UPPSALA
UNIVERSITET

UPTEC X 19043
Examensarbete 30 hp
November 2019

Characterization of the Recombination Landscape in Red-breasted and Taiga Flycatchers

Bella Vilhelmsson Sinclair

UPPSALA
UNIVERSITET

Teknisk- naturvetenskaplig fakultet
UTH-enheten

Besöksadress:
Ångströmlaboratoriet
Lägerhyddsvägen 1
Hus 4, Plan 0

Postadress:
Box 536
751 21 Uppsala

Telefon:
018 – 471 30 03

Telefax:
018 – 471 30 00

Hemsida:
http://www.teknat.uu.se/student

Abstract

Characterization of the Recombination Landscape in Red-Breasted and Taiga Flycatchers

Bella Vilhelmsson Sinclair

Between closely related species there are genomic regions with a higher level of differentiation compared to the rest of the genome. For a time it was believed that these regions harbored loci important for speciation but it has now been shown that these patterns can arise from other mechanisms, like recombination.

The aim of this project was to estimate the recombination landscape for red-breasted flycatcher (*Ficedula parva*) and taiga flycatcher (*F. albicilla*) using patterns of linkage disequilibrium. For the analysis, 15 red-breasted and 65 taiga individuals were used. Scaffolds on autosomes were phased using fastPHASE and the population recombination rate was estimated using LDhelmet. To investigate the accuracy of the phasing, two re-phasings were done for one scaffold. The correlation between the re-phases were weak on the fine-scale, and strong between means in 200 kb windows.

2,176 recombination hotspots were detected in red-breasted flycatcher and 2,187 in taiga flycatcher. Of those 175 hotspots were shared, more than what was expected by chance if the species were completely independent (31 hotspots). Both species showed a small increase in the rate at hotspots unique to the other species.

The low number of shared hotspots might indicate that the recombination landscape is less conserved between red-breasted and taiga flycatchers than found between collared and pied flycatcher. However, the investigation of the phasing step indicate that the fine-scale estimation, on which hotspots are found, might not be reliable. For future analysis, it is important to use high-quality data and carefully chose methods.

Handledare: Madeline Chase
Ämnesgranskare: Niclas Backström
Examinator: Jan Andersson
ISSN: 1401-2138, UPTEC X19 043

Sammanfattning

Hur nya arter uppstår är en fråga som intresserar forskare. Det vanligaste sättet arter uppstår är genom geografisk isolation, så kallad allopatrisk artbildning. Två populationer som är isolerade från varandra kan av slumpen utvecklas åt olika håll, om de sedan får kontakt igen kan de vara så pass olika att de inte längre kan få avkomma. En annan form av artbildning, som är mer omtvistad, är sympatrisk artbildning där en ny art uppstår utan geografisk isolering. Vilka genetiska mekanismer som ligger bakom denna form av artbildning är inte väl undersökt. En teori är att en genvariant, kallad allel, kan vara fördelaktig för den ena population och skadlig för den andra. Detta skulle göra att individer som bär på denna allel mer sällan kan få livsduglig avkomma med den andra populationen, vilket skulle minska det genetiska utbytet mellan de två populationerna.

Biologiska processer som denna lämnar mönster i genomet som kan studeras. Sympatrisk artbildning visar sig som genomiska regioner med ett högt antal skillnader mellan två populationer, en hög differentiering. Dessa regioner med hög differentiering sticker ut ur det genomiska landskapet som öar och har benämnts som “speciation islands”, på grund av antagandet att de hade betydelse för artbildningsprocessen. Det har dock visats sig (på senare tid) att det finns andra biologiska processer som kan ge upphov till samma mönster av differentiering, och en av dessa processer är rekombination.

Rekombination sker när könsceller bildas (spermier och ägg) genom överkorsning mellan kromosomerna i ett kromosompar. Detta resulterar i kromosomer med en blandning av genetiskt material från föräldraindividerna. När man kartlägger en arts rekombinationslandskap uppskattar man hur många överkorsningar som sker på olika platser i genomet. I vissa däggdjur styrs positionen för överkorsningarna av ett protein som heter PRDM9, vilket ger upphov till regioner med en mycket högre nivå av rekombination jämfört med resten av genomet, dessa regioner kallas för “hotspots”. Vissa andra djur, t.ex. fåglar, som saknar detta protein kan fortfarande ha hotspots, men positionen av dessa är istället korrelerade till andra element, t.ex. promotorer i genomet.

Halsbandsflugsnappare och svartvit flugsnappare är två arter som används för att studera artbildning. Mellan dessa arter har man hittat “speciation islands” och funnit en stark korrelationen mellan rekombination och graden av differentiering. Man har också funnit att rekombinationslandskapet är välbevarat mellan dessa arter med en stor andel gemensamma hotspots. För att få en bättre bild av hur robusta dessa resultat är över ett längre evolutionärt tidsspann är det av intresse att studera fler artpar.

Mindre flugsnappare och tajgaflugsnappare är ett artpar som är nära släkt med halsbands- och svartvit flugsnappare. De var tills nyligen betraktade som samma art, men efter genetiska studier har man funnit att de är distinkta. Syftet med detta projektet var att kartlägga rekombinationslandskapet hos mindre- och tajgaflugsnappare.

När närliggande alleler har en större chans att ärvas tillsammans än om slumpen avgjorde, säger man att dessa alleler är i linkage disequilibrium (LD). Detta kan ske då två alleler har en större fördelaktig effekt då de förekommer tillsammans. Eftersom rekombination sker genom att byta arvsmassa mellan homologa kromosomer kan rekombination bryta kopplingen mellan alleler. Förekomsten och styrkan av LD påverkas alltså av rekombination vilket gör att man kan uppskatta rekombinationslandskapet i en population från mönster av LD.

Med denna metod uppskattades rekombinationslandskapet för mindre- och tajgaflugsnappare. Resultaten indikerade att rekombinationslandskapet mellan mindre- och tajgaflugsnappare är mindre konserverat än mellan halsbands- och svartvit flugsnappare. Men detta är som sagt bara en indikation eftersom metoderna för att uppskatta rekombinationslandskapet i de två artparen skiljer sig åt, framförallt i användningen av högkvalitativ data där studien på halsbands- och svartvit flugsnappare använde en striktare filtrering. Eftersom man uppskattar rekombinationslandskapet baserat på signaler i genomet kan bakgrundsbrus påverka resultatet och leda till felaktiga slutsatser.

Detta projekt kan användas som ett preliminärt resultat för att det finns intressanta skillnader mellan de två olika artparen av flugsnappare. Det pekar också på vikten av att använda högkvalitativ data och lämpliga metoder.

Contents

Abbreviations and acronyms

bp - Basepairs
cM - Centimorgan
kb - Kilo basepairs
LD - Linkage disequilibrium
PRDM9 - Positive-regulatory domain zinc finger protein 9
SNP - Single nucleotide polymorphism
VCF - Variant call format

1 Background

A central question to the field of speciation research is how new species emerge. Fundamental to this question, is how one defines species. The most common definition is the biological species concept, which is based on reproductive compatibility. If two individuals can produce viable fertile offspring, they are reproductively compatible and of the same species (Mayr 1942). Reproductive compatibility between populations of the same species is maintained through gene flow, which is the exchange of genes between populations. Therefore speciation occurs as populations become reproductively incompatible, which arises from reproductive barriers. A reproductive barrier reduces or stops gene flow and can be a physical barrier, for example morphology or geography, or a genetic barrier. A genetic barrier can for example be caused by incompatibility between gene variants or possibly a single locus, however, the latter is debated.

For a long time it was questioned if speciation was possible in sympatric populations, populations that share the same environment. This is contrary to the most common form of speciation, when populations diverge in isolation, called allopatric speciation. In allopatric speciation, genetic drift, mutations and selection can lead to divergence, and reproductive isolation can then arise as a by-product of these processes. For sympatric speciation, reproductive barriers have to arise during the influence of gene flow. Some examples of sympatric speciation have been observed (Campbell 2015, Wolf & Ellegren 2017) and can arise, for example through changes in ploidy (a change in the number of chromosomes) or strong sexual selection (Campbell 2015). The common pattern for both forms of speciation is that gene flow has to be interrupted, either through geographic isolation or other mechanisms.

How reproductive barriers function on the genomic scale is not well understood (Ravinet *et al.* 2017). One method to find genomic regions that are important for speciation is to compare the genomes of closely related species or divergent populations. Regions with more differences between the two populations are said to have a higher level of genomic differentiation. From these kind of comparisons the idea of ”speciation islands” appeared (Cruickshank & Hahn 2014). Speciation islands are genomic regions that have a higher level of differentiation than the genomic background between two populations. The hypothesis is that at loci with alleles beneficial to one of the populations and deleterious to the other population, gene flow between the two populations will be reduced. This reduction of gene flow is due a reduced fitness in individuals that experience gene flow at these loci, which result in a lower probability of survival for these individuals.

