From DEPARTMENT OF LABORATORY MEDICINE
Karolinska Institutet, Stockholm, Sweden

ANALYSIS OF GENOMIC DATA AS AN APPROACH TO UNDERSTANDING MIGRATION IN SONG BIRDS

John Boss

Karolinska Institutet

Stockholm 2015

Published by Karolinska Institutet.
Printed by Printed by E-Print AB 2015

ISBN 978-91-7549-820-1X

Analysis of genomic data as an approach to understanding migration in song birds

THESIS FOR Licentiate DEGREE

By

John Boss

Principal Supervisor:
Professor Antony Wright
Karolinska Institutet
Department of Laboratory medicine
Division of Clinical Research Center

Co-supervisor:
Accociate professor Mats Grahn
Södertörns University
Department of School of Natural Sciences,
Technology and Environmental Studies

Examination Board:
Accociate professor Peter Savolainen
KTH Royal Institute of Technology
Department of Scilifelab
Division of Division of gene technology

Professor Sören Nylin
Stockholms Universitet
Department of Zoology
Division of Ecology

Dr Hong Jiao
Karolinska Institutet
Department of Biosciences and Nutrition

ABSTRACT

Many species of birds migrate every year thousands of kilometers, relying on sight, memory, magnetic sensors and instincts to find their way across continents. Many juvenile birds travel complicated migration routes without the guidance of more experienced adults. To successfully accomplish this they need instincts that utilizes multiple navigational senses together with a time dependent schedule. Little is known of the genetics behind migration behavior and which cellular processes are involved. Improved sequencing methods allow us to investigate migration traits from a cellular and genetic perspective. This have given us new insight of the mechanisms of migration and, in time, will let us understand the evolutionary origin of this behavior. In this thesis I focuses on the possibilities of using population genetics to discover the cellular mechanisms involved in migration.

I'm using two subspecies of the small songbird Willow warbler Phylloscopus trochilus to explore the genetics behind the migration behavior. The two subspecies Phylloscopus trochilus trochilus and Phylloscopus trochilus acredula differ significantly in their migration routes while in the same time show few genetic or phenotypic differences. Here I compare genotype and gene expression differences between this subspecies in order to find candidate genes involved in the genetics of migration.

In Paper I we sequence mRNA from brain samples of 16 birds, 8 from each subspecies, using 454-pyrosequencing. We detect three areas of recent selection pressure corresponding to regions in chromosome 1, 3 and 5 on the Zebra finch genome.

In Paper II we compare mRNA expression levels between migrating and breeding birds as well as between the two subspecies. We use a custom microarray probe design based on expressed sequence tags from the Zebra finch to measure the mRNA levels of 22,109 probe sets. We find 14 probe sets with subspecies differences and 3045 that change between the breeding and migrating seasons.

In conclusion, we provide a list of genes and chromosome regions with possible importance for migration and migration behavior. Further studies needs to pair candidate genes with phenotypic differences utilizing laboratory controlled behavior or gene specific sequencing and position tracking

LIST OF SCIENTIFIC PAPERS

I. **Lundberg M, Boss J, Canbäck B, Liedvogel M, Larson KW, Grahn M, Åkesson S, Bensch S, Wright A. 2013**
Characterisation of a transcriptome to find sequence differences between two differentially migrating subspecies of the willow warbler Phylloscopus trochilus. BMC Genomics 14: 33

II. **Gene expression in the brain of a migratory songbird during breeding and migration. manuscript**
John Boss, Miriam Liedvogel, Max Lundberg, Peter Olsson, Nils Reischke, Sara Naurin, Susanne Åkesson, Dennis Hasselquist, Anthony Wright and Mats Grahn and Staffan Bensch

CONTENTS

LIST OF ABBREVIATIONS

ADCYAP1	Adenylate Cyclase Activating Polypeptide 1
DI	Differentiation index
EST	Expressed sequence tag
FDR	False discovery rate
GO	Gene onthology
NGS	Next generation sequencing
PPP3CA	protein phosphatase 3, catalytic subunit, alpha isozyme
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variants
UTR	Untranslated region

1 INTRODUCTION

Next generation sequencing (NGS) has opened up new scientific possibilities by allowing production of genetic and gene expression data for a progressively diminishing investment in time and cost. This trend is making the techniques available to a bigger part of the scientific community and to society at large. Sequencing methods are today used, among other things, to evaluate the health of biomes, identify rare genetic diseases and personalize medical treatment in relation to the patient's own DNA sequence. There are challenges associated with this increase of sequence data and diminishing sequencing cost. The biggest hurdle has become the processing and analysis of data. As the mass of data quickly increases the methodology for handling and analyzing it is still being developed and standardized. Multiple companies provides software solutions with easy to use interfaces and a readymade analysis protocols that processes sequence data and present understandable results. Unfortunately, they cannot be used in all study designs and don't always perform the appropriate algorithms. Bioinformatics experts on the other hand have often little time to inform themselves of the details of the studies they are assisting. It is increasingly important for scientists to have a firm grasp of the bioinformatics field in order to understand the strengths and weaknesses of data.

