGT)(CTCA)(TGAG)	0
Tetra	(GAAG)(CTTC)(AAGG)(CCTT)(AGGA)(TCCT)(GGAA)(TTCC)	4
Tetra	(TAAT)(ATTA)(AATT)(TTAA)	0
Tetra	(GAGG)(CCTC)(AGGG)(CCCT)(GGGA)(TCCC)(GGAG)(CTCC)	1
Tetra	(AACC)(GGTT)(ACCA)(TGGT)(CCAA)(TTGG)(CAAC)(GTTG)	0
Tetra	(AGAC)(GTCT)(GACA)(TGTC)(ACAG)(CTGT)(CAGA)(TCTG)	0
Tetra	(CTAA)(TTAG)(TAAC)(GTTA)(AACT)(AGTT)(ACTA)(TAGT)	0
Tetra	(GATA)(TATC)(ATAG)(CTAT)(TAGA)(TCTA)(AGAT)(ATCT)	0
Tetra	(CATC)(GATC)(ATCC)(GGAT)(TCCA)(TGGA)(CCAT)(ATGG)	15
Tetra	(TACA)(TGTA)(ACAT)(ATGT)(CATA)(TATG)(ATAC)(GTAT)	0
Tetra	(AGCA)(TGCT)(GCAA)(TTGC)(CAAG)(CTTG)(AAGC)(GCTT)	0
Tetra	(CTGG)(CCAG)(TGGC)(GCCA)(GGCT)(AGCC)(GCTG)(CAGC)	0
Tetra	(TGAT)(ATCA)(GATT)(AATC)(ATTG)(CAAT)(TTGA)(TCAA)	0
Tetra	(TGAC)(GTCA)(GACT)(AGTC)(ACTG)(CAGT)(CTGA)(TCAG)	0
Tetra	(CTGC)(GCAG)(TGCC)(GGCA)(GCCT)(AGGC)(CCTG)(CAGG)	0
Tetra	(TGCG)(CGCA)(GCGT)(ACGC)(CGTG)(CACG)(GTGC)(GCAC)	1
Tetra	(GTCC)(GGAC)(TCCG)(CGGA)(CCGT)(ACGG)(CGTC)(GACG)	2
Tetra	(AGGT)(ACCT)(GGTA)(TACC)(GTAG)(CTAC)(TAGG)(CCTA)	0
Tetra	(GGGT)(ACCC)(GGTG)(CACC)(GTGG)(CCAC)(TGGG)(CCCA)	0
Tetra	(TAAG)(CTTA)(AAGT)(ACTT)(AGTA)(TACT)(GTAA)(TTAC)	0
Tetra	(ATGC)(GCAT)(TGCA)(GCTA)(TAGC)(CATG)	0
Tetra	(GCTC)(GAGC)(CTCG)(CGAG)(TCGC)(GCGA)(CGCT)(AGCG)	0
Tetra	(GACC)(GGTC)(ACCG)(CGGT)(CCGA)(TCGG)(CGAC)(GTCG)	0
Tetra	(GCGG)(CCGC)(CGGG)(CCCG)(GGGC)(GCCC)(GGCG)(CGCC)	0
Tetra	(GAGC)(CGTC)(AGCG)(CGCT)(GCGA)(TCGC)(CGAG)(CTCG)	0
	Total Number	**65029**

The first two columns show the type of microsatellite and microsatellite motifs. The third column shows the total number of microsatellites longer than or equal to 5 repeat units.

In some studies, a high mutation rate in the microsatellites is thought to create a variation that does not have a direct deleterious effect on the individual but contributes to adaptation potential of a population (Kashi, et al., 1997; King, et al., 1997, Brohede, 2003). Since coding regions and the UTRs are important in protein expression and regulation, an occurrence of microsatellites in these regions can be, in some cases, harmful, as a polymorphism in a microsatellite can result in the change in expression of a protein or even change in protein regulation. Especially, an expansion in trinucleotide repeats can alter protein expression and result in neurodegenerative diseases with dementia or mental retardation (Margolis, et al. 1999), for example, Huntington's disease (Kremer, et al. 1991). Therefore, the occurrence of microsatellites within coding regions including UTRs is expected to be a lot less as compared to the microsatellites in the rest of the genome. Table 3 shows the number of microsatellites within the coding regions when the thresholds for minimum number of repeats was set to 5.

3.2 Comparison between the flycatcher and zebra finch genome based on microsatellite data

3.2.1 Total number of microsatellites in the zebra finch genome

In case of Zebra Finch genome, the total number of microsatellite was a lot larger than the flycatcher genome.

Table 4 shows that there are more than 9 million microsatellites at a threshold of 5. This total number of microsatellites significantly decreases to a little more than 0.8 million when the threshold is changed to 8. In case of Zebra Finch, the number of microsatellites with motif (CATC) is also significantly larger than the rest of the tetranucleotide microsatellites. $(TTGT)_n$ is the most common amongst the tetranucleotide microsatellites. The trend for $(GC)_n$ is the same in case of Zebra Finch as was in case of flycatcher, being the least common ones among the dinucleotide microsatellites.

However, the fact that the number of microsatellites in the zebra finch genome is larger than the number of microsatellites in the flycatcher genome might not be due to a biological reason but due to the fact that it is harder to assemble larger repeat with short NGS reads (Ellegren, personal communication) since repeats have always presented technical challenges for sequence alignment and assembly programs. From a computational perspective, repeats create ambiguities in alignment and assembly, which, in turn, can produce biases and errors when interpreting results (Treangen, et al. 2012). NGS technologies typically generate short reads with higher error rates. This means that assemblies that have longer repeats and duplications suffer from this short read length (Alkan, et al., 2010). In addition, the chance of unique alignment or assembly is reduced not only by the presence of repeat sequences in complex genomes, but also by shared homologies within closely related gene families and pseudogenes (Voelkerding, et al., 2009).

Table 4: Total number of microsatellites in the Zebra Finch genome.

Repeat Type	Motif	Repeat unit >=5	Repeat Unit>=8	Repeat Unit >=10
Mono	(A)(T)	7598322	754648	334167
Mono	(G)(C)	1367337	30534	10352
Di	(AT)(TA)	32120	8007	4889
Di	(GC)(CG)	347	5	1
Di	(TC)(GA)(CT)(AG)	23888	1904	923
Di	(TG)(CA)(GT)(AC)	38884	5508	3176
Tri	(CAT)(ATG)(ATC)(GAT)(TCA)(TGA)	1214	264	146
Tri	(CAA)(TTG)(AAC)(GTT)(ACA)(TGT)	2502	314	147
Tri	(AAT)(ATT)(ATA)(TAT)(TAA)(TTA)	4575	1462	1159
Tri	(AAG)(CTT)(AGA)(TCT)(GAA)(TTC)	919	231	179
Tri	(GCT)(AGC)(CTG)(CAG)(TGC)(GCA)	2504	291	179
Tri	(TCC)(GGA)(CCT)(AGG)(CTC)(GAG)	3680	578	203
Tri	(GCG)(CGC)(CGG)(CCG)(GGC)(GCC)	871	37	11
Tri	(CTA)(TAG)(TAC)(GTA)(ACT)(AGT)	786	172	107
Tri	(CAC)(GTG)(ACC)(GGT)(CCA)(TGG)	544	25	6
Tri	(GTC)(GAC)(TCG)(CGA)(CGT)(ACG)	14	1	0
Tetra	(TTGT)(ACAA)(TGTT)(AACA)(GTTT)(AAAC)(TTTG)(CAAA)	2175	149	35
Tetra	(AATG)(CATT)(ATGA)(TCAT)(TGAA)(TTCA)(GAAT)(ATTC)	151	74	63
Tetra	(AATA)(TATT)(ATAA)(TTAT)(TAAA)(TTTA)(AAAT)(ATTT)	1121	245	125
Tetra	(AAGA)(TCTT)(AGAA)(TTCT)(GAAA)(TTTC)(AAAG)(CTTT)	1075	714	652
Tetra	(TCAC)(GTGA)(CACT)(AGTG)(ACTC)(GAGT)(CTCA)(TGAG)	66	34	31
Tetra	(GAAG)(CTTC)(AAGG)(CCTT)(AGGA)(TCCT)(GGAA)(TTCC)	981	579	464
Tetra	(TAAT)(ATTA)(AATT)(TTAA)	119	4	0
Tetra	(GAGG)(CCTC)(AGGG)(CCCT)(GGGA)(TCCC)(GGAG)(CTCC)	249	29	8
Tetra	(AACC)(GGTT)(ACCA)(TGGT)(CCAA)(TTGG)(CAAC)(GTTG)	243	52	29
Tetra	(AGAC)(GTCT)(GACA)(TGTC)(ACAG)(CTGT)(CAGA)(TCTG)	262	27	5
Tetra	(CTAA)(TTAG)(TAAC)(GTTA)(AACT)(AGTT)(ACTA)(TAGT)	30	8	6
Tetra	(GATA)(TATC)(ATAG)(CTAT)(TAGA)(TCTA)(AGAT)(ATCT)	878	691	633
Tetra	(CATC)(GATC)(ATCC)(GGAT)(TCCA)(TGGA)(CCAT)(ATGG)	1664	937	767
Tetra	(TACA)(TGTA)(ACAT)(ATGT)(CATA)(TATG)(ATAC)(GTAT)	265	48	16
Tetra	(AGCA)(TGCT)(GCAA)(TTGC)(CAAG)(CTTG)(AAGC)(GCTT)	51	6	0
Tetra	(CTGG)(CCAG)(TGGC)(GCCA)(GGCT)(AGCC)(GCTG)(CAGC)	8	1	1
Tetra	(TGAT)(ATCA)(GATT)(AATC)(ATTG)(CAAT)(TTGA)(TCAA)	125	22	3
Tetra	(TGAC)(GTCA)(GACT)(AGTC)(ACTG)(CAGT)(CTGA)(TCAG)	21	4	2
Tetra	(CTGC)(GCAG)(TGCC)(GGCA)(GCCT)(AGGC)(CCTG)(CAGG)	42	2	0
Tetra	(TGCG)(CGCA)(GCGT)(ACGC)(CGTG)(CACG)(GTGC)(GCAC)	3	2	1
Tetra	(GTCC)(GGAC)(TCCG)(CGGA)(CCGT)(ACGG)(CGTC)(GACG)	53	8	4
Tetra	(AGGT)(ACCT)(GGTA)(TACC)(GTAG)(CTAC)(TAGG)(CCTA)	21	7	0
Tetra	(GGGT)(ACCC)(GGTG)(CACC)(GTGG)(CCAC)(TGGG)(CCCA)	20	0	0
Tetra	(TAAG)(CTTA)(AAGT)(ACTT)(AGTA)(TACT)(GTAA)(TTAC)	16	2	0
Tetra	(ATGC)(GCAT)(TGCA)(GCTA)(TAGC)(CATG)	3	0	0
Tetra	(GCTC)(CAGC)(CTCG)(CGAG)(TCGC)(GCGA)(CGCT)(AGCG)	2	0	0
Tetra	(GACC)(GGTC)(ACCG)(CGGT)(CCGA)(TCGG)(CGAC)(GTCG)	1	0	0
Tetra	(GCGG)(CCGC)(CGGG)(CCCG)(GGGC)(GCCC)(GGCG)(CGCC)	6	0	0
Tetra	(GAGC)(CGTC)(AGCG)(CGCT)(GCGA)(TCGC)(CGAG)(CTCG)	0	0	0
Tetra	(ACGA)(TCGT)(CGAA)(TTCG)(GAAC)(GTTC)(AACG)(CGTT)	2	2	1
Tetra	(AGCT)(GCTA)(TAGC)(CTAG)	4	0	0
	Total Number	**9088158**	**807626**	**358490**