As a consequence of linkage between loci some loci are more often inherited together than would be expected if all loci were inherited independently, this is called linkage disequilibrium (LD).

Because of LD, gene flow will be restricted at loci nearby loci under selection. This will lead to a larger region with reduced gene flow that can diverge between the two species, leading to higher differentiation, while the rest of the genome is homogenized by gene flow (Burri *et al.* 2015, Cruickshank & Hahn 2014, Ravinet *et al.* 2017, Wolf & Ellegren 2017). It was thus believed that regions of higher differentiation contained loci important for speciation, but with more studies it has been shown that these patterns can arise through other mechanisms (Cruickshank & Hahn 2014, Wolf & Ellegren 2017) and the importance of so called speciation islands is now unclear.

LD influences the effect of selection, as selection on one locus will impact linked loci. A beneficial locus in linkage with a deleterious locus can be removed from the population because of the selection on the deleterious locus and the opposite can also happen, where a deleterious mutation in linkage with a beneficial locus can be fixed in the population. When an allele is removed or fixed by selection and brings loci in linkage to extinction of fixation it is called a selective sweep when positive selection is acting, and background selection when purifying selection is acting. Selective sweeps and background selection reduce the nucleotide diversity, and a reduced diversity can increase the differentiation between two populations, giving similar patterns as a speciation island.

Recombination can break linkage between loci, and therefore mediate how large regions are affected by linked selection (Ravinet *et al.* 2017). The prediction is that in regions of lower recombination, linked selection is more pronounced and this would decrease the diversity which results in an increase in differentiation (Cruickshank & Hahn 2014, Cutter & Payseur 2013, Wolf & Ellegren 2017). The occurrence of speciation islands is predicted to be correlated with the recombination rate, something that has been found for example in flycatchers (Burri *et al.* 2015). In a more recent study on flycatchers (Rettelbach *et al.* 2019), it was found that background selection explains most, but not all, of the patterns of diversity and that diversity is strongly correlated with recombination rate. They conclude that selection against gene flow, as a cause of regions with higher differentiation, can be rejected. For now, it looks like regions of elevated genetic differentiation say little about speciation and are mostly affected by the recombination landscape.

1.1 Recombination

Recombination is the exchange of genetic material between parental chromosomes through crossovers during the meiosis (Campbell 2015, Figure 1). This produces new genetic combinations that can be evolutionary beneficial. Recombination can also break apart already existing beneficial combinations which makes recombination itself subject to selection (Burri *et al.* 2015, Stapley *et al.* 2017).

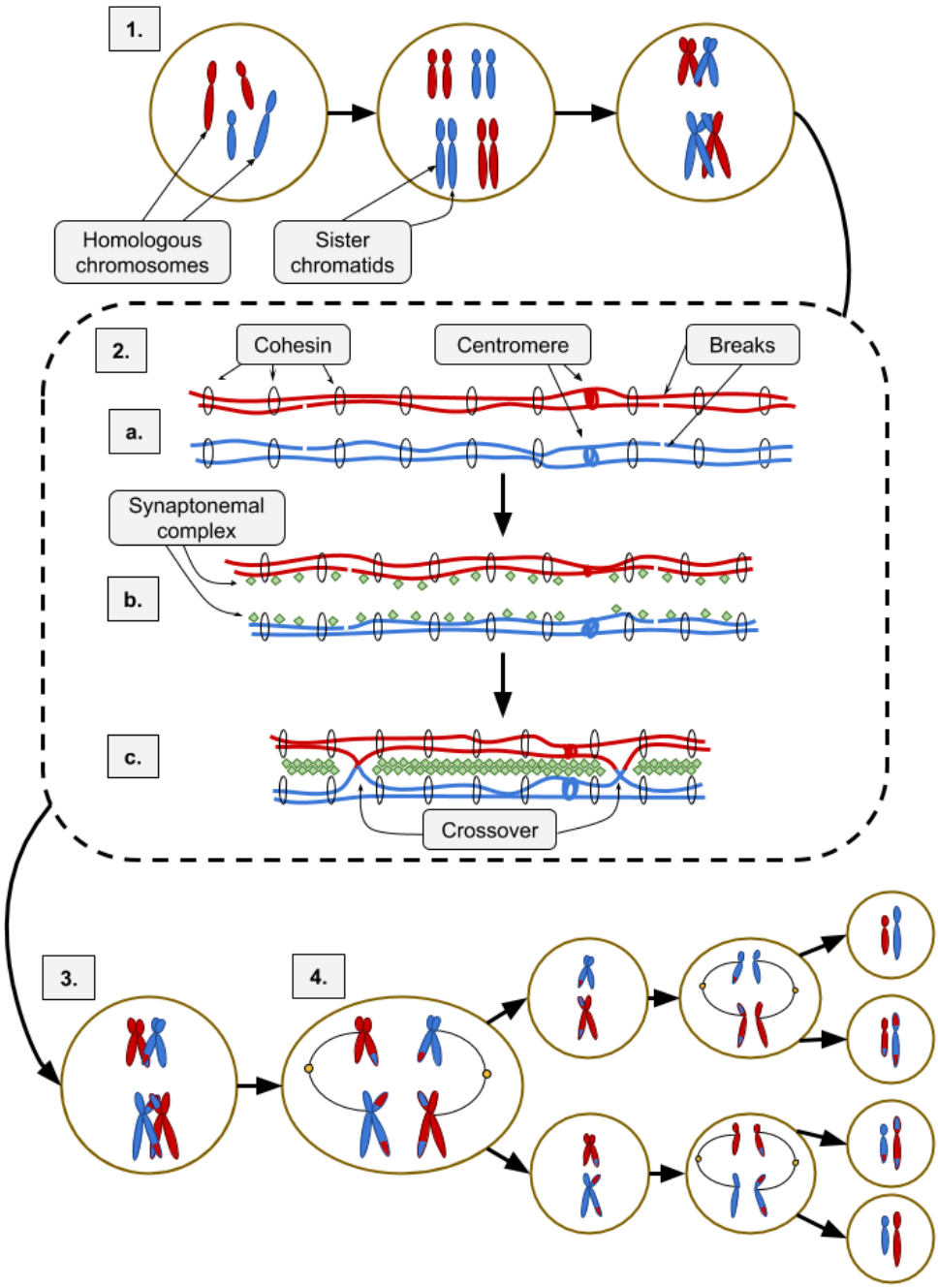

Figure 1: A sketch of the different steps in the meiosis, detailing the crossover event. **1.** In the beginning of the meiosis, the chromosomes are duplicated and homologous chromosomes line up. **2a.** Sister chromatids are held together by the protein cohesin and double stranded breaks are created at the same positions in non-sister chromatids. The chromosomes condense more and more in each step. **2b.** Synaptonemal complexes hold the homologous chromosomes together. **2c.** The double stranded breaks are closed by joining the ends of non-sister chromatids, creating crossover. **3.** The synaptonemal complex is dissolved but the homologous chromosomes stay attached because of the cohesin between sister chromatids. **4.** In the last steps of the meiosis, chromatids segregate to individual cells.

In some species, in order for proper segregation between chromosomes in the meiosis, one crossover per chromosome is obligatory (Stapley *et al.* 2017). This is for example believed to be the case in flycatcher (Smeds *et al.* 2016). This results in a higher mean recombination rate in smaller chromosomes, giving rise to a correlation between chromosome size and recombination rate. This correlation is particularly pronounced in birds, which have a large variation in chromosome size (Stapley *et al.* 2017).

A linkage map is a map of the genome measured in centimorgans (cM) where one cM denotes that two loci have a 1% chance to be separated by a crossover event. If recombination was random across the genome, the chance of a crossover between two loci would be linearly dependent on the physical distance, but this is not the case as the chance of a crossover varies over the chromosomes. Linkage maps are constructed using genetic markers from a large pedigree and makes it possible to anchor and order scaffolds along chromosomes.

Pedigree based recombination rate is calculated by sequencing individuals in a pedigree and inferring crossover events on a per generation basis (Kawakami *et al.* 2014). Another form of recombination rate can be estimated from LD, which gives a historical view of the recombination rate for the whole population. This measurement of the recombination rate is called the population-scaled recombination rate and is in this report referred to as LD recombination rate. It is possible to estimate the recombination rate from LD because recombination breaks linkage between loci, LD is therefore affected by the recombination rate. A weakness in using LD to estimate the recombination rate is that the population demographic history (for example expansions and bottlenecks) affect LD, which can make the estimate non-reliable if the population under investigation has gone through demographic events (Stapley *et al.* 2017).