Migration

Many species of animals have advanced abilities for orientation and navigation. Examples are the migration of Atlantic and Pacific salmon back from the sea to the river where they were born as well as the migration behavior of the Monarch Butterfly, which stretches over generations as it guides them from Northern USA down to Mexico and back again (Zhu et al., 2009). Birds also are experts in global navigation. The feat of pigeons to find their way back to their original nest without local knowledge has long been used by humans. Some bird species regularly navigate over hundreds of miles of open water with exact precision (Horton et al., 2014) or solo navigate their way across continents with incredible accuracy (Willemoes et al., 2014).

Bird Navigation

Birds navigate during migration using multiple senses, memory cues and instincts. The migration routes can be complex with multiple stopover sites and can cover as much as 80,000 km (Egevang et al., 2010) annually. In some cases it is thought that migrants follow the routes of older birds taught for generations or fly alone guided by instincts (Mellone et al., 2011; Willemoes et al., 2014). Birds undertaking their first solo migration rely on direction and time dependent instincts (Berthold, 1991), while adults may have gained an improved geographical sense (Berthold et al., 1992; McKinnon et al., 2014). The directional instinct is inherited as a multifactorial genetic trait (Wiltschko and Wiltschko, 2013). As a consequence, nestlings tend to inherit an intermediary migration route in regards to their parents.

Crossbreeding studies have confirmed this by showing intermediary flight paths and directions for hybrid birds compared to their parents (Delmore and Irwin, 2014). Polymorphisms within the ADCYAP1 gene have been shown to account for 5% of the observed migration-related restlessness in two European Blackcap populations (Mueller et al., 2011). The ADCYAP1 gene produces a secreted protein that activates cyclic adenosine monophosphate (cAMD) levels, which further activates translation of target transcripts (NCBI).

The aim of this project was to use genomics and associated bioinformatics analysis methods as an approach to investigate migration mechanisms in birds. To approach this question we studied genetic polymorphism and differential gene expression in two closely related subspecies of the Willow warbler with different migration behavior.

2 GENOMICS DATA ACQUISITION

DNA Microarray

DNA microarray systems provide relatively fast and inexpensive ways to measure mRNA levels for tens of thousands of genes simultaneously in the same sample (Hacia et al., 1998). The microarray method today competes with mRNA sequencing as the best way to measure cellular RNA levels. Microarrays requires no sample preparation, such as rRNA removal, but don't reach the same level of sensitivity as RNA sequencing methods. Probe sequences designed to represent the different sequences to be detected in samples are attached at different known microarray locations on a glass or silica surface (Zhao et al., 2014). Usually, mRNA sequences in the sample are converted to fluorescently labeled cDNA, which is subsequently denatured and allowed to hybridize to the probe sequences on the array. Since the rate of hybridization is proportional to the concentration of cDNAs, the relative levels of each transcript in different samples can be determined by the strength of their hybridization signal. More details about the principles involved can be found in(Duggan et al., 1999). The designed microarray probes must bind very specifically to cDNA sequences in the sample therefore the hybridization conditions are adjusted so that binding of probe sequences to non-cognate sequences in the sample is minimized. If non-cognate transcripts can bind to probes, or if single nucleotide variants (SNV) make the intended transcript bind with less strength, the resulting signals may be unreliable. Multiple probes are therefore designed for every transcript in order to minimize these problems (Naurin et al., 2008). Still, cross-hybridization of unspecific transcripts and analogue signal strength retrieved makes a general background signal unavoidable and this has to be corrected for during data analysis(Quackenbush, 2001).

Next Generation Sequencing

The new high throughput chip based DNA sequencing technologies have enabled sequencing of hundreds of thousands reads simultaneously at a progressively falling cost. The general improvement against earlier methods is that sequence data is collected by imaging methods, in real time, as the sequence is determined. The classical sequencing methods, called chain-termination or Sanger sequencing, used elongation terminating nucleotides (di-deoxynucleotidetriphosphates) to interrupting elongation at multiple stages along the sequence. By interrupting with different defect base types in 4 samples and then comparing the fragment lengths one can determine the nucleotide sequence (Sanger et al., 1977). This method produce longer and more accurate reads than Next generation sequencing (NGS) methods but takes much longer to produce the same amount of data due to the necessity of determining the lengths of all the termination products (Schuster, 2008).
In 2007 NGS started to drastically decrease sequencing costs. During a few years the until 2012 the average cost of a sequencing a human genome decreased from ten million US dollars to less than ten thousand US dollars (figure 1).