The first two columns show the type of microsatellite and microsatellite motif. The three left columns show the total number of microsatellites when the threshold for minimum number of repeat units was set to 5, 8 and 10 in SciRoKo.

3.2.2 Genomic distribution of repeat length per repeat motif in zebra finch genome

Table 5 shows the genomic distribution of repeat length per repeat motif in the Zebra Finch genome. The number of microsatellites that are longer than 40 repeat units is a lot more as compared to flycatcher, where no microsatellites were greater than 45 repeat units long. A large number of microsatellites are longer than 20 repeat units, especially noticeable in case of tri- and tetranucleotide microsatellites when compared to the tri- and tetranucleotide microsatellites in the flycatcher genome.

Table 5: Genomic distribution of repeat length per repeat motif in the Zebra Finch genome.

Repeat Type	Microsatellite Motif	Length = 5	Length = 10	Length = 15	Length = 20	Length > 20	Length >= 25	Length >= 30	Length >= 35	Length >= 40	Length >= 45	Length >= 50
Mono	(A)(T)	4707636	91812	20852	4598	26339	13704	5495	2792	738	155	39
Mono	(G)(C)	1024431	2941	634	66	87	12	5	2	2	1	1
Di	(AT)(TA)	14994	833	288	182	868	284	68	9	2	2	1
Di	(GC)(CG)	259	0	0	0	0	0	0	0	0	0	0
Di	(TC)(GA)(CT)(AG)	16485	232	48	25	88	37	19	8	3	3	1
Di	(TG)(CA)(GT)(AC)	23874	516	181	102	656	253	46	11	3	1	0
Tri	(CAT)(ATG)(ATC)(GAT)(TCA)(TGA)	566	26	10	5	13	3	2	0	0	0	0
Tri	(CAA)(TTG)(AAC)(GTT)(ACA)(TGT)	1442	35	7	4	14	5	0	0	0	0	0
Tri	(AAT)(ATT)(ATA)(TAT)(TAA)(TTA)	2022	71	104	56	280	87	17	2	0	0	0
Tri	(AAG)(CTT)(AGA)(TCT)(GAA)(TTC)	479	17	7	3	88	59	38	24	19	10	7
Tri	(GCT)(AGC)(CTG)(CAG)(TGC)(GCA)	1573	33	11	3	24	6	2	0	0	0	0
Tri	(TCC)(GGA)(CCT)(AGG)(CTC)(GAG)	1948	76	5	0	3	0	0	0	0	0	0
Tri	(GCG)(CGC)(CGG)(CCG)(GGC)(GCC)	611	6	0	0	0	0	0	0	0	0	0
Tri	(CTA)(TAG)(TAC)(GTA)(ACT)(AGT)	247	14	8	2	3	0	0	0	0	0	0
Tri	(CAC)(GTG)(ACC)(GGT)(CCA)(TGG)	376	5	1	0	0	0	0	0	0	0	0
Tri	(GTC)(GAC)(TCG)(CGA)(CGT)(ACG)	11	0	0	0	0	0	0	0	0	0	0
Tetra	(TTGT)(ACAA)(TGTT)(AACA)(GTTT)(AAAC)(TTTG)(CAAA)	1433	24	0	0	0	0	0	0	0	0	0
Tetra	(AATG)(CATT)(ATGA)(TCAT)(TGAA)(TTCA)(GAAT)(ATTC)	52	10	8	0	0	0	0	0	0	0	0
Tetra	(AATA)(TATT)(ATAA)(TTAT)(TAAA)(TTTA)(AAAT)(ATTT)	558	46	5	0	0	0	0	0	0	0	0
Tetra	(AAGA)(TCTT)(AGAA)(TTCT)(GAAA)(TTTC)(AAAG)(CTTT)	222	30	28	19	414	313	235	168	128	88	64
Tetra	(TCAC)(GTGA)(CACT)(AGTG)(ACTC)(GAGT)(CTCA)(TGAG)	18	5	2	0	2	1	0	0	0	0	0
Tetra	(GAAG)(CTTC)(AAGG)(CCTT)(AGGA)(TCCT)(GGAA)(TTCC)	180	39	24	15	216	148	105	72	54	34	25
Tetra	(TAAT)(ATTA)(AATT)(TTAA)	84	0	0	0	0	0	0	0	0	0	0
Tetra	(GAGG)(CCTC)(AGGG)(CCCT)(GGGA)(TCCC)(GGAG)(CTCC)	149	4	0	0	0	0	0	0	0	0	0
Tetra	(AACC)(GGTT)(ACCA)(TGGT)(CCAA)(TTGG)(CAAC)(GTTG)	129	4	1	0	0	0	0	0	0	0	0
Tetra	(AGAC)(GTCT)(GACA)(TGTC)(ACAG)(CTGT)(CAGA)(TCTG)	121	5	0	0	0	0	0	0	0	0	0
Tetra	(CTAA)(TTAG)(TAAC)(GTTA)(AACT)(AGTT)(ACTA)(TAGT)	12	2	1	0	1	0	0	0	0	0	0
Tetra	(GATA)(TATC)(ATAG)(CTAT)(TAGA)(TCTA)(AGAT)(ATCT)	92	55	77	5	18	5	1	0	0	0	0
Tetra	(CATC)(GATC)(ATCC)(GGAT)(TCCA)(TGGA)(CCAT)(ATGG)	345	87	77	13	63	24	9	4	2	2	2
Tetra	(TACA)(TGTA)(ACAT)(ATGT)(CATA)(TATG)(ATAC)(GTAT)	120	8	0	0	0	0	0	0	0	0	0
Tetra	(AGCA)(TGCT)(GCAA)(TTGC)(CAAG)(CTTG)(AAGC)(GCTT)	32	0	0	0	0	0	0	0	0	0	0
Tetra	(CTGG)(CCAG)(TGGC)(GCCA)(GGCT)(AGCC)(GCTG)(CAGC)	5	0	0	0	0	0	0	0	0	0	0
Tetra	(TGAT)(ATCA)(GATT)(AATC)(ATTG)(CAAT)(TTGA)(TCAA)	42	1	0	0	0	0	0	0	0	0	0
Tetra	(TGAC)(GTCA)(GACT)(AGTC)(ACTG)(CAGT)(CTGA)(TCAG)	8	2	0	0	0	0	0	0	0	0	0
Tetra	(CTGC)(GCAG)(TGCC)(GGCA)(GCCT)(AGGC)(CCTG)(CAGG)	27	0	0	0	0	0	0	0	0	0	0
Tetra	(TGCG)(CGCA)(GCGT)(ACGC)(CGTG)(CACG)(GTGC)(GCAC)	1	0	0	0	0	0	0	0	0	0	0
Tetra	(GTCC)(GGAC)(TCCG)(CGGA)(CCGT)(ACGG)(CGTC)(GACG)	19	2	0	0	0	0	0	0	0	0	0
Tetra	(AGGT)(ACCT)(GGTA)(TACC)(GTAG)(CTAC)(TAGG)(CCTA)	8	1	0	0	0	0	0	0	0	0	0
Tetra	(GGGT)(ACCC)(GGTG)(CACC)(GTGG)(CCAC)(TGGG)(CCCA)	15	0	0	0	0	0	0	0	0	0	0
Tetra	(TAAG)(CTTA)(AAGT)(ACTT)(AGTA)(TACT)(GTAA)(TTAC)	10	0	0	0	0	0	0	0	0	0	0
Tetra	(ATGC)(GCAT)(TGCA)(GCTA)(TAGC)(CATG)	2	0	0	0	0	0	0	0	0	0	0
Tetra	(GCTC)(CAGC)(CTCG)(CGAG)(TCGC)(GCGA)(CGCT)(AGCG)	2	0	0	0	0	0	0	0	0	0	0
Tetra	(GACC)(GGTC)(ACCG)(CGGT)(CCGA)(TCGG)(CGAC)(GTCG)	1	0	0	0	0	0	0	0	0	0	0
Tetra	(GCGG)(CCGC)(CGGG)(CCCG)(GGGC)(GCCC)(GGCG)(CGCC)	6	0	0	0	0	0	0	0	0	0	0
Tetra	(GAGC)(CGTC)(AGCG)(CGCT)(GCGA)(TCGC)(CGAG)(CTCG)	0	0	0	0	0	0	0	0	0	0	0
Tetra	(ACGA)(TCGT)(CGAA)(TTCG)(GAAC)(GTTC)(AACG)(CGTT)	0	1	0	0	0	0	0	0	0	0	0
Tetra	(AGCT)(GCTA)(TAGC)(CTAG)	2	0	0	0	0	0	0	0	0	0	0

The first two columns show the type of microsatellite and microsatellite motif, respectively. The rest of the columns show the number of microsatellites of a particular length, given in terms of repeat units (5, 10, 15, 20, >=20, >=25, >=30, >=35, >=40, >=45 and >=50).