1.1.1 Recombination Hotspots

The recombination rate is not constant across the genome and in many species so called recombination hotspots exist, which are regions of a few hundred base-pairs that have many times higher recombination rate than the surrounding genome (Stapley *et al.* 2017). In several mammals, the location of hotspots are governed by the PRDM9 gene, which is a protein that binds to specific motifs in the genome during the meiosis and promotes crossovers (de Massy 2013). Due to the fast evolution of the binding motifs of PRDM9, hotspots in mammals change rapidly and can even differ between sub-species of mice (de Massy 2013). In other taxonmical groups, for example birds that lack the PRDM9 gene, hotspots are more conserved and linked to different genomic features, such as the location of promoters (Singhal *et al.* 2015, Stapley *et al.* 2017).

1.2 *Ficedula* Flycatchers

The collared flycatcher (*Ficedula albicollis*) and pied flycatcher (*F. hypoleuca*) have been the focus of many studies, and are used as a model organism for speciation research (Ellegren *et al.* 2012). The males of both collared and pied flycatchers have black and white patterns, and are often referred to as black and white flycatchers (not to be confused with the Swedish name for pied flycatcher which directly translates to black and white flycatcher). The collared and pied flycatchers are believed to have diverged less than 1 million years ago (Nadachowska-Brzyska *et al.* 2013). A high-density linkage map is available for the collared flycatcher, and has been used to make the most recent genome assembly for collared flycatcher (Kawakami *et al.* 2014). In 2017 the LD recombination rate for collared and pied flycatcher was estimated (Kawakami *et al.* 2017), giving a fine-scale view of the recombination rate in flycatchers. They found that the landscape is highly conserved between the two species, which also has been seen between other birds (Stapley *et al.* 2017), and hotspots were enriched for different genomic features. A heterogenous landscape of differentiation between collared and pied flycatchers has been observed (Ellegren *et al.* 2012). Much of the variation has been shown to be explained by linked selection in regions of low recombination (Burri *et al.* 2015), highlighting the importance of controlling for the recombination rate when looking for patterns of speciation.

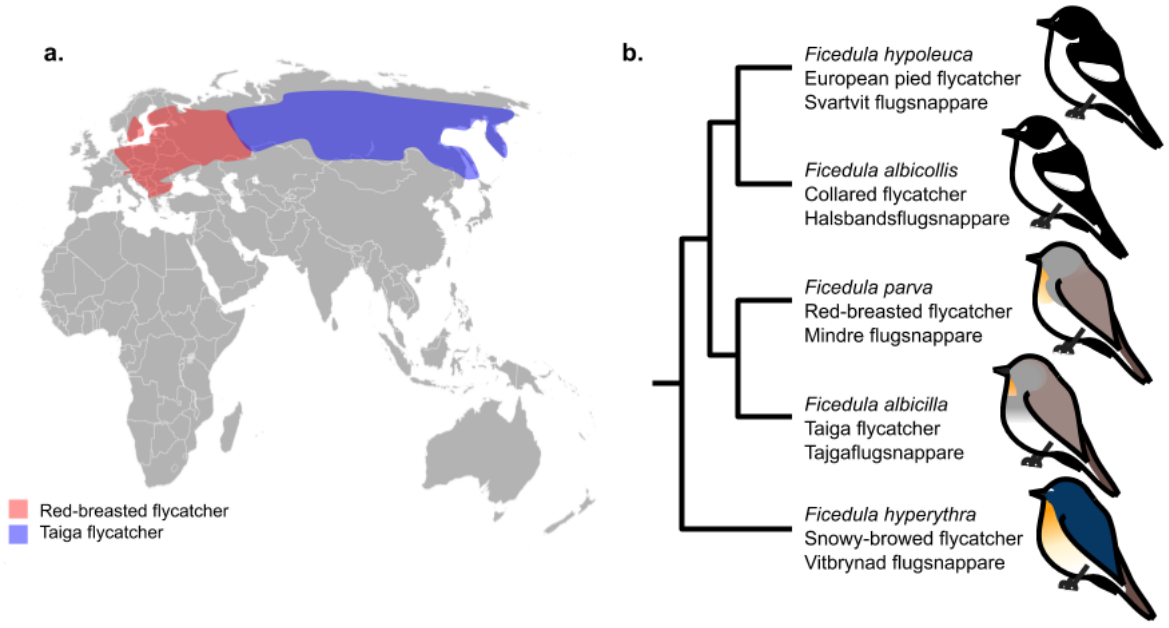

Figure 2: a. A rough map of the breeding regions of red-breasted and taiga flycatchers (BirdLife International 2017 & 2018). **b.** Cladogram of the *Ficedula* flycatchers mentioned in this report. Relationship between species from (Moyle *et al.* 2015).

Red-breasted flycatcher (*F. parva*) and taiga flycatcher (*F. albicilla*) are another species pair of flycatchers, which are in the same genus as the collared and pied flycatcher (Moyle *et al.* 2015, see Figure 2). It was previously believed that taiga flycatcher was a sub-species of the red-breasted flycatcher, but after genetic analyses (Hung & Zink 2014) they were found to be distinct species. The red-breasted and taiga flycatchers are very similar phenotypically but have minor differences in morphology, and different distributions. The red-breasted flycatcher breeds in Eastern Europe while the taiga flycatcher breeds in Siberia.

1.3 Theory

A common approach in speciation genomic studies is to look at SNPs in the genome. A SNP is a single nucleotide polymorphism, a position in the genome where individuals in a population can have different alleles. SNPs are found by performing variant calling, which extracts all differences between a sample sequence and the reference genome. The allele combination one individual has at a SNP is called its genotype. The minor allele frequency is the frequency of the least common allele.

1.3.1 Phasing

A haplotype can refer to a sequence block that is inherited as a single unit or to a whole chromosome, in this report it refers to the latter. Each individual has two haplotypes for each chromosome and phasing is the process of assigning alleles to haplotypes. From the sequencing it is not possible to know which allele was on which copy of the chromosome but this can be statistically computed either from pedigree information or patterns of LD (Browning & Browning 2011).

1.4 Project Aim

The aim of the project was to estimate the LD recombination rate for red-breasted and taiga flycatchers and look at the conservation between the landscapes.

2 Methods

All custom scripts used can be found in the Git-repository: https://github.com/Bella2347/Git-exjobb.

2.1 Data

The collection of samples and generation of data files had been performed prior to this project. The origin and generation of data is described below. The variant calling and filtration was performed by Madeline Chase.

Whole genome sequencing was performed on 15 samples of red-breasted flycatcher, and 72 samples of taiga flycatcher. Samples were provided by Bell Museum of Natural History, University of Minnesota. Red-breasted flycatcher samples were collected from five populations in Western Russia, and taiga flycatcher samples were collected from 13 populations in Eastern Russia and Mongolia. Additionally, to identify the ancestral state, sequence data from one individual of snowy-browed flycatcher (*F. hyperythra*), collected in Indonesia, and 95 individuals of collared flycatcher, collected on the island Gotland, was included in the analysis.

Raw sequencing reads of each species were mapped to the collared genome assembly version fAlb15 using BWA 0.7.15 (Li & Durbin 2009). Variant calling was then performed following GATK best practices protocol. Duplicates were marked and bam files for each sample were merged using Picardtools 2.0.1. In order to perform base quality score recalibration, an initial round of variant calling was performed, and the filtered SNPs were then used for recalibration. Individual gVCF files were generated by GATK HaplotypeCaller 3.7, and joint genotyping was performed with the GenotypeGVCFs tool. A set of hard filters were applied to the raw SNPs ("QD lt; 2.0 || FS gt; 60.0 || MQ lt; 40.0 || MQRankSum lt; -12.5 || ReadPosRankSum lt; -8.0"), and base quality score recalibration was performed on the BAM files. This first step of variant calling was performed separately for the 95 collared flycatchers, the 15 red-breasted samples, 72 taiga samples and the single snowy-browed sample.

After performing base quality score recalibration, another round of variant calling was performed. Again, individual gVCF files were generated by GATK HaplotypeCaller, but joint genotyping was now performed with all samples combined. The same set of hard filters was applied to the raw SNPs. Then, a series of additional filtration steps were performed in order to retain only high-quality genotypes. First, multi-allelic sites were removed and genotypes with a read depth less than five were marked as missing. In order to exclude potentially collapsed regions, genotypes

with a read depth of greater than 186 were also considered missing. A minimum genotype quality score was included, and was set to 30 for both autosomal genotypes, and male genotypes on the Z-chromosome. This threshold was lowered to 15 for female genotypes on the Z-chromosome. After filtration, seven individuals of taiga flycatcher were identified as having extremely high percentages of missing data, and were thus removed from further analysis.

After the filtering the data consisted of 15 samples from red-breasted flycatcher, 65 samples from taiga flycatcher, 95 samples from collared flycatcher and one sample from snowy-browed flycatcher summarizing to 85,871,984 SNPs.

2.1.1 Sub-setting the Data

Only scaffolds that were oriented and mapped to chromosomes were extracted from the data, using a python script, to reduce the processing time in further steps. In the reference assembly used, 93.4% of scaffolds were mapped and oriented to chromosomes (Kawakami *et al.* 2014) and those that were not were mainly small scaffolds that mapped poorly to the reference genome, making the data loss when un-mapped scaffolds were excluded rather small.