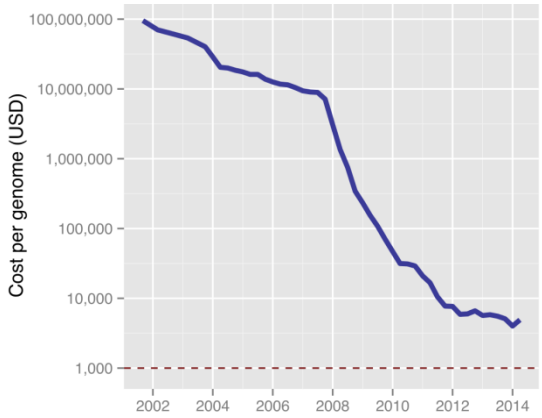

Figure 1: Estimated genome sequencing cost of a human genome between 2002-2014. As calculated by the United states National human genome research institute for a standard of 30x coverage using Illumina methods.

Competition between a small number of companies providing NGS solutions led to several different sequencing methods with slight variations in sequence read number and read accuracy. It is important to select the method that optimally fits each particular study design.

Illuminas dye sequencing

Illuminas dye sequencing is currently one of the used NGS methods. DNA material is cut into fragments of about 150-300bp long and adapters are ligated to both ends of the fragments. The ends of each fragment are then paired to primers attached on a glass slide and amplified by PCR using primers specific for the adapters. As the sequences separate during the denaturation step in each PCR cycle both ends can re-attach and continue to amplify the sequence. This creates groups of identical sequence fragments that can be sequenced using fluorescent dye coupled nucleotides. It is important to adjust the concentration of DNA in the sample to avoid that these groups or "islands" get too close together, which would cause mixed sequence reads, or that they get to far apart and waste sequencing potential. Illuminas dye sequencing currently produces more reads than any other method, In January 2014 Illumina claimed to have reached the 1000 dollar per human genome milestone which could be a crucial step in the field of personalized medicine.

Pyrosequencing

Pyrosequencing produces long sequence reads (up to 700bp). Amplification is performed in a similar way to the Illumina method but the DNA to be sequenced is attached to beads, which are transferred into individual micro-wells in a plate. Nucleotides are added one by one and any nucleotide addition at each position is detected by detecting pyrophosphate that is released as a result. The Pyrophosphate is used as a substrate to produce ATP which in turn

drives the conversion of luceiferin into oxyluciferin and detectable light. The long reads of Pyrosequencing can be very beneficial during sequence assembly but the total amount of data produced per run is much lower than for the Illumina procedure.

pH change sequencing

pH reading senses the specific change different nucleotides bases create in the local ion concentration upon the matching of the next base to a replicated sequence. pH sequencing is supplied commercially by Life Technologies. Library amplification is performed using emulsion PCR (emPCR). DNA fragmens are mixed with breads coated with complimentary primers and PCR reagents. The PCR reactions are isolated from one another using an emulsion of oil and water to form microdroplets, where each microdroplet contains a single bead. Beads containing amplified DNA are then sorted into wells and replicated using regular reagents. The local change of pH created by incorporation of a nucleotide into the fragments is sensed using a semiconductor silicon chip. The pH change sequencing method produces about 200bp long sequence reads with a ca. 1% average error rate. Most errors are deletion or insertion misreads. Latest models of pH change sequencing produce around 1Gb of data in a few hours. The method is relatively inexpensive as it does not require specially modified nucleotides or other reagents and uses inexpensive semiconductor technology for detection but the method produces less data than its competitors.

Paired-end and Mate-pairs

DNA sequence assembly software often fails to utilize reads representing multiple repeats or paralogous genes because there is no unique genomic location to which they can be mapped unambiguously. Since such sequences arise from more than one genomic location they can sometimes be detected as having artificially high read coverage. Paired-end and mate-pair libraries facilitate correct identification and assembly of these regions by linking them to an adjacent unique sequence.