3.2.3 Comparison between the flycatcher and zebra finch genome based on microsatellite data

Table 6, 7 and 8 show a summary and comparison of microsatellites in the flycatcher and zebra finch genome. As can be seen from Table 8, the mononucleotide microsatellites in the zebra finch genome, with a minimum length of 5 repeat units, is 1.19 times more than the total number of mononucleotide microsatellites in the flycatcher genome. On average, the number of microsatellites in the zebra finch genome is approximately 1.06 times more than the number of microsatellites in the flycatcher genome, at a minimum length of 5 repeat units. However, at a minimum length of 10 repeat units, in case of tetranucleotide microsatellites, zebra finch genome has 3.469 times (Table 8) more microsatellites than the flycatcher genome, where flycatcher has 61 tetranucleotide (Table 6) microsatellites and zebra finch has 2,846 tetranucleotide microsatellites (Table 7).

Table 6: Abundance of microsatellites in the flycatcher assembly (Ellegren, personal communication).

Repeat type	Minimum length (repeat units)	
	5	**10**
Mononucleotide	7,527,980	270,528
Dinucleotide	83,162	5,564
Trinucleotide	15,651	616
Tetranucleotide	12,325	61

Table 7: Abundance of microsatellites in the zebra finch genome (Ellegren, personal communication).

Repeat type	Minimum length (repeat units)	
	5	**10**
Mononucleotide	8,965,659	344,519
Dinucleotide	95,239	8,989
Trinucleotide	17,609	2,137
Tetranucleotide	9,657	2,846

Table 8: Relative occurrence of microsatellites in zebra finch genome compared to the flycatcher genome (Ellegren, personal communication).

Repeat type	Minimum length (repeat units)	
	5	**10**
Mononucleotide	1.19	1.27
Dinucleotide	1.145	1.616
Trinucleotide	1.125	3.469
Tetranucleotide	0.784	46.66

3.3 Identification of polymorphism at the microsatellite loci

The genome assembly was first used to identify polymorphism in step 4.3.1. Since the assembly had a high coverage, there was a chance of sampling both the alleles present in an individual for a particular microsatellite and to be able to estimate heterozygosity based on a difference in the length of the microsatellite in these alleles. With different reads aligning to alleles of different lengths, it becomes obvious if the individual is heterozygous or not based on the number of reads. A somewhat equal number of reads aligning to two different alleles suggest that two different alleles with two different microsatellites lengths occur within the individual. This, therefore, means that the individual is heterozygous for that microsatellite.

3.3.1 Polymorphism in the genome assembly

Table 9 shows the occurrence of polymorphism at the microsatellite loci in the flycatcher genome. In this case, only a few loci at Scaffold 1 are shown. Allele 1 is the microsatellite length in terms of repeat units found in the genome assembly whereas Count 1 shows the total number of reads having the same length of microsatellite as was found in the assembly. The rest of the alleles show the different lengths of microsatellites found in reads that had unique sequences at both ends of the microsatellites and that aligned to the microsatellite loci in the flycatcher genome.

A microsatellite with two different alleles that had the same read count was considered polymorphic (highlighted in yellow). However, if more than two alleles were found with an equal read count, the locus were not considered in this analysis. In total there are 1307 polymorphic loci in the genome assembly out of a total of approximately 0.1 million loci (since mononucleotide microsatellites were excluded from polymorphism calculations). The column labeled “Total number of reads covering the whole repeat region” shows the total number of reads aligned to the assembly at the microsatellite loci, having unique sequences at both ends.

This approach of identifying polymorphism at a microsatellite locus cannot be reliable since it is unclear what the difference in the total number of reads aligning to two different alleles should be considered to identify a locus polymorphic. In this case, for a locus to be considered polymorphic an exact equal number of reads mapping on to two different alleles is considered the right choice, which is likely to be wrong. This is because with relatively few reads, there are many more possibilities to get not exactly the same number of reads for the two alleles, compared to only a single possibility to get an equal number of reads. Since a difference of even a single read would misinterpret the polymorphic nature of a particular microsatellites locus.

3.3.2 Polymorphism at the microsatellite loci in unrelated individuals of pied and collared flycatcher

In order to find a more suitable approach to identify degree of polymorphism at particular microsatellite locus heterozygosity estimations were used by counting from one individual to 10 individuals of each species. Table 10 and Table 11 show the expected heterozygosity estimates for each microsatellite locus in 10 individuals from each species. The tables show the mean length of a particular microsatellite found within all the 10 individuals of a particular species. The allele frequencies are calculated based on the number of individuals that show a particular allele and the total number of individuals having an allele for a particular microsatellite. The expected heterozygosity is calculated based on the allele frequency data. The values for expected heterozygosity lie between 0 and 1. These values give an idea about the degree of polymorphism instead of rigidly specifying if a locus is polymorphic or not. The expected heterozygosity estimations can be useful when comparing the degree of polymorphism within species between different types of microsatellites (section 3.3.3) as well as between species.

For each table (Table 10 and 11) Column 1 and 2 show the scaffold and start and stop positions of a microsatellite (data for a few loci in Scaffold 1 are shown in this case). The third column shows the motif. The fourth column shows the number of repeat units in the assembly for a particular microsatellite. The fifth column shows the mean length of a microsatellite for all the individuals. Columns labeled 1-10 show the number of repeat units for each microsatellite for each of the 10 individuals. The “Allele Freq” columns show the calculations for the allele frequency for each allele found in the individuals by taking into account the number of individuals that have a particular allele and the total number of individuals that have an allele for a particular locus and the number of individuals. The last column shows the expected heterozygosity for a particular locus.

Table 9: Polymorphism content at microsatellite loci in the genome assembly.

Scaffold	Locus	Motif	Allele1	Count1	Allele 2	Count 2	Allele 3	Count3	Allele 4	Count4	Total number of reads covering the whole repeat region	Polymorphic?
S00001	18366716-18366801	AAT	29	2							2	No
S00001	2565442-2565492	TG	26	7							7	No
S00001	6618605-6618681	TAG	26	1							1	No
S00001	1318918-1318964	CT	24	1	22	1					2	Yes
S00001	20552996-20553040	AC	23	6	22	1	24	1			8	No
S00001	2433818-2433847	AT	15	1	16	1					2	Yes
S00001	3166576-3166636	AATG	15	16							16	No
S00001	3170178-3170206	TA	15	1							1	No
S00001	4869504-4869533	AT	15	1							1	No
S00001	5098456-5098485	TA	15	1	13	3	12	1	20	1	6	No
S00001	7002458-7002515	TTGG	15	5							5	No
S00001	10430612-10430641	AT	15	1							1	No
S00001	10860954-10860982	TA	15	1	14	2	11	1			4	No
S00001	11006506-11006534	GT	15	14	13	17	14	1	12	1	33	No
S00001	11442328-11442388	ATCC	15	2	16	2					4	Yes
S00001	12965163-12965206	TAA	15	0	24	1	16	1	13	1	3	No
S00001	6076498-6076524	GA	14	15	17	1	18	1	19	1	18	No
S00001	6107989-6108043	TGAA	14	10	18	1					11	No
S00001	10343136-10343189	TATT	14	3							3	No
S00001	10461118-10461145	TG	14	13	16	1	13	2	20	4	20	No
S00001	14741172-14741199	AT	14	1	13	1	15	1	12	1	4	No
S00001	15686176-15686203	AC	14	18	33	1	13	2	7	1	22	No
S00001	19488427-19488480	TTGA	14	36	13	1					37	No
S00001	24331352-24331379	TA	14	5	10	1	11	2	12	3	11	No
S00001	609057-609082	AT	13	0	15	1					1	No
S00001	933375-933424	TCCT	13	3							3	No
S00001	1021313-1021351	ACT	13	2	15	1	16	5	17	2	10	No
S00001	1503671-1503695	TG	13	3	11	17	14	19	17	1	40	No
S00001	1524961-1524986	CA	13	11	12	2	15	7	16	12	32	No
S00001	1777928-1777952	GA	13	2	9	5					7	No
S00001	2167042-2167066	GT	13	15	11	2	12	3			20	No
S00001	2367874-2367899	AT	13	2							2	No
S00001	3492729-3492753	AT	13	0	6	2					2	No
S00001	4809831-4809855	AT	13	4	12	4					8	Yes

The first column shows the scaffold ID (in this case only Scaffold 1 is shown). The second column shows the start and end position of a microsatellite separated by a “-“. The third column shows the microsatellite motifs. The columns labeled “Allele” show the different alleles found in the reads when the reads were aligned to the assembly. Allele 1 shows the microsatellite length found in the assembly. The columns labeled “Count” show the number of reads aligned to the assembly with unique sequences at both ends of a microsatellite. The third column shows the total number of reads covering a microsatellite locus. The last column, labeled “Polymorphism?” shows whether or not the individual is polymorphic for that particular microsatellite locus. The polymorphic loci are highlighted yellow in this case.

Table 10: Expected heterozygosity estimates for 10 collared individuals.

Scaffold	Locus	Motif	RU_ Original	Mean length	1	2	3	4	5	6	7	8	9	10	Allele freq 1	Allele freq 2	Allele freq 3	Allele freq 4	Allele freq 5	Allele freq 6	Allele freq 7	Allele freq 8	Allele freq 9	Heterozygosity
S00001	107627-107641	AAC	5	5	5		5	5		5	5	5	5	5	1									0
S00001	226263-226309	ATCC	12	11.11	13	10	7		16	11	9	14	9	12	0.11	0.11	0.11	0.11	0.11	0.22	0.11	0.11		0.8669
S00001	463627-463641	CAA	5	5	5			5	5	5	5	5	5	5	1									0
S00001	530404-530431	TAT	9	9.6			11		9	10		9		9	0.2	0.6	0.2							0.56
S00001	573543-573563	TTAT	5	5	5				5	5	5			5	1									0
S00001	586719-586740	TCC	7	6.8	7	5	7	7	7	7	7	7	7	7	0.9	0.1								0.18
S00001	609057-609082	AT	13	15										15	1									0
S00001	822090-822112	CAAC	6	6.17		6			6	6		7	6	6	0.83	0.17								0.2822
S00001	860588-860614	AT	14	14		13			14	15		14		14	0.2	0.6	0.2							0.56
S00001	887357-887387	TTTA	8	7.84	10		7	8		7		7		8	0.17	0.5	0.33							0.6122
S00001	933375-933424	TCCT	13	13										13	1									0
S00001	935113-935144	CAAA	8	9					8	9	7	10		11	0.2	0.2	0.2	0.2	0.2					0.8
S00001	1014916-1014937	GAT	7	7.4	8	7				8	7			7	0.4	0.6								0.48
S00001	1015014-1015035	GAT	7	8										8	1									0
S00001	1021313-1021351	ACT	13	14.58	13		13	17	17	17			13	12	0.43	0.43	0.14							0.6106
S00001	1099159-1099175	GAA	6	6.22	6	6		7	6	6	7	6	6	6	0.78	0.22								0.3432

Table 11: Expected heterozygosity estimates for 10 pied individuals.