Red-breasted and taiga individuals were extracted to separate VCF-files using a python script. Using VCFtools (Danecek *et al.* 2011) both files were filtered on a minor allele frequency of one to remove non-variant sites. After sub-setting the data 14,302,132 SNPs remained for red-breasted flycatcher and 37,804,612 SNPs for taiga flycatcher.

2.2 Phasing

Plink v1.9b (Chang *et al.* 2015) and a python script were used to create input files for fastPHASE (Scheet & Stephens 2006). No information on gender for the samples were available for Plink and Plink could therefore not create correct input files for the Z-chromosome. The Z-chromosome was therefore excluded from further analysis.

2.2.1 Setting Parameters

In a previous version of fastPHASE it was possible to get an estimate of the optimal number of clusters (**K**) to use during the phasing. In the latest version (v1.5) this was not longer possible, fastPHASE v1.4 was therefore used.

For two scaffolds for each species the optimal values of **K** were found by testing values from two to 20 with a step-size of two. For the red-breasted flycatcher scaffolds N00001 (393,323 SNPs) and N00039 (95,682 SNPs) were used. Scaffold N00001 was chosen to try to capture the maximum complexity in the data, which was believed to be in the larger scaffolds. Scaffold N00039 was chosen to capture the middle end of the complexity. The optimal values of **K** were 6 and 4 for each scaffold respectively. A **K** of 6 was then used to phase all files.

For taiga flycatcher scaffold N00020 (490,170 SNPs) and N00200 (46,688 SNPs) were used. Scaffold N00020 was chosen because fastPHASE has a character limit of 500,000 and the larger scaffolds exceeded this limit and those input files therefore had to be tweaked before they could be used by fastPHASE. Scaffold N00200 was used as the lower end of the complexity. The optimal values of **K** were 18 and 8 respectively, indicating that this data-set was more complex than the red-breasted data-set. A **K** of 12 was then used to phase all scaffolds. An intermediate value of **K** was used to accommodate for the variation in **K** and to keep the computation time down since it increases with **K**.

Both red-breasted and taiga samples were phased using the default parameter settings, 20 random starts and 35 iterations. The red-breasted samples were phased without using sub-population information and the taiga samples were phased with two sub-population labels. The sub-population labels for taiga flycatcher were obtained through a PCA analysis preformed by Madeline Chase. FastPHASE can impute missing genotypes, which is the default. This was done for both species.

2.2.2 Evaluating the Accuracy of the Phasing

To test how reliable the phasing was, two re-phasings were done for scaffold N00100 (Chromosome 13). One re-phasing was preformed with imputation and one without imputation. For these re-phased scaffolds all the same steps were preformed as for the first phasing. The conformity between the estimates based on different phases was examined by calculating the Pearson correlation coefficient (r) between them.

2.3 LD Estimation of the Recombination Rate

The LD recombination rate was estimated using LDhelmet v1.10 (Chan *et al.* 2012). Theta was set to 0.0031, based on estimates of π calculated from the same samples. Theta is the population scaled mutation rate which is calculated from the nucleotide diversity. π is the average pairwise difference between two randomly chosen sequences in a sample and is a measurement of diversity.

The recommended grid of p-values and number of padé coefficients were calculated. The mutation matrix used was established by Mugal *et al.* (2013) from chicken intergenic regions (Table 1). A mutation matrix specify the probability of a mutation from one allele to another.

LDhelmet calculates the recombination rate as [1/bp]. To translate this unit into [cM/Mb], which is the unit for per-generation recombination rate, the rate in each 200 kb window had to be divided by the genome-wide effective population size. The effective population size was not known for red-breasted and taiga flycatchers and therefore all the values of the recombination rate in this report are in [1/bp]. This results in that the magnitude of the recombination rate was not directly comparable between the two species or to the rate in the linkage map.

Table 1: Mutation matrix (Mugal *et al.* 2013).

	A	C	G	T
A	0.1125	0.1478	0.5654	0.1743
C	0.2061	0.0000	0.1939	0.5999
G	0.5999	0.1939	0.0000	0.2061
T	0.1743	0.5654	0.1478	0.1125

2.3.1 Ancestral State

To derive the ancestral state for each SNPs, 95 samples from collared flycatchers and one sample from *F. hyperythra* were used as out-groups. If all samples in two of the three groups (two out-groups and one in-group) were monomorphic for the same allele, this allele was set as the ancestral state. For all SNPs detected between these species, 32.1% sites were polarized.

The prior probability of the ancestral allele was set to 0.97 and to 0.01 for the other alleles. For SNPs without an ancestral state the prior probability was set to 0.25 for all alleles. Along with the mutation matrix, this information was used to improve the accuracy of LDhelmet.

2.3.2 Running LDhelmet

LDhelmet had a haplotype limit of 50, therefore only the 25 best sequenced taiga flycatcher samples were used. The best samples were chosen to have the lowest percentage missing data and the highest coverage. Sequence and position files were generated with a python script. Five independent runs were made for each species with a block penalty of 10 (Kawakami *et al.* 2017) and a burn in of 200,000 with 2,000,000 iterations.

For unknown reasons, no estimations for scaffold N00040-N00043 were obtained and they were therefore excluded from further analysis.

2.4 Analysing the LD Recombination Rate

The average recombination rate over all five runs was calculated and used for analysis if nothing else is stated. All analyses were done using R (R Core Team 2018).

2.4.1 Calculations in 200 kb Windows

The average LD recombination rate was calculated in 200 kb non-overlapping windows, which is the scale at which the linkage map had signal. Pearson correlation coefficients were calculated between the LD recombination rates of the two species and the linkage map, which was based on collared flycatcher. Since a correlation between chromosome length and recombination rate had been observed for flycatchers (Kawakami *et al.* 2017), a partial correlation while controlling for chromosome length was calculated. This was done by regressing the estimated LD recombination rate and the linkage map on chromosome length and then calculating the Pearson correlation between the residuals from each regression.

2.4.2 Detecting Hotspots

Hotspots were identified by looking at the recombination rate between pairs of SNPs. If the rate between a pair of SNPs was at least 10 times higher than the average in the surrounding 200 kb, it was saved as a potential hotspot. Adjacent potential hotspots were merged, and if their total length was at least 750 bp it was deemed as a hotspot. The SNP density in the surrounding 200 kb had to be at least 1 SNP/1 kb. These limits were chosen based on the definition used by Kawakami *et al.* (2017). This was done to be able to compare the results obtained to what was found for collared and pied flycatcher.

In the ends of the scaffold the surrounding window could be less than 200 kb. How this affected the hotspot detection was investigated by using different window lengths, from 10 kb to 400 kb. What was found was that in shorter windows, fewer hotspots were detected (Figure 3). These end cases were therefore not handled as a special case, the window was only truncated.

The detection of hotspots was done separately for each run of LDhelmet, and only hotspots that were detected in all five runs with an overlap of at least 750 bp were used as hotspots.

To find hotspots that were shared between red-breasted and taiga flycatchers, a hotspot was said to be shared if the middle of the hotspot was at most 3 kb apart. A permutation test was done to find the expected number of shared hotspots by chance. The test was preformed as follows: Hotspots

were assigned randomly along the scaffolds. For each scaffold there were as many randomly assigned hotspots as there were true hotspots on that scaffold. This was done for both species and then the number of randomly assigned hotspots that were at most 3 kb apart were counted. 1,000 permutations were done.

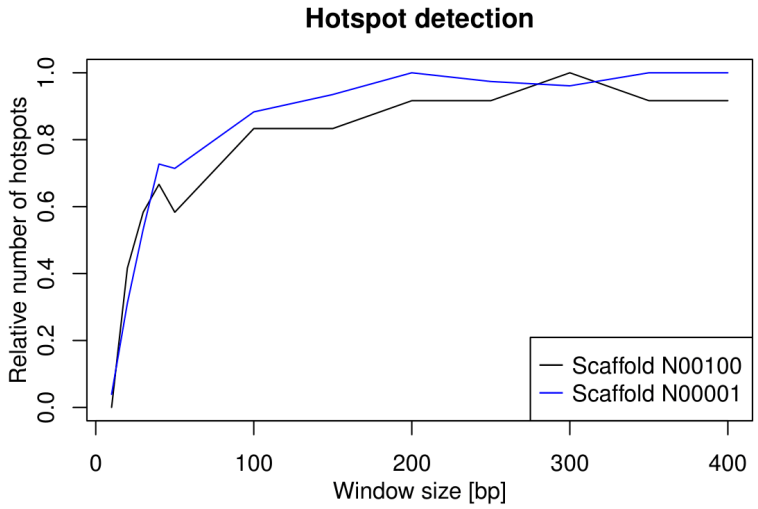

Figure 3: The effect of the length of the window, used to normalize the rate when detecting hotspots, on the number of hotspots detected. The y-axis is normalized by dividing the number of hotspots detected by the maximum number of hotspots detected using any window size.