Scaffold	Locus	Motif	RU Original	Mean length	1	2	3	4	5	6	7	8	9	10	Allele freq 1	Allele freq 2	Allele freq 3	Allele freq 4	Allele freq 5	Allele freq 6	Allele freq 7	Allele freq 8	Allele freq 9	Heterozygosity
S00001	107627-107641	AAC	5	5			5	5		5	5			5	1									0
S00001	226263-226309	ATCC	12	12.87	14	10	12		12	13	12		14	14	0.38	0.13	0.38	0.13						0.6774
S00001	463627-463641	CAA	5	5		5							5		1									0
S00001	530404-530431	TAT	9	8.5	8									9	0.5	0.5								0.5
S00001	573543-573563	TTAT	5	5								5			1									0
S00001	822090-822112	CAAC	6	6		6								6	1									0
S00001	887357-887387	TTTA	8	6	6										1									0
S00001	935113-935144	CAAA	8	7.5	5					10					0.5	0.5								0.5
S00001	1014916-1014937	GAT	7	7			7	7			7		7	7	1									0
S00001	1105884-1105898	CAA	5	7		7				7					1									0
S00001	1243759-1243778	TCCT	5	5.22	5	5	5	5	5		5	5	7	5	0.89	0.11								0.1958
S00001	1255906-1255928	ATTT	6	6	6			6						6	1									0
S00001	1362090-1362104	GTT	5	5		5	5			5	5	5			1									0
S00001	1412537-1412559	AG	12	12.8	15	12	12	14	12	15	12	12	12	12	0.2	0.7	0.1							0.46
S00001	1495245-1495272	CA	14	13.53		19						13		9	0.33	0.33	0.33							0.6733

3.3.3 Relationship between heterozygosity and mean length

Mean expected heterozygosity was plotted against mean repeat length to find the relationship between the mean length of a microsatellite and expected heterozygosity. It was observed that the relationship varies amongst the categories of microsatellites (di-, tri- and tetra-nucleotide microsatellites). Figure 3 shows the relationship between mean expected heterozygosity and mean length of all the different categories of microsatellites.

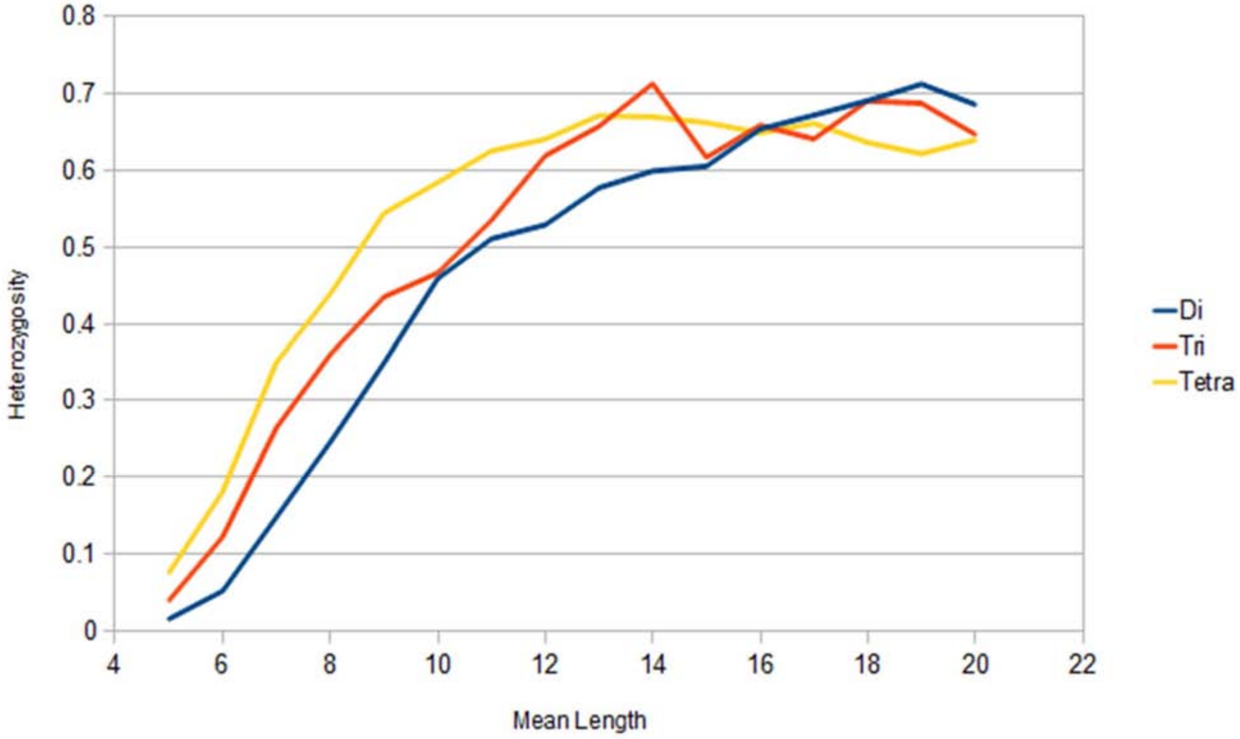

Figure 3: Mean expected heterozygosity against mean length in terms of number of repeat units for 10 collared individuals.

The blue line indicates the relationship between the mean length and expected heterozygosity for dinucleotide microsatellites, the red line shows the relationship for trinucleotide microsatellites and the yellow line shows the relationship for tetranucleotide microsatellites.

As can be seen from the plot, the relationship between mean expected heterozygosity and mean length tends to be sigmoidal. As opposed to the other microsatellites, tetranucleotide microsatellites show a somewhat different behavior where expected heterozygosity increases with an increase in number of repeat units and then becomes steady after a mean length of approximately 13 repeat units. Whereas in case of dinucleotide repeat units, the mean expected heterozygosity keeps on increasing with an increase in number of repeat units till approximately 20 repeat units after which the heterozygosity starts decreasing. In case of trinucleotide microsatellites, there are various peaks seen, however, a negative slope in case of trinucleotide repeats starts at about 14 repeat units.

Figure 4 and 5 show the best fit plots (red) for the heterozygosity curves (black) for mean length against mean expected heterozygosity for di-, tri- and tetra- nucleotide repeats, respectively.

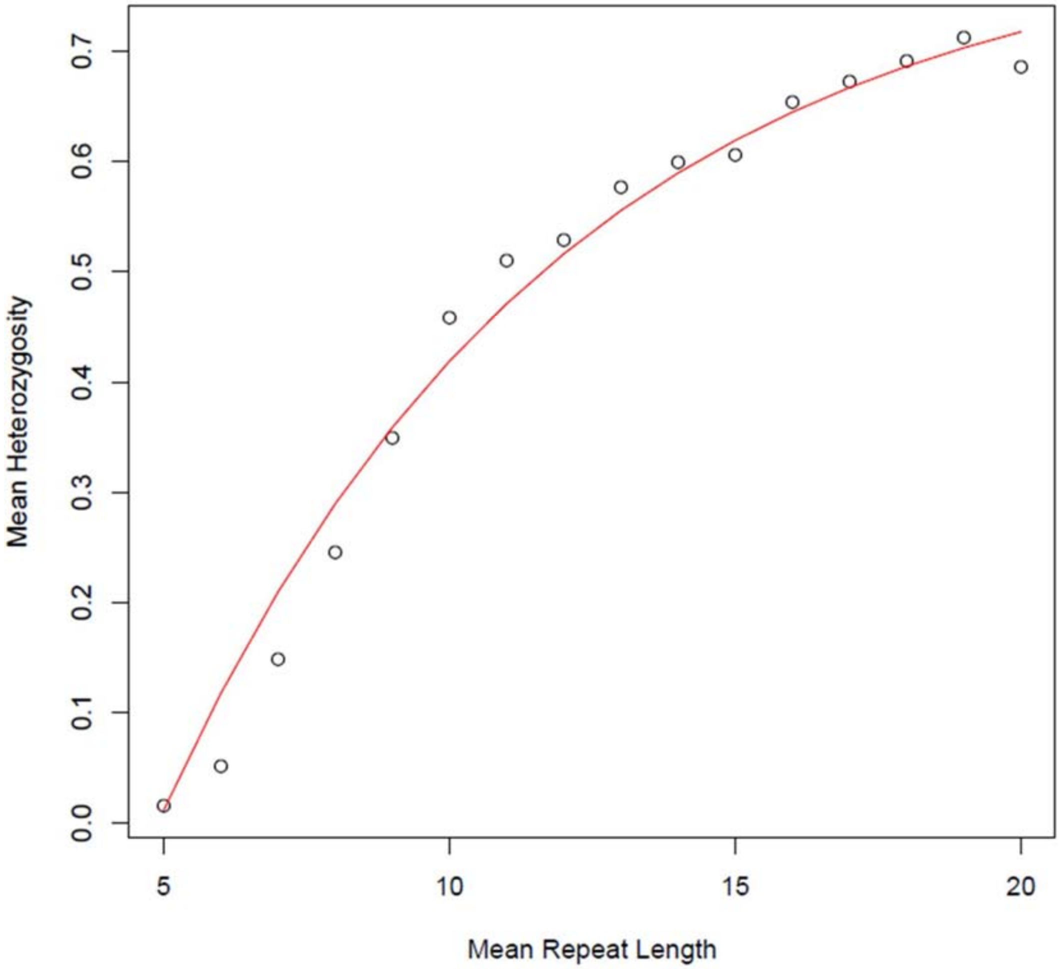

Figure 4: Best fit plots for the mean expected heterozygosity against mean length curves for dinucleotide microsatellites.