3 Results

3.1 Comparison to the Linkage Map

When the mean recombination rate was calculated for each scaffold and grouped by chromosomes, a negative correlation between chromosome size and recombination rate was found in both red-breasted and taiga flycatcher (Figure 4). This is a pattern which has been found in other birds (Stapley *et al.* 2017) and was therefore expected in this data-set.

The correlation between the two species and the linkage map (Pearson correlation coefficient $r = 0.480$ for red-breasted and $r = 0.484$ for taiga flycatcher) was lower than the correlation between the two species ($r = 0.677$)(Table 2, Figure 5). This was expected since red-breasted and taiga flycatchers are more closely related to each other than to collared flycatcher, on which the linkage

map is constructed. When controlling for chromosome length, the correlation decreases between red-breasted flycatcher and the linkage map and between red-breasted and taiga flycatcher (Table 2, Figure 6).

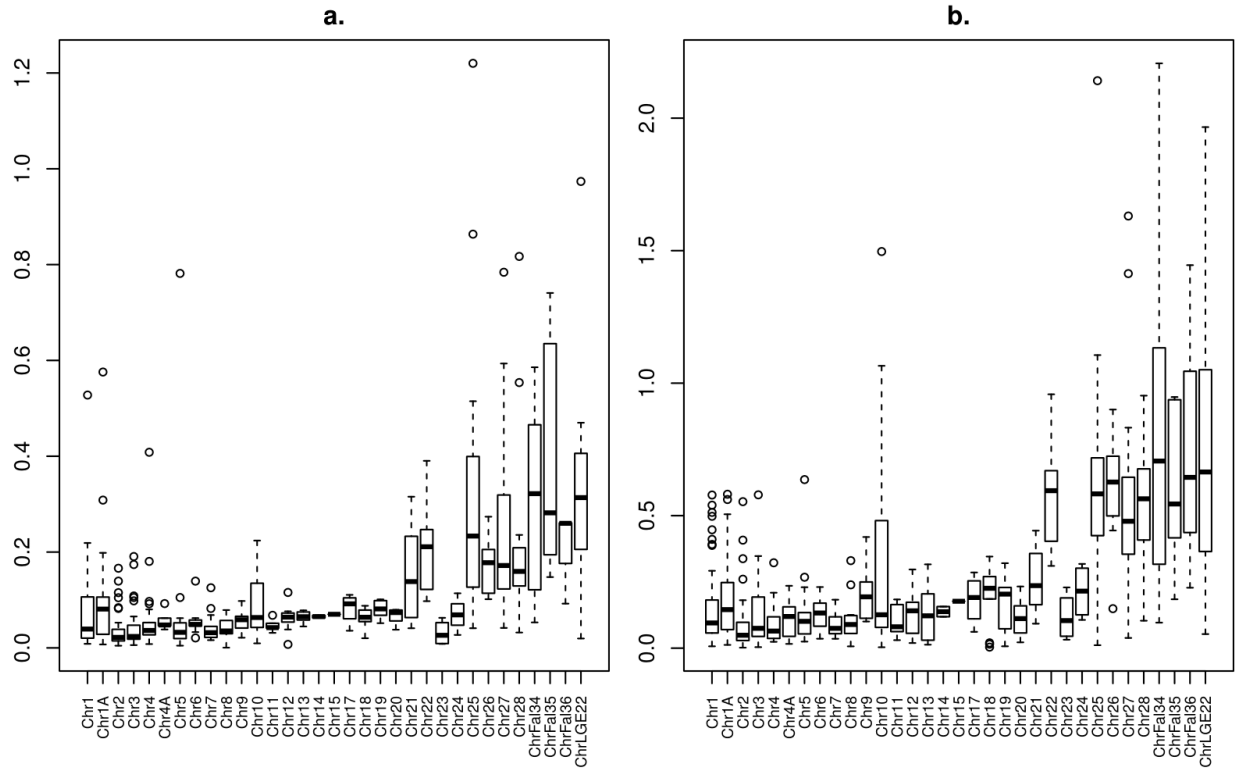

Figure 4: Mean LD recombination rate in **a.** red-breasted and **b.** taiga flycatcher per scaffold grouped by chromosome.

Table 2: Pearson correlation coefficients for complete and partial correlation between the LD recombination rate in each species and the linkage map based on collared flycatcher. The partial correlation is controlling for chromosome length. 95-percentile confidence interval in parentheses, p-value < 2.2e-16 for all correlations.

	Correlation	Partial correlation
Red-breasted and linkage map	0.480 (0.458 - 0.501)	0.364 (0.339 - 0.388)
Taiga and linkage map	0.484 (0.462 - 0.505)	0.483 (0.462 - 0.505)
Red-breasted and taiga	0.677 (0.662 - 0.692)	0.553 (0.533 - 0.572)

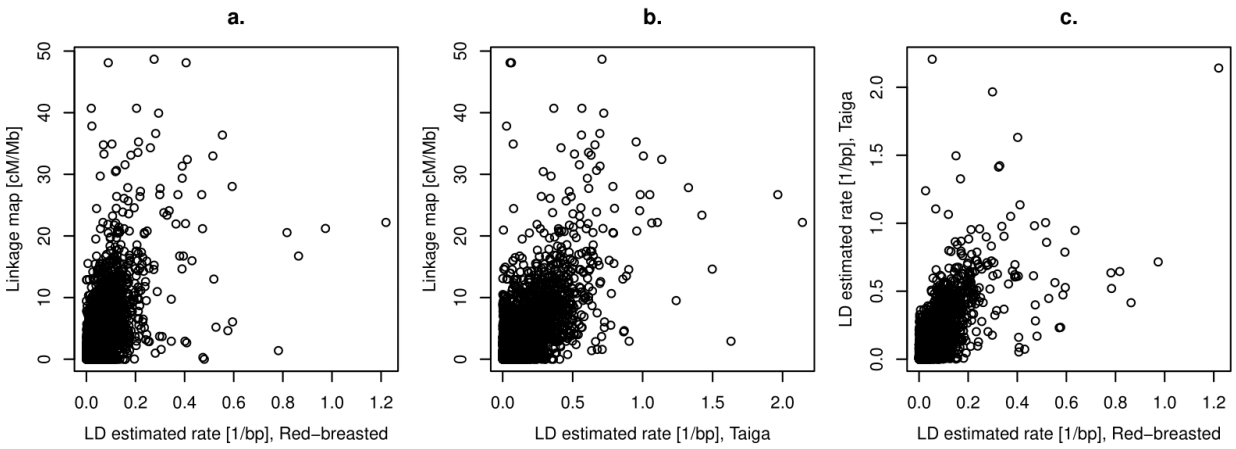

Figure 5: Relationship between LD recombination rates for red-breasted and taiga flycatchers and the linkage map based on collared flycatcher. Outliers have been left out. **a.** Red-breasted flycatcher against the linkage map. **b.** Taiga flycatcher against the linkage map. **c.** Red-breasted flycatcher against taiga flycatcher.

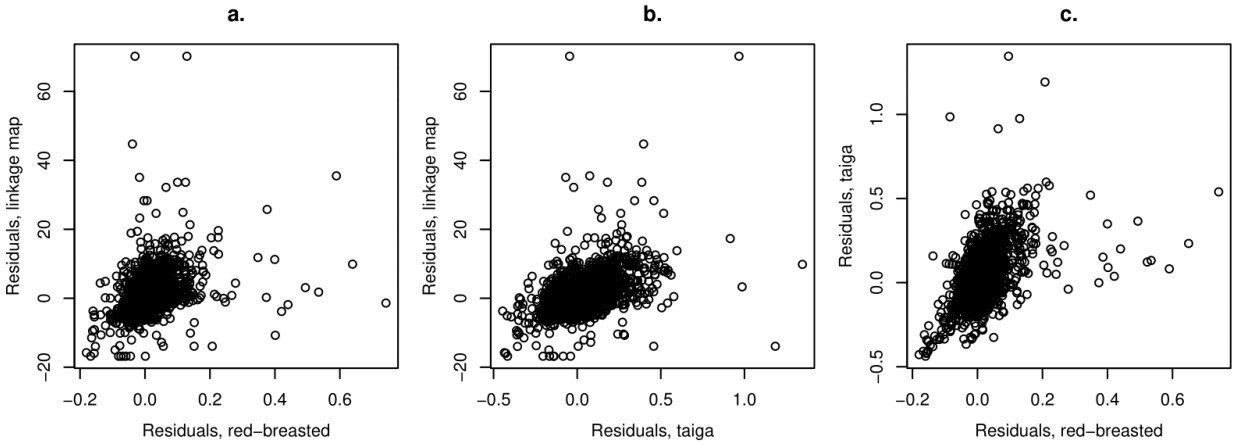

Figure 6: Residuals on which partial correlation has been calculated. **a.** Residuals from regressing red-breasted LD recombination rate and the linkage map on chromosome length. **b.** Residuals from regressing taiga LD recombination rate and the linkage map on chromosome length. **c.** Residuals from regressing red-breasted and taiga LD recombination rate on chromosome length.

3.2 Hotspots

2,176 hotspots were found in red-breasted flycatcher and 2,187 in taiga flycatcher. These numbers are in the range of what has been found in collared (400-1,482) and pied flycatchers (1,485-3,085) (Kawakami *et al.* 2017). The average length of hotspots for red-breasted flycatcher was 1,570 bp and for taiga flycatcher 1,589 bp (collared flycatcher: 1,640 bp). For hotspots found in red-breasted

flycatcher, 175 were shared with taiga flycatcher. The same number for taiga was 176, which means that one hotspot in red-breasted flycatcher is at most 3 kb from two hotspots in taiga flycatcher. The hotspot in red-breasted flycatcher that corresponds to two hotspots in taiga flycatcher is long, 4,895 bp, and the two hotspots in taiga flycatcher are only separated by 468 bp. The number of shared hotspots are lower than expected but still more than expected by chance which was tested using a permutation test (Figure 7).

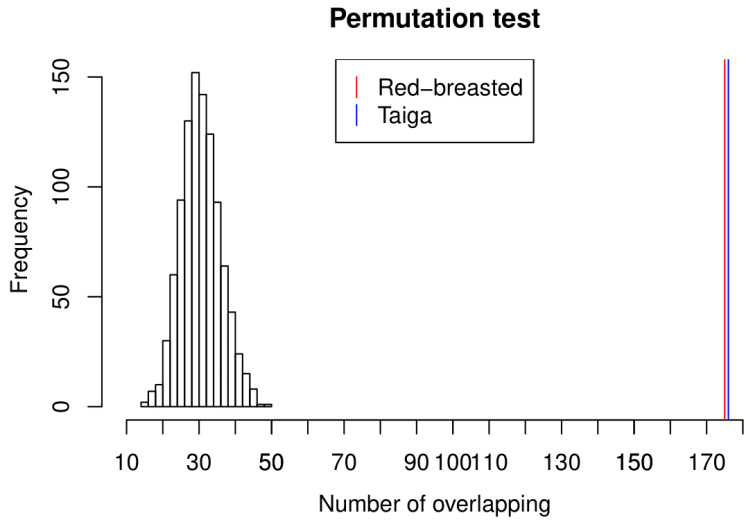

Figure 7: Expected number of overlapping hotspots between 2,176 hotspots for red-breasted flycatcher and 2,187 for taiga flycatcher over 1000 permutations.

Hotspots shared between the two species showed a similar increase in the rate relative to the mean rate in 40 kb regions surrounding hotspots. At the positions of hotspots unique to one species, the other species showed an increase in the relative rate (Figure 8).

3.3 Re-phasing

For red-breasted flycatcher, four hotspots were shared between the second and third phase using a 750 bp overlap, no hotspots were shared between all phases. The same pattern was found for taiga flycatcher.

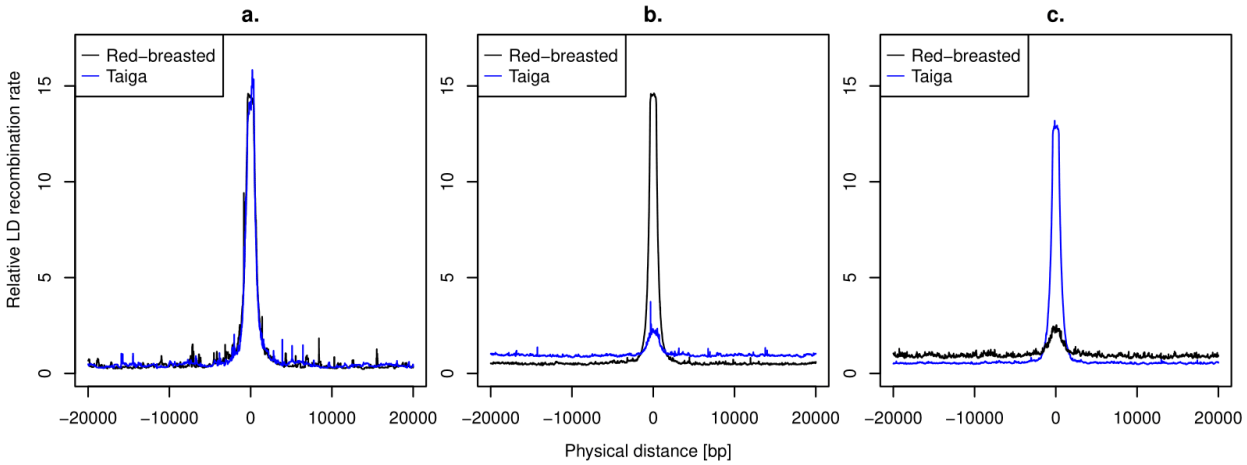

Figure 8: Mean LD recombination rate in 40 kb surrounding shared hotspots between the two species and unique hotspots for each species. The rate is relative to the mean in the 40 kb window. **a.** The relative rate for both species in 40 kb surrounding shared hotspots. **b.** The relative rate for both species in 40 kb surrounding hotspots unique to red-breasted flycatcher. **c.** The relative rate for both species in 40 kb surrounding hotspots unique to taiga flycatcher.

The correlation between the LD recombination rate based on the different phases of scaffold N00100 was lower for red-breasted flycatcher (r = 0.297-0.715, Table 3) than for taiga flycatcher (r = 0.557-0.769, Table 4, Figure 9). The stronger correlation for taiga flycatcher was expected since the phasing was done with 65 individuals compared to 15 individuals for red-breasted flycatcher and an increase in sample size is known to improve the accuracy of the phasing.

Table 3: Pearson correlation coefficients between LD recombination rate based on different phases for scaffold N00100 and red-breasted flycatcher. 95-percentile confidence interval in parentheses, p-value < 2.2e-16 for all correlations.

Red-breasted flycatcher	Phasing 1	Phasing 2	Phasing 3
Phasing 1 with imputation		0.355 (0.347 - 0.362)	0.297 (0.289 - 0.305)
Phasing 2 with imputation			0.715 (0.710 - 0.719)
Phasing 3 without imputation			

Table 4: Pearson correlation coefficients between LD recombination rate based on different phases for scaffold N00100 and taiga flycatchers. 95-percentile confidence interval in parentheses, p-value < 2.2e-16 for all correlations.

Taiga flycatcher	Phasing 1	Phasing 2	Phasing 3
Phasing 1 with imputation		0.677 (0.673 - 0.681)	0.557 (0.552 - 0.561)
Phasing 2 with imputation			0.769 (0.766 - 0.772)
Phasing 3 without imputation			