The red line shows the regression line that was obtained through a trial and error methodology. The black circles show the data points for dinucleotide motifs from Figure 3 above.

The regression line in Figure 4 was drawn with C_1 equal to 0.1 which was decided based on different trials and then selecting the line that seemed the best fit for the dinucleotide microsatellites.

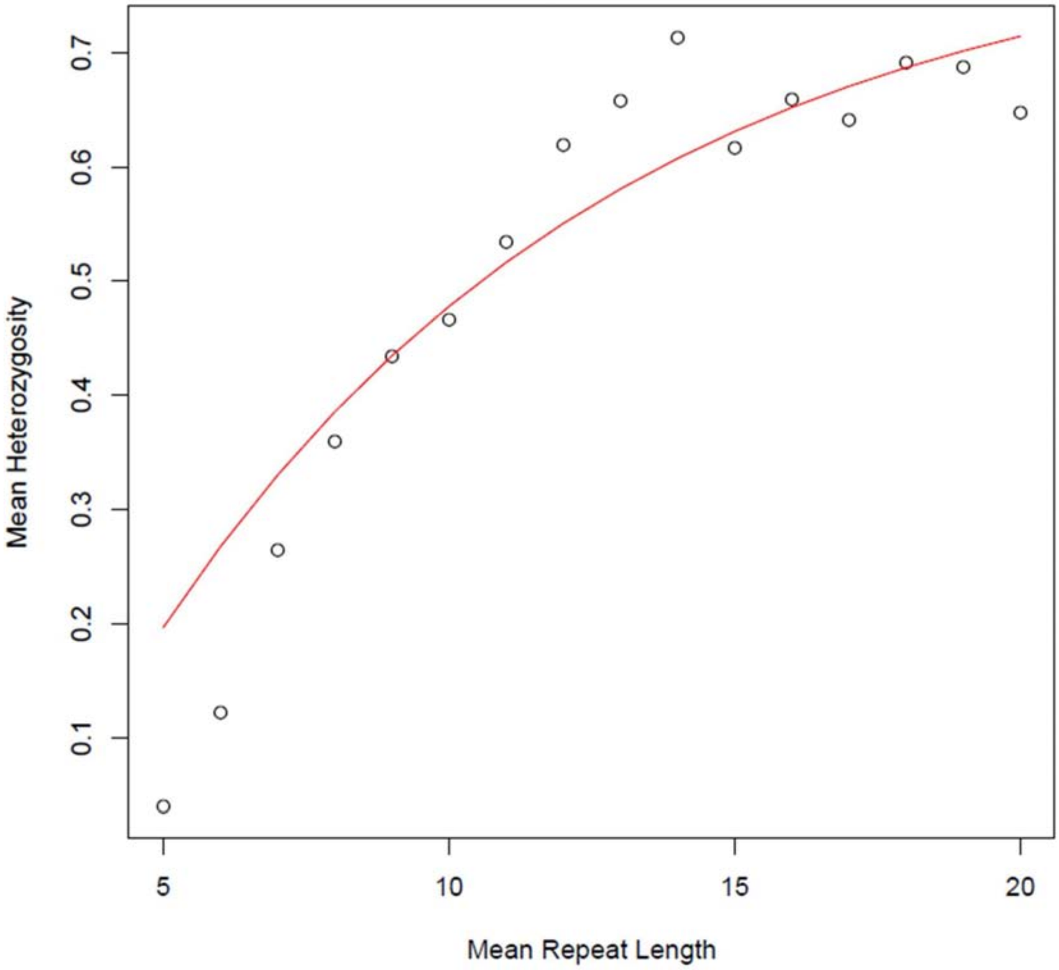

Figure 5: Best fit plots for the mean expected heterozygosity against mean length curves for trinucleotide microsatellites.

The red line shows the regression line that was obtained through a trial and error methodology. The black circles show the data points for trinucleotide motifs from Figure 3 above.

The regression line in Figure 5 was drawn with C_1 equal to 0.1 which was decided based on different trials and then selecting the line that seemed the best fit for the trinucleotide microsatellites.

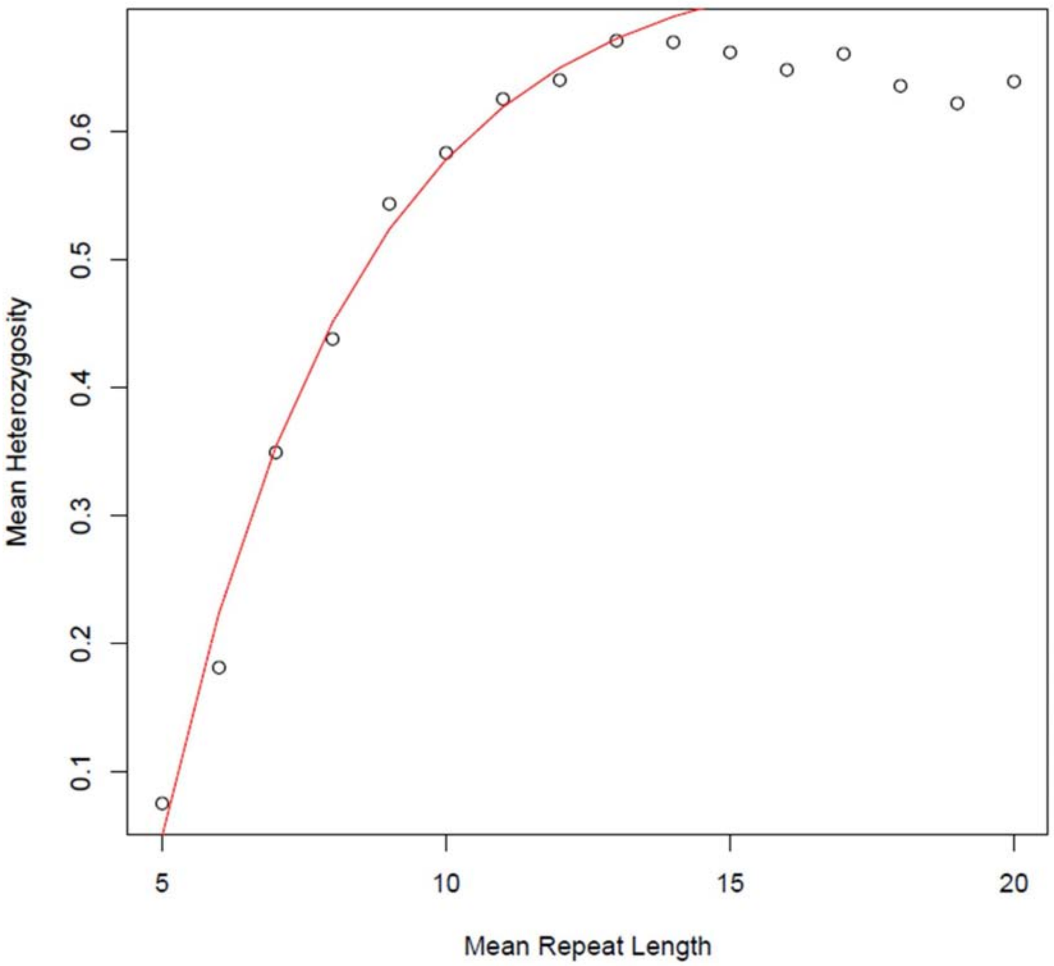

Figure 6: Best fit plots for the mean expected heterozygosity against mean length curves for tetranucleotide microsatellites.

The red line shows the regression line that was obtained through a trial and error methodology. The black circles show the data points for tetranucleotide motifs from Figure 3 above.

The regression line in Figure 6 was drawn with C_1 equal to 0.1 which was decided based on different trials and then selecting the line that seemed the best fit for the tetranucleotide microsatellites. However, in this case the data points for a mean length greater than 13 repeat units were not considered since the regression line that was obtained after considering these plots (Figure 7) was not a best fit for smaller mean length.

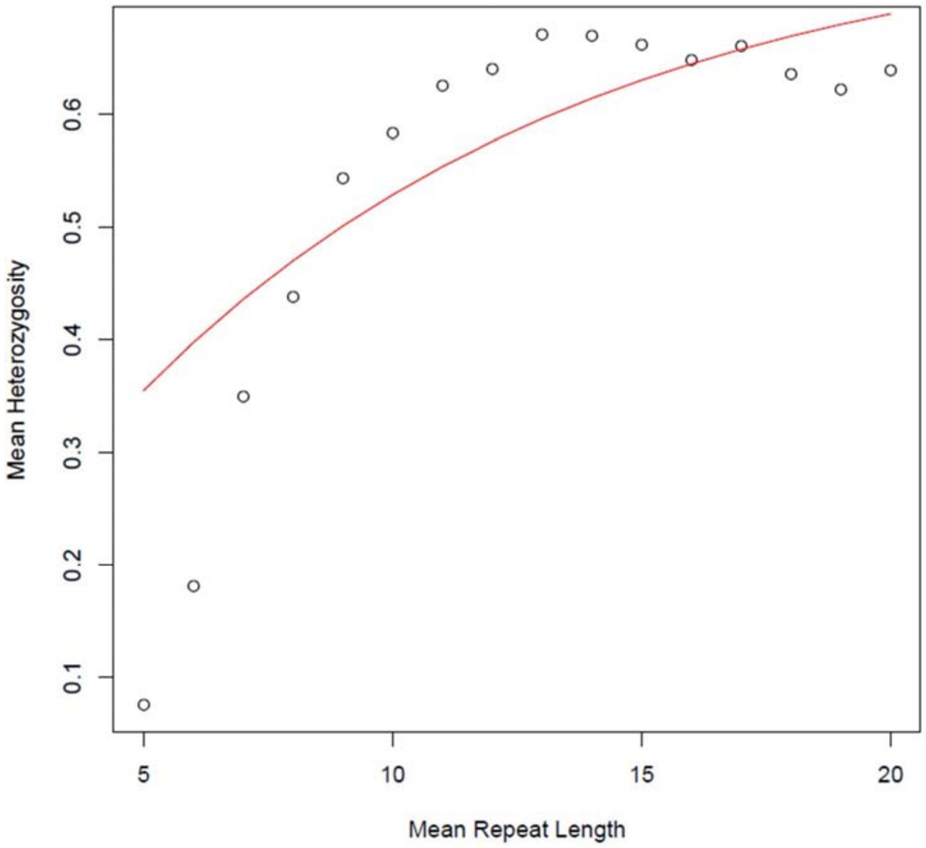

Figure 7: Best fit plots for the mean expected heterozygosity against mean length curves for tetranucleotide microsatellites when all the data points were considered.

The red line shows the regression line that was obtained through a trial and error methodology. The black circles show the data points for tetranucleotide motifs from Figure 3 above. The regression line is not a best fit for smaller mean repeat lengths.

The equations of these lines were used to compute the deviation for all the loci for each type of microsatellite. The values from the equation of the line were subtracted from the expected heterozygoisty values calculated earlier to compute the devation for each locus. These deviations were then plotted over the chromosomes using the same procedure that was used to map the expected heterozygosity over the chromosomes (Figure 8).

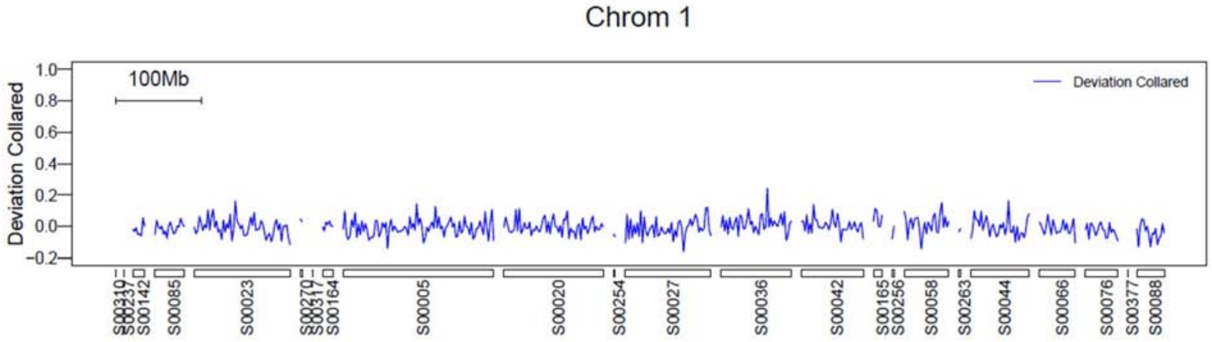

Figure 8: Deviation for dinucleotide repeats in collared flycatcher individuals for chromosome 1.

The x label shows the scaffolds that map onto chromosome 1.

Unfortunately, a clear pattern could not be observed in case of microsatellites deviations and heterozygosity, as were produced by others in the group for SNP (single nucleotide polymorphism) data. There is a lot of variation among windows, suggesting that there is a lot of noise in the data.

3.3.4 Variability distribution per chromosome

Table 12 shows the expected heterozygosity distribution over the chromosomes in the flycatcher genome.

Table 12: Variability distribution per chromosome in terms of expected heterozygosity.

Chromosome	Heterozygosity
Chr1	0.124367309
Chr1A	0.118563246
Chr1B	0.150373485
Chr2	0.128491809
Chr3	0.122185197
Chr4	0.123040461
Chr4A	0.107337915
Chr5	0.111913736
Chr6	0.104264105
Chr7	0.10951666
Chr8	0.110298655
Chr9	0.105739051
Chr10	0.099212876
Chr11	0.103223355
Chr12	0.103604492
Chr13	0.114705085
Chr14	0.108456155
Chr15	0.111079075
Chr17	0.090007274
Chr18	0.100004708
Chr19	0.09397696
Chr20	0.083488029
Chr21	0.111344138
Chr22	0.098122964
Chr23	0.115090724
Chr24	0.123339833
Chr25	0.165275265
Chr26	0.135788022
Chr27	0.08102161
Chr28	0.128416308
ChrLGE22	0.120067719
ChrZ	0.115092346

Expected heterozygosity was plotted for each 200 kb window per chromosome. Figure 9 shows and example of Chromosome 1. The peaks show how the expected heterozygosity is distributed over the chromosome and the labels on the x-axis show the scaffold IDs of the scaffolds that map onto chromosome 1. The peaks are almost similar between pied and collared flycatchers with a few variations.

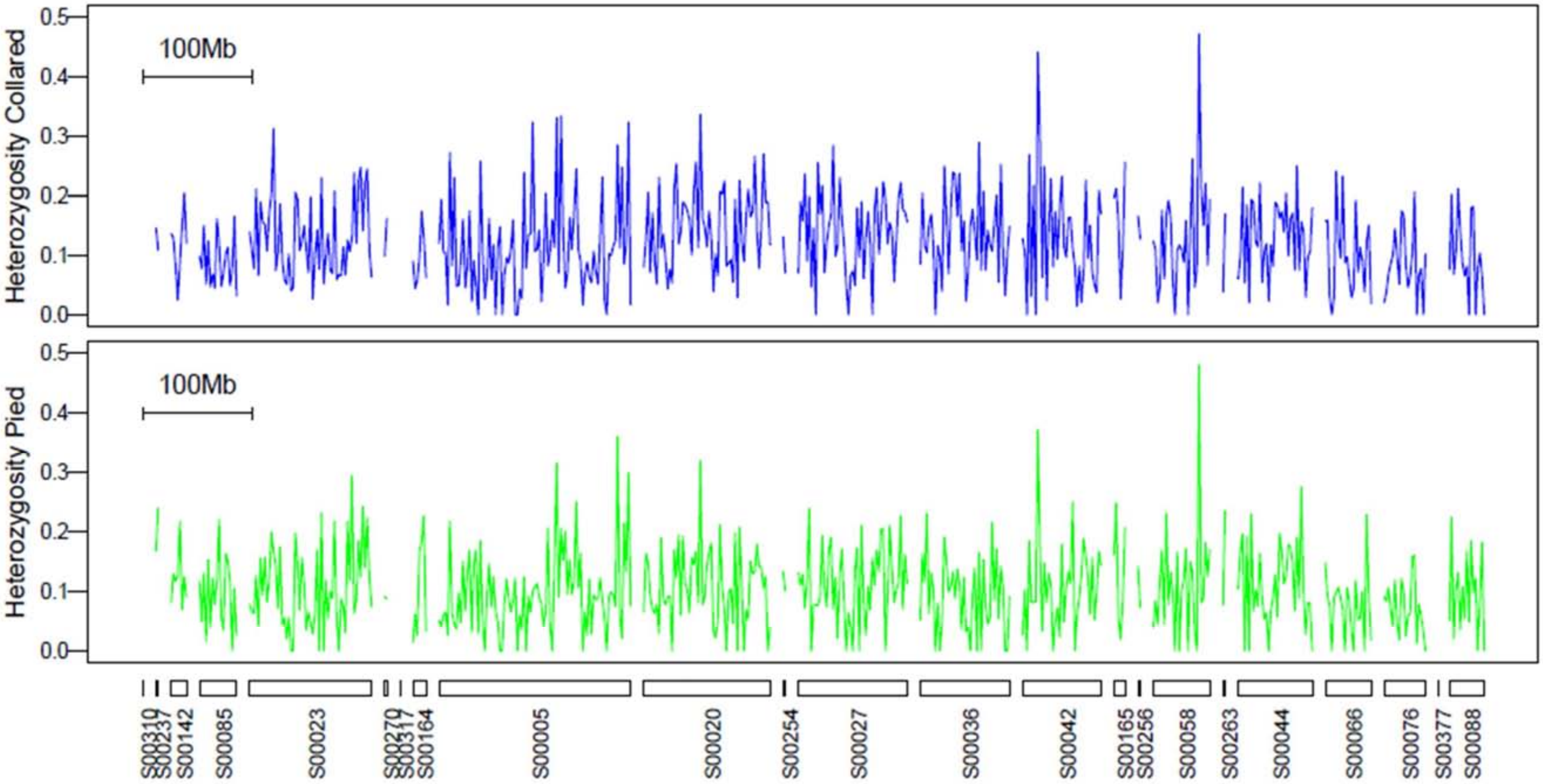

Figure 9: Expected heterozygosity distribution over Chromosome 1.

The horizontal labels show the scaffold IDs and the green plot show the expected heterozygosity distribution for pied flycatcher whereas the blue plot shows the expected heterozygosity for collared flycatcher.

3.3.5 Autosomal variability vs. sex chromosome variability

Table 13 shows a comparison between the expected heterozygosity at the autosomes and the sex chromosome which seem quite similar in this case.

Table 13: Autosomal variability vs. sex chromosome variability in terms of expected heterozygosity.

	Mean Heterozygosity
Autosomes	0.112977943
Sex Chromosome	0.115092346

This negates the idea proposed through the theory of effective population size. According to Wrieght (1931, 1938) effective population size is “the number of breeding individuals in an idealized population that would show the same amount of dispersion of allele frequencies under random genetic drift or the same amount of inbreeding as the population under consideration”. Under neutrality, the number of autosomes in a population of birds tends to be three fourth the number of sex chromosomes, i.e.:

$$N_e\ Z\text{-}A=3/4N_e\ A$$

where N_e denotes the effective population size, Z denotes the sex chromosome and A denotes the autosomes. As the effective population size of Z chromosome is less than that of the autosomes, the expected heterozygosity is also expected to be less. However, in this case it is almost equal. This can be due to the fact that all the individuals in the population are males, and chances of mutations in a male a higher than those of females as demonstrated by male-biased mutation rates (Backström, personal communication). However, further investigation into this is still required to have a clear picture.

3.3.6 Density of microsatellites per chromosome

Table 14 shows the density of microsatellites per chromosome. As can be seen, shorter the chromosome, lower is the microsatellite density and thus the total number of microsatellites on that chromosome.

The chromosomes that are shorter in size tend to have a lower number of microsatellites than the larger chromosomes. This might be due to the fact that smaller chromosomes are more tightly packed with coding sequences and chances of finding microsatellites within coding sequences are a lot less as compared to the rest of the genome (Ellegren H. personal communication).

Table 14: Microsatellite density per chromosome.