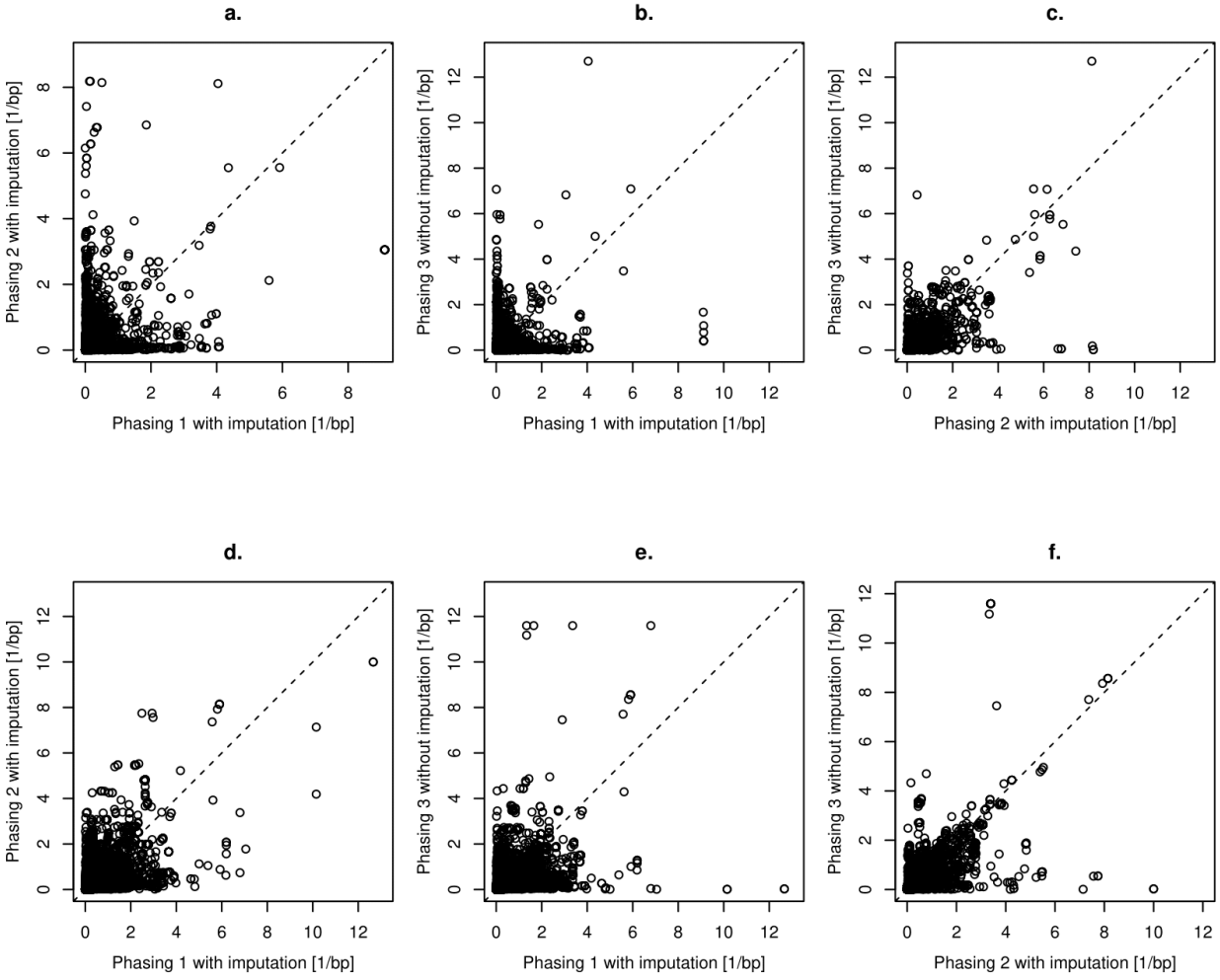

Figure 9: Relationship between the LD recombination rate based on three independent phases for scaffold N00100 for red-breasted flycatcher (**a, b, c**) and taiga flycatcher (**d, e, f**). **a. and d.** Phasing 1 with imputation of missing genotypes and phasing 2 with imputation. **b. and e.** Phasing 1 with imputation and phasing 3 without imputation. **c. and f.** Phasing 2 with imputation and phasing 3 without imputation.

The correlation between the different phases was higher when comparing the mean rate in 200 kb windows (Table 5, Figure 10) than when the rate for SNP pairs were compared. These results indicate that the mean LD recombination rate in 200 kb windows is more resistant to the variation in phasing than the LD recombination rate on the level of SNP pairs.

Table 5: Pearson correlation coefficients between the mean LD recombination rate in 200 kb windows based on different phases for scaffold N00100 and red-breasted flycatcher. 95-percentile confidence interval in parentheses, p-value = 1.34e-8 for all correlations.

Red-breasted flycatcher	Phasing 1	Phasing 2	Phasing 3
Phasing 1 with imputation		0.965 (0.898 - 0.988)	0.965 (0.899 - 0.988)
Phasing 2 with imputation			0.952 (0.865 - 0.984)
Phasing 3 without imputation			

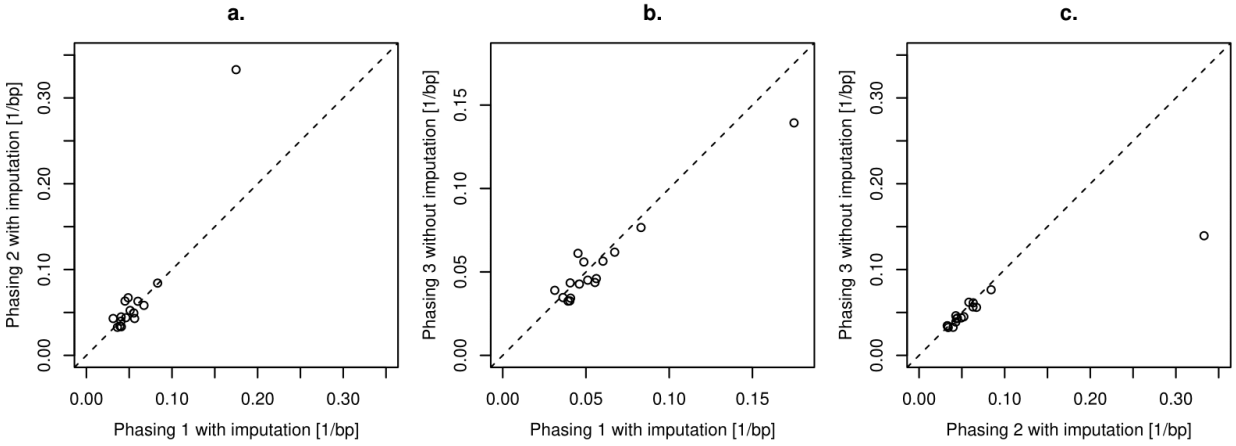

Figure 10: Relationship between the mean LD recombination rate in 200 kb windows based on three independent phases for scaffold N00100 for red-breasted flycatcher. **a.** Phasing 1 with imputation of missing genotypes and phasing 2 with imputation. **b.** Phasing 1 with imputation and phasing 3 without imputation. **c.** Phasing 2 with imputation and phasing 3 without imputation.

4 Discussion

The aim of this project was to characterize the recombination landscape for red-breasted and taiga flycatcher. Firstly, phase was imputed from 15 red-breasted and 65 taiga individuals. Secondly, the recombination rate was estimated using the phase for all red-breasted individuals and the 25 best taiga individuals. From the estimated rate, recombination hotspots were found in each species and shared hotspots were found between species. In 200 kb windows, corresponding to the windows of a linkage map based on collared flycatcher, the mean recombination rate was calculated. Using these windows, the correlation between the species recombination landscapes was calculated.

To investigate the accuracy of the phasing step, two re-phasings of one scaffold was done for both species. The estimated recombination rates, based on different phases, were compared to find the consensus between different runs of the phasing method.

4.1 Conservation of the Recombination Landscape

The number of shared hotspots between red-breasted and taiga flycatchers was 8%, which was low compared to the number of shared hotspots between collared (39%) and pied flycatcher (45%, Kawakami *et al.* 2017). In the study of collared and pied flycatcher, four populations of each species were used and shared hotspots had to be shared between at least one population of collared flycatcher and one population of pied flycatcher. This means that one hotspot had four populations which it could be shared with, increasing the chance of a hotspot being shared.

A study that compared hotspot location in zebra finch (*Taeniopygia guttata*) and long-tailed finch (*Poephila acuticauda*) found that 73% of zebra finch hotspots were shared with long-tailed finch, the expected number of shared hotspots was 4.4%. Their definition of hotspots differed a bit from the definition used here but they had the same definition of shared hotspots.

A factor that is expected to affect the number of shared hotspots is the divergence time between the species pairs. Collared and pied flycatcher has a divergence time of less than 1 million years (Nadachowska-Brzyska *et al.* 2013) and zebra finch and long-tailed finch has a divergence time of about 2.6 - 3.2 million years (Singhal *et al.* 2015), while red-breasted and taiga flycatcher has a divergence time of about 2.5 - 4.2 millions years (Hung & Zink 2014). With the higher divergence time between red-breasted and taiga flycatcher we would expect fewer shared hotspots, but not as few that was found here. Compared to what has been found in other birds, the number of shared hotspots between red-breasted and taiga flycatcher is low.

At the position of hotspots unique to one species, the rate in the other species is increased compared to the background. This has been seen in other flycatchers (Kawakami *et al.* 2017) and is expected only because of how hotspots are defined as some regions will not be classified as a hotspots since the rate is slightly less than 10 times the background. These regions will contribute to the patterns of an increase in the rate for one species at the location of hotspots unique to the other without this being a consistent pattern for all hotspots.

The occurrence of one hotspot in red-breasted flycatcher corresponding to two hotspots in taiga flycatcher, indicate that the definition of hotspots might result in a loss of the signal of conservation. The length of the red-breasted hotspot and the short region separating the two taiga hotspots points to that the two taiga hotspots are in fact one hotspot. The definition of hotspots split the hotspot in taiga flycatcher in two, showing that information might be lost depending on how hotspot are defined.

The correlation of mean LD recombination rate in 200 kb windows between red-breasted and taiga flycatchers was $r = 0.677$ (partial correlation: $r = 0.553$) and the same correlation between collared and pied flycatchers was $r = 0.79$ (Kawakami *et al.* 2017). This, together with the low number of shared hotspots between red-breasted and taiga flycatchers, indicates that the recombination landscape might be less conserved in this species pair than between collared and pied flycatchers. However, the results of the re-phasing questions how reliable the fine-scale estimation of the LD recombination rate is, on which hotspots were found. Since no hotspots were shared between all estimates based on different phases this indicate that the noise introduced by the phasing might affect the detection of hotspots considerably. On the other hand, more hotspots were shared than expected by chance, indicating that the signal in the data is partly robust.