Chromosome	Microsatellite Count	Length	Density
Chr1	757699	1.07E+08	0.0070921
Chr1A	500080	73837706	0.0067727
Chr1B	3624	1282761	0.0028252
Chr2	1103000	1.56E+08	0.0070783
Chr3	721959	1.03E+08	0.0070356
Chr4	446367	61352190	0.0072755
Chr4A	133400	21315098	0.0062585
Chr5	430114	64236895	0.0066957
Chr6	247532	38752799	0.0063875
Chr7	255222	37494583	0.0068069
Chr8	214463	32160144	0.0066686
Chr9	167914	26906897	0.0062406
Chr10	141878	22339215	0.0063511
Chr11	142372	21707955	0.0065585
Chr12	135005	21969076	0.0061452
Chr13	94593	14654906	0.0064547
Chr14	105645	17369186	0.0060823
Chr15	103245	17308273	0.0059651
Chr17	68143	12363331	0.0055117
Chr18	65368	11674962	0.005599
Chr19	69300	11875408	0.0058356
Chr20	94577	16358631	0.0057815
Chr21	44950	8201983	0.0054804
Chr22	28388	5238787	0.0054188
Chr23	44951	7876571	0.0057069
Chr24	46344	8194758	0.0056553
Chr25	4291	1929498	0.0022239
Chr26	24011	4698996	0.0051098
Chr27	12517	3804572	0.00329
Chr28	27627	5865327	0.0047102
ChrLGE22	8788	2193776	0.0040059
ChrZ	455460	66629407	0.0068357

The first columns shows the chromosome name, the second column shows the total number of microsatellites on a particular chromosome, the third column shows the length of the chromosome and the fourth column shows the density of microsatellites per chromosome.

3.3.7 Degree of genetic differentiation between pied and collared flycatcher

The degree of genetic differentiation between the pied and the collared flycatcher was estimated using F_{st} calculations. Figure 10 below shows the plot for F_{st} mapped onto chromosome 1. As can be seen, no particular pattern can be observed in this case as opposed to the patterns obtained through SNP data by other group members (Figure 11). The random pattern in Figure 10 can be due to a lot of noise in the microsatellite data. The values of F_{st} lie between 0 and 1. In this case the values lie between 0.1 and 0.4 on average which show the diversity of the randomly chosen alleles within the pied and collared flycatcher relative to those found in the entire population. It expresses the proportion of genetic diversity due to allele frequency differences among populations.

However, in future, the windows in which a high F_{st} value for SNPs is observed can be compared with the ones that were obtained for microsatellite data to analyze how they look like in the microsatellite loci. The F_{st} values at each locus instead of windows can also be observed manually to see how they are different from the SNP data that was used by others in the group.

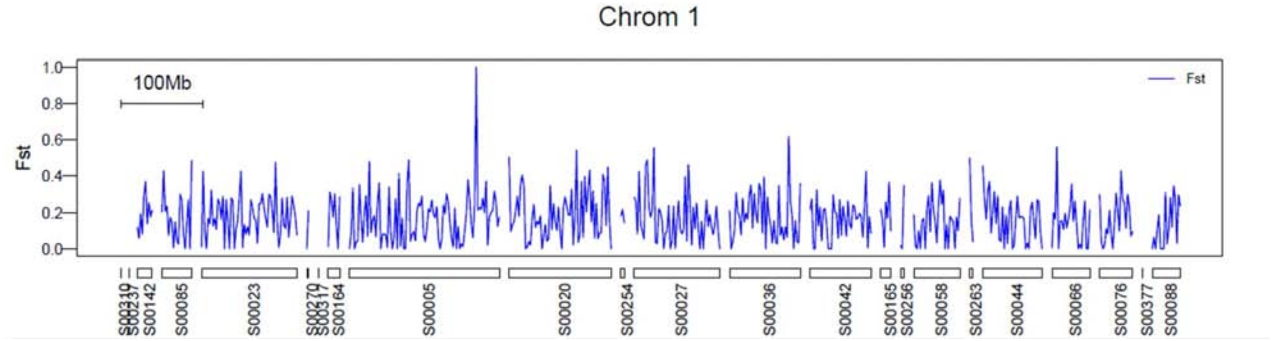

Figure 10: F_{st} distribution on chromosome 1 for pied and collared flycatcher.

The x label shows the scaffolds that make up the chromosome and their orientation.

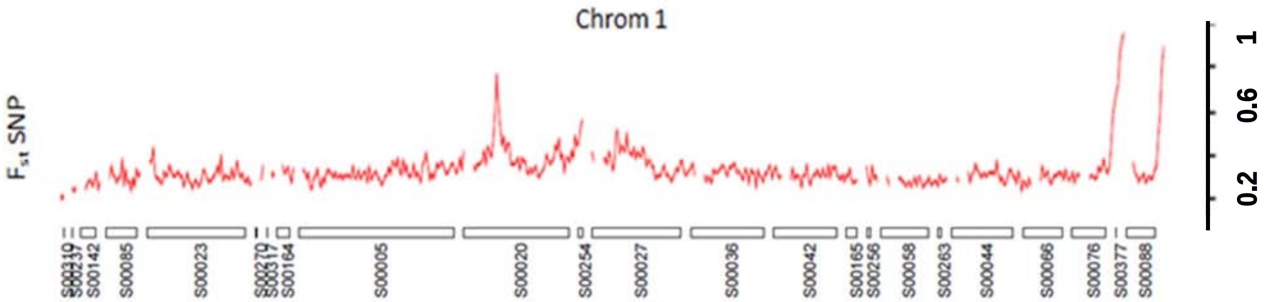

Figure 11: F_{st} pattern observed through SNP data.

This was estimated by other members in the group on the collared and pied flycatcher data.

4 Conclusions

Obtaining a clear pattern of differentiation between two species is challenging, especially when using microsatellite data to do so. I applied different bioinformatics tools and techniques on the genome sequences of pied and collared flycatchers to identify microsatellites and the degree of polymorphism at each of the identified loci excluding the mononucleotide microsatellite loci. I used the microsatellite allele frequency data to calculate the degree of genetic differentiation between the two species in terms of F_{st} and the degree of polymorphism in terms of expected heterozygosity. However, when the expected heterozygosity, F_{st} and the deviation from the expected heterozygosity at the microsatellite loci was mapped onto the chromosomes, a clear pattern was not observed. This was highly contradictory to the patterns obtained when analyzing the SNP data, as done by other members of the group. This was probably due to a large amount of microsatellite data which contributed to a lot of noise and lack of knowledge about the loci that should be considered important and the techniques that can be applied to the microsatellite data efficiently in order to reduce this noise.

However, I believe this project has nevertheless laid the ground for future investigations in terms of microsatellites in the flycatchers as well as any other species. A pipeline has already been created which can be elaborated further, with slight modifications, by the group members to manipulate the microsatellite data. This study may provide a valuable entrance point into further studies of this ecological model and I hope that it will provide a great deal of help for future microsatellite based analyses in the laboratory.

5 Acknowledgements

I would like to thank everyone who has helped me during this project. I would like to thank my supervisor, Hans Ellegren, for accepting me as a Masters student and for giving me a chance to learn more about this field. I would like to thank him for his guidance and patience throughout the project and for giving me a chance to experience how research is actually carried out, with failures and trials till we get the results that we want. I would like to thank him for the time he gave me to discuss the results and the knowledge he gave me through his experiences in our discussion sessions.

I would like to sincerely thank Linnea Smeds for her guidance and help that she provided me throughout the project. I would also like to thank her for taking time out to find where I am in the project and explaining things that I could not understand. I would like to thank her for her scripts and for helping me understand those scripts so I could use them efficiently.

I would like to thank Pall Isolfur Olason for his making me aware of the different bioinformatics tools that I can use to make work easier, accurate and fast. I would like to thank him for helping me out with problems with Uppmax.

I would like to thank Reto Burri for his help with F_{st} analysis and for being so patient with my questions and problems and for modifying his scripts so I could use them for my own data.

I would like to thank Lucie Gattepaille for her help with the deviation problems and for all her efforts to help me get the right equations for the plots.

I would like to thank Holger Schielzeth for his help with statistical problems. I would like to thank Severin Uebbing for his suggestions about the results and the comments during the presentation sessions that brought new aspects about microsatellite data into my mind.

I thank my parents and brothers for being there for me and for being patient and supportive. I would like to thank them for their guidance with every major decision I made during the time I have spent here.

I would like to express my gratitude to my friend, Raza ul Haq Akif, for his help and support with all the programming and computer problems throughout the project and for helping me learn how to program efficiently.

Last but not least I would like to thank my course coordinators, Lars-Göran Josefsson and Margareta Krabbe, for helping me with my decision to continue my studies as a PhD student. I would like to thank them for their support and understanding with all the arrangements and changes that were made after I got the position. I express sincere gratitude to Lars-Goran Josefsson for being an excellent support during the project and for always guiding me right away with all the little problems and questions I had.

6 References

[1] Molecular Biology Web Book. Mutation by Replication Errors: http://www.web-books.com/MoBio/Free/Ch7F3.htm. Date visitied: October 4, 2011.

[2] Novocraft.com. Frequently Asked Questions: http://novocraft.com/wiki/faq1#q31. Date visited: October 9, 2011.

[3] Sputnik. Available at: http://espressosoftware.com/sputnik/.

[4] SciRoKo. Available at: http://kofler.or.at/bioinformatics/SciRoKo/index.html.

[5] RepeatMasker. Available at: http://www.repeatmasker.org/.

Alatalo, R., Carlson, A., Lundberg, A. and Ulfstrand, S. 1981. The conflict between male polygamy and female monogamy: the case of the pied flycatcher *Ficedula hypoleuca*. American Naturalist. 117:738-753.

Alatalo, R. V., Gustafsson, L. and Lundberg, A. 1982. Hybridization and breeding success of collared and pied flycatchers on the island of Gotland. The Auk. 99:285-291.

Alkan, C., Sajjadian, S. and Eichler, E. 2010. Limitations of next-generation genome sequence assembly. Nature Methods. 8:61-65.

Amos, W. and Rubinstzein, D. C.1996 Microsatellites are subject to directional evolution. Nature. 12: 13-14.

Arzimanoglou, I., Gilbert, F. and Barber, H. 1988. Microsatellite instability in human solid tumors. Cancer. 82(10):1808-1820.

Bachtrog, D., Weiss, S., Zangerl, B., Brem, G. and Schlotterer, C. 1999. Distribution of dinucleotide microsatellites in the *Drosophila melanogaster* genome. Molecular Biology and Evolution. 16, 602-610.

Backström, N., Lindell, J., Zhang, Y., Palkopoulou, E., Qvarnström, A., Sætre, P. and Ellegren, H. 2010. A high-density scan of the Z chromosome in Ficedula flycatchers reveals candidate loci for diversifying selection. Evolution. 64:3461-3475.

Barbara, T., Palma-Silva, C., Paggi, G. Bered, F., Fay, M. and Lexer, C. 2007. Cross-species transfer of nuclear microsatellite markers: potentials and limitations. Molecular Ecology. 16:3759-3767.

Bell, G. I. and Jurka, J. 1997. The length distribution of perfect dimer repetitive DNA is consistent with its evolution by an unbiased single-step mutation process. J. Molecular Evolution. 44:414-421.

Benet, A., Molla, G. and Azorin, F. 2000. D (GA x TC) (n) microsatellite DNA sequences enhance homologous DNA recombination in SV40 minichromosomes. Nucleic Acids Research. 28(23):4617-4622.

Blanquer-Maumont, A. and Crouauroy, B. 1995. Polymorphism, monomorphism, and sequences in conserved microsatellites in primate species. Molecular Evolution. 41:492-497.

Brohede J. 2003. Rates and pattern of mutation in microsatellite DNA. Comprehensive summaries of Uppsala dissertations from the Faculty of Science and Technology, Uppsala University.

Buschiazzo, E., Gemmell, N. J. 2006. The rise, fall and renaissance of microsatellites in eukaryotic genomes. Bioessays. 28:1040-1050.

Buschiazzo, E. and Gemmell, N. 2010. Conservation of human microsatellites across 450 million years of evolution. Genome Biology and Evolution. 2:153-165.

Cantarel, B. L., Korf, I., Robb, S., Parra, G., Ross, E., Moore, B., Holt, C., Alvarado, A. S. and Yandell, M. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. 2008. Genome Research. 18: 188-196.

Calabrese, P. and Durrett, R. 2003. Dinucleotide repeats in the Drosophila and human genomes have complex, length dependent mutation processes. Molecular Biology and Evolution. 20:715–725.

Consortium IHGS. 2001. Initial sequencing and analysis of the Human Genome. Nature. 409 (6822):860-921.

Coulson, T. N., Pemberton, J. M., Albon, S. D., Beaumont, M., Marshall, T. C., Slate, J., Guinness, F. E. and Clutton-Brock, T. H. 1998. Microsatellites reveal heterosis in red deer. Proceedings of the Royal Society B: Biological Sciences. 265: 489-495.36.

Cracraft, J. and Barker, F. K. 2009. Passerine birds (Passeriformes). In: Hedges SB & Kumar S (eds.) the Timetree of Life. 423-431. Oxford University Press, Oxford, UK.

Crawford, A. et al. 1998. Microsatellite evolution: testing the ascertainment bias hypothesis. Molecular Evolution. 46:256-260.

Ellegren, H., Lifjeld, J. T., Slagsvold, T. and Primmer, C. R. 1995. Handicapped males and extra pair paternity in Pied Flycatchers: a study using microsatellite markers. Molecular Ecology. 4: 739-744.

Ellegren H. 2004. Microsatellites: simple sequences with complex evolution. Nature. 5 (6): 435-445.

Ezenwa, V. et al. 1998. Ancient conservation of trinucleotide microsatellite loci in polistine wasps. Molecular Phylogenetics and Evolution. 10:168-177.

FitzSimmons, N., Moritz, C. and Moore SS. 1995. Conservation and dynamics of microsatellite loci over 300 million years of marine turtle evolution. Molecular Biology and Evolution. 12:432-440.

Gemmell, N., Allen, P., Goodman, S. and Reed, J. 1997. Interspecific microsatellite markers for the study of pinniped populations. Molecular Ecology. 6:661-666.

Goldstein, D. and Schlotterer, C. 1999. Microsatellites evolution and applications. Oxford University Press.

González-Martínez, S. et al. 2004. Cross-amplification and sequence variation of microsatellite loci in Eurasian hard pines. Theoretical and Applied Genetics. 109:103-111.

Guillemaud, T., Almada, F., Serrao, S. and Cancela, M. 2000. Interspecific utility of microsatellites in fish: a case study of (CT)n and (GT)n markers in the shanny Lipophrys pholis (Pisces: Blenniidae) and their use in other Blennioidei. March Biotechnol (NY). 2:248-253.

Harr, B., Zangerl, G. B. and Schlotterer C. 1998. Conservation of locus-specific microsatellite variability across species: a comparison of two Drosophila sibling species, *D. melanogaster* and *D. simulans.* Molecular Biology and Evoution. 15:176-184.

Harr, B. and Schlötterer, C. 2000. Long microsatellite alleles in drosophila melanogaster have a downward mutation bias and short persistence times, which cause their genome-wide underrepresentation. Genetics. 155:1213-1220.

Jarne, P. and Lagoda, P. 1996. Microsatellites, from molecules to populations and back. Trends in Ecology & Evolution. 11(10):424-429.

Kashi, Y., King, D. and Soller, M. 1997. Simple sequence repeats as a source of quantitative genetic variation. Trends in Genetics. 13: 74-78.

King, D., Soller, M. and Kashi, Y. 1997. Evolutionary tuning knobs. Endeavour. 21:36-40.

Kofler, R., Schlotterer, C. and Lelly, T. 2007. SciRoKo: a new tool for whole genome microsatellite search and investigation. Bioinformatics. 23(13): 1683-1685.

Kremer, E. J., Pritchard, M., Lynch M., Yu, S., Holman, K., Baker, E., Warren, S. T., Schlessinger, D., Sutherland, G. R. and Richards, R. I. 1991. Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CCG)n. Science. 252: 1711-1714.

Kruglyak, S., Durrett, R. T., Schug, M. D. and Aquadro, C. F. 1998. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. PNAS USA. 95: 10774–10778.

Landry, P. A., Koskinen, M. T. and Primmer, C. R. 2002. Deriving evolutionary relationships among populations using microsatellites and $(\delta\mu)^2$: all loci are equal, but some are more equal than others. Genetics. 161: 1339-1347.

Leclercq, S., Rivals, E. and Jarne, P. 2007. Detecting microsatellites within genomes: significant variation among algorithms. BMC Bioinformatics. 8:125

Lifjeld, J. T., Slagsvold, T. and Lampe, H. M. 1991. Low frequency of extra-pair paternity in pied flycatchers revealed by DNA fingerprinting. Behau. Ecology Sociohiology. 29: 95 101.

Margolis, R. L., McInnis, M. G., Rosenblatt, A. and Ross, C. A. 1999. Trinucleotide repeats expansion and neuropsychiatric disease. Arch. Gen. Psychiatry. 56: 1019-1031.

Martin, P., Makepeace, K., Hill, S., Hood, D. and Moxon, E. 2005. Microsatellite instability regulates transcription factor binding and gene expression. PNAS, USA. 102(10):3800-3804.

Mitas, M. 1997. Trinucleotide repeats associated with human disease. Nucleic Acids Research. 25(12):2245-2254.

Moore, S., Hale, P. and Byrne, K. 1998. NCAM: a polymorphic microsatellite locus conserved across eutherian mammal species. Animal Genetics. 29:33-36.

Morgante, M., Hanafey, M. and Powell, W. 2002. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nature. 30: 194-200.

Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature. 420,520-562.

Primmer, C., Moller, A. and Ellegren, H. 1996. A wide-range survey of cross-species microsatellite amplification in birds. Molecular Ecology. 5:365-378.

Primmer, C., Moller, A. and Ellegren, H. 1996. New microsatellites from the pied flycatcher *Ficedula hypoleuca* and the swallow *Hirundo rustica* genomes. Hereditas. 124: 281-283.

Primmer, C. R., Saino, N., Moller, A. P. and Ellegren, H. 1988. Unraveling the process of microsatellite evolution through analysis of germ line mutations in barn swallows Hirundo rustica. Molecular Biology and Evolution. 15:1047–1054.

Quinlan, A. R. and Hall, I. M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26: 841–842.

Qvarnström, A., Rice, A., and Ellegren, H. 2010. Speciation in *Ficedula* flycatchers. Philosophical Transactions of the Royal Society B. 365:1841-1852.

Rico, C., Rico, I. and Hewitt, G. 1996. 470 million years of conservation of microsatellite loci among fish species. Proceedings of the Royal Society B : Biological Sciences. 263:549-557.

Sainudiin, R., Durrett, R., Aquadro, C. and Nielsen, R. 2004. Microsatellite mutation models: insights from a comparison of humans and chimpanzees. Genetics. 168:383-95.

Schlötterer, C., Amos, B., Tautz, D. 1991. Conservation of polymorphic simple sequence loci in cetacean species. Nature. 354:63-65.

Schlötterer, C. 2002. A microsatellite-based multilocus screen for the identification of local selective sweeps. Genetics 160: 753-763.

Schlotterer, C. 2004. The evolution of molecular markers- just a matter of fashion? Nature. 5:63-69.

Stephan, W. and Kim, Y. 1998. Persistence of microsatellite arrays in finite populations. Molecular Biology and Evolution 15:1332-1336.

Svedin, N., Wiley, C., Veen, T., Gustafsson, L. and Qvarnström, A. 2008 Natural and sexual selection against hybrid flycatchers. Proceedings of the Royal Society B. 275:735-744.

Tachida, H. and Iizuka, M. 1992. Persistence of repeated sequences that evolve by replication slippage. Genetics. 131:471-478.

Toth, G., Gaspari, Z. and Jurka, J. 2000. Microsatellites in different eukaryotic genomes: surveys and analysis. Genome Research. 10:967-981.

Treangen J. T. and Salzberg S. 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Reviews Genetics. 13:36-46.

Uebbing, S. 2011. Evolutionary investigations of transcriptome data in two flycatcher species. Degree Project in Biology. Uppsala University.

Voelkerding K., Dames S. and Durtschi J. 2009. Next-generation sequencing: from basic research to diagnostics. Clinical Chemistry. 55:641–658.

Weber, J. L. 1990. Informativeness of human (dC-dA)n.(dG-dT)n polymorphisms. Genomics 7:524-530.

Wiehe, T. 1998. The effect of selective sweeps on the variance of the allele distribution of a linked multiallele locus: hitchhiking of microsatellites. Theoretical Population Biology. 53: 272-283.

Xu, X., Peng, M. & Fang, Z. 2000. The direction of microsatellite mutations is dependent upon allele length. Nature Genetics. 24**:**396–399.