4.2 Implications of the Phasing Method

The higher correlation between 200 kb windows of different phases, compared to the fine-scale correlation, shows that there is a signal in the data that is found using these methods. The low correlation on the fine-scale might be improved by using high-quality data. The SNP density in this data-set was generally high which makes it possible for more stringent filtering without loosing coverage over the genome. If this project was re-done with only high-quality SNPs it would also be possible to investigate how much the quality of the SNPs used affects the estimate.

FastPHASE infers phase based on the assumption that linked variants in the genome cluster together (Scheet & Stephens 2006). Haplotypes on the chromosome scale are inferred by placing the variants that cluster together on the same haplotype and combining the linked regions by finding overlap between them until haplotypes of the length of chromosomes are obtained. Variants cluster together based on LD and the LD varies over the genome, which should mean that the accuracy of the phasing varies with the LD. In regions with strong LD, the clustering should be more consistent giving a better phasing accuracy. Since recombination breaks LD, recombination hotspots have lower LD which could make the phasing harder in these regions. A possibility is that the strength of the recombination is wrongly estimated since the phasing is less accurate in regions with high recombination. One could also argue that if an allele is unlinked the information that allele gives is that it is unlinked. It should therefore not matter if it is assigned to the wrong haplotype, since the unlinked property of the allele makes it occur in random combinations with other alleles over the samples, a problem would be if the phasing introduced a false linkage for that allele. It is hard to know how common these errors are or how the accuracy is affected by LD, unless one would study the phasing method in great detail which requires a detailed knowledge of the statistical methods used. This is outside the scope of this report.

The re-phasing without imputation does not say much in itself about the effect of imputing missing genotypes. To find this effect, one would have to plan an experiment specifically for this and look at more than one re-phase without imputation. The prediction would be that without imputation, LDhelmet has a weaker prediction, which might show as a higher variance between the LDhelmet runs based on the same phase or a higher variance in runs. When using imputation, the imputation should be different between the different phases resulting in a higher variance between runs based on different phases. The variance between runs based on the same phase, when using imputation, should be lower, since the imputation gives a false sense of accuracy of the phase. The safest approach would be to phase without imputation, since guessing in the earlier steps of the process can introduce false signal from noise.

4.3 Further Analysis

For future analysis, more rigorous filtering steps should be applied to ensure higher quality SNPs since this is believed to increase the accuracy of the results. The phasing steps should also be investigated to find the best approach available. It would also be interesting to analyse the position of collared hotspots in contrast to the positions found in red-breasted and taiga flycatchers. A more detailed comparison with collared flycatcher might answer how the evolution of the recombination rate looks on the time-scale between the two species pairs.

Another way to increase the accuracy of the results would be to control for missing data. From a quick look at the coverage over different scaffold in this data-set, it looks like some scaffolds have a high amount of missing data. If these scaffolds were excluded the correlations would probably be more trustworthy.

It would also be interesting to try to find another definition of hotspots. The definition used here uses hard limits and regions with a rate slightly less than 10 times the background will not be classified as a hotspot, which affects the number of shared hotspots found. It might be possible to use different ranges of intensity instead of one hard limit, especially when investigating the rate in one species using the hotspots from the other. If instead of defining shared hotspots as was done here, one could group hotspots of one species according to the rate in the other species to see how many hotspots show a signal in the other species, regardless if the intensity is at least 10 times the background. This would give an indication of how many hotspots seems to share an origin between the two species, saying more about the conservation of the recombination landscape than what can be said now.

These results indicate that there might be interesting patterns between red-breasted and taiga flycatchers, and that the recombination landscape may be less conserved between these two species than in other comparisons of birds. Most importantly, the results in this report highlight the insecurity in statistical phasing and the need for high-quality data and a careful choice of methods.

5 Acknowledgements

I want to thank my supervisor, Madeline Chase, for giving me the tools and material to execute this project. My subject reader, Niclas Backström for giving advice and showing an interest in this project. I also want to thank Hans Ellegren for letting me do my project in his group and Carina Mugal for advice on how to determine the ancestral state and detect hotspots.

References

BirdLife International 2017. Ficedula albicilla (amended version of 2016 assessment). The IUCN Red List of Threatened Species 2017: e.T22734119A119301073. http://dx.doi.org/10.2305/IUCN.UK.2017-3.RLTS.T22734119A119301073.en. Downloaded on 17 October 2019.

BirdLife International 2018. Ficedula parva. The IUCN Red List of Threatened Species 2018: e.T22735909A132037161. http://dx.doi.org/10.2305/IUCN.UK.2018-2.RLTS.T22735909A132037161.en. Downloaded on 17 October 2019.

Browning SR, Browning BL. 2011. Haplotype phasing: Existing methods and new developments. Nature Reviews Genetics 12: 703–714.

Burri R, Nater A, Kawakami T, Mugal CF, Olason PI, Smeds L, Suh A, Dutoit L, Bureš S, Garamszegi LZ, Hogner S, Moreno J, Qvarnström A, Ružić M, Sæther S-A, Sætre G-P, Török J, Ellegren H. 2015. Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of Ficedula flycatchers. Genome Research 25: 1656–1665.

Campbell NA (ed). 2015. Biology: a global approach, 10. ed., Global ed. Pearson, Harlow.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience, doi 10.1186/s13742-015-0047-8.

Cruickshank TE, Hahn MW. 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Molecular Ecology 23: 3133–3157.

Cutter AD, Payseur BA. 2013. Genomic signatures of selection at linked sites: unifying the disparity among species. Nature reviews Genetics 14: 262–274.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R. 2011. The variant call format and VCFtools. Bioinformatics 27: 2156–2158.

de Massy B. 2013. Initiation of Meiotic Recombination: How and Where? Conservation and Specificities Among Eukaryotes. Annual Review of Genetics 47: 563–599.

Ellegren H, Smeds L, Burri R, Olason PI, Backström N, Kawakami T, Künstner A, Mäkinen H, Nadachowska-Brzyska K, Qvarnström A, Uebbing S, Wolf JBW. 2012. The genomic landscape of species divergence in Ficedula flycatchers. Nature 491: 756–760.

Hung C-M, Zink RM. 2014. Distinguishing the effects of selection from demographic history in the genetic variation of two sister passerines based on mitochondrial–nuclear comparison. Heredity 113: 42–51.

Kawakami T, Smeds L, Backström N, Husby A, Qvarnström A, Mugal CF, Olason P, Ellegren H. 2014. A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution. Molecular Ecology 23: 4035–4058.

Kawakami T, Mugal CF, Suh A, Nater A, Burri R, Smeds L, Ellegren H. 2017. Whole-genome patterns of linkage disequilibrium across flycatcher populations clarify the causes and consequences of fine-scale recombination rate variation in birds. Molecular Ecology 26: 4158–4172.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England) 25: 1754–1760.

Mayr E. 1942. Systematics and the origin of species: from the viewpoint of a zoologist. Columbia University Press, New York.

Moyle RG, Hosner PA, Jones AW, Outlaw DC. 2015. Phylogeny and biogeography of Ficedula flycatchers (Aves: Muscicapidae): Novel results from fresh source material. Molecular Phylogenetics and Evolution 82: 87–94.

Mugal CF, Arndt PF, Ellegren H. 2013. Twisted signatures of GC-biased gene conversion embedded in an evolutionary stable karyotype. Molecular Biology and Evolution 30: 1700–1712.

Nadachowska-Brzyska K, Burri R, Olason PI, Kawakami T, Smeds L, Ellegren H. 2013. Demographic Divergence History of Pied Flycatcher and Collared Flycatcher Inferred from Whole-Genome Re-sequencing Data. PLoS Genetics, doi 10.1371/journal.pgen.1003942.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ravinet M, Faria R, Butlin RK, Galindo J, Bierne N, Rafajlović M, Noor M a. F, Mehlig B, Westram AM. 2017. Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. Journal of Evolutionary Biology 30: 1450–1477.

Rettelbach A, Nater A, Ellegren H. 2019. How Linked Selection Shapes the Diversity Landscape in Ficedula Flycatchers. Genetics 212: 277–285.

Scheet P, Stephens M. 2006. A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. American Journal of Human Genetics 78: 629–644.

Singhal S, Leffler EM, Sannareddy K, Turner I, Venn O, Hooper DM, Strand AI, Li Q, Raney B, Balakrishnan CN, Griffith SC, McVean G, Przeworski M. 2015. Stable recombination hotspots in birds. 6.

Smeds L, Mugal CF, Qvarnström A, Ellegren H. 2016. High-Resolution Mapping of Crossover and Non-crossover Recombination Events by Whole-Genome Re-sequencing of an Avian Pedigree. PLoS Genetics, doi 10.1371/journal.pgen.1006044.

Stapley J, Feulner PGD, Johnston SE, Santure AW, Smadja CM. 2017. Variation in recombination frequency and distribution across eukaryotes: patterns and processes. Philosophical Transactions of the Royal Society B: Biological Sciences 372: 20160455.

Wolf JBW, Ellegren H. 2017. Making sense of genomic islands of differentiation in light of speciation. Nature Reviews Genetics 18: 87–100.