Paired-end sequencing connects reads by ligating unique adapters to the two ends of DNA fragments of approximately known size (<1kb)(Mardis, 2013). After cluster formation using any of the NGS methods the sequence at each end of fragments can be independently determined. When assembling reads, pairs of such sequences can be linked to each other at a distance of the estimated fragment length. Mate-pair sequencing works in a similar way but can be used with longer sequences, up to 20 kb. The longer fragments are first turned into a circular DNA by ligating the fragments end to end with inclusion of a single adapter sequence at the junction. Using unique priming sites in the adapter sequence the two ends of the fragment can then be sequenced independently. Paired-end and mate-pairs sequencing is recommended when performing de novo sequencing of complex genomes or when sequencing multiple repeats. Linking reads is not beneficial for all projects however, as the

more complex library preparation required will decrease the quality of the reads and the quantity of data obtained (Schatz et al., 2010).

Whole-genome sequencing

Whole genome sequencing is used in projects studying genome-wide events or sequencing species for which the genome sequences is not known. Assembling reads for whole genome sequencing can be approached with two main methods, namely reference genome based assembly and de novo assembly. Reference genome based assembly aligns reads against an existing reference genome sequence from the same species or a very closely related species. This decreases the computer memory requirement and calculation time for the assembly software but sequences that can't align to the reference genome will not be assembled. De novo assembly aligns all available reads against each other, this is necessary when no related genome is available. The sequence depth needed depends much on the complexity of the genome, read length and quality. Illuminas dye sequencing and 454 sequencing provides a read quality around 99.9%, giving 1 error per 1000 bases but read quality can vary greatly across genomes (Dohm et al., 2008) and is highly variable, depending on the sequencing method used. Quality is also compromised by PCR amplification by a known C-G nucleotide bias, C-G base rich sequences has been shown to interrupt polymerase replication with an increased frequency compared to A-T rich or mixed regions (Sims et al., 2014). A minimum coverage of 35 reads per base pair is today recommended for reliable identification of single nucleotide variants (SNV) (Sims et al., 2014). Illuminas dye sequencing or pyrosequencing works well for sequencing complex genomes such as mammals (Sims et al., 2014). For previously unsequenced species it is recommended to use pair-end and mate-pair library preparation (Mardis, 2013).

RNAseq

It is estimated that as much as three quarters of the human genome is at some point transcribed (Djebali et al., 2012). Sequencing the transcriptome thus provides much information about the genome and can identify splice variants created during gene transcription (Wang et al., 2009). Transcriptome sequencing limits read data to active sequences of the genome. A problem associated with using transcriptome samples to collect sequence data is that a few highly expressed genes dominate the transcriptome. To achieved an adequately sequence depth over rare transcripts the amount of sequenced data must increase or the transcripts levels has to be equalized. This process is called sample normalization. Normalization preferentially decreases the frequency of abundant mRNA transcripts by a process of re-naturation elimination (Zhulidov et al., 2005). The re-hybridization rate of complementary strands of cDNA in solution is highly dependent on their concentration. By degrading double stranded DNA from the sample, shortly after the re-hybridization process has begun, cDNA sequences can be differentially degraded according to abundance. Repeating this process multiple times in between PCR amplification runs tends

to equalize sequence frequencies and facilitate a more even sequence coverage. One problem when normalizing samples concerns homologous sequences. Transcripts of identical or close to identical sequences will be degraded as though they represent a single unique transcript even though they originate from multiple loci on the genome. This has the effect of artificially lower coverage at such loci.

mRNA sequencing

mRNA sequencing is a way to focus the sequencing on protein-coding genes and thereby increase the read depth for such genomic regions. mRNA is purified through binding of the poly A sequence that marks sequences for translation to immobilized poly T. Filtering for the poly A chain on RNA sequences eliminates rRNA that otherwise can account for as much as 85% of the transcriptome (Morlan et al., 2012). The mRNA sequences are converted into cDNA and sequenced. A drawback using poly A capturing is that the method slightly degrades mRNA and decreases the sample quality. This prevents poly A treatment on sensitive samples that have been formalin fixed or stored imbedded in paraffin (Zhao et al., 2014). Transcriptome sequencing has another drawback of not producing overlapping contigs. As only part of the genome is sequenced, the read created contigs are thus expected to be interrupted due to intergenic regions between genes. This means that contigs cannot be combined into longer scaffolds and the absence of longer genomic sequences prevents chromosome positioning. The lack of positioning data also and makes alternative splicing harder to predict.

RNAseq to measure mRNA levels

The DNA microarray approach is an established method for measuring the relative levels of transcripts in different samples but recently there has been increasing competition from the RNAseq approach. RNAseq has the advantage of not requiring any pre-identification of probe sequences or species-specific chip design and construction. This enables RNAseq to detect novel transcripts that have not previously been described as well as experiments on novel species without extensive genome-sequence information (Sims et al., 2014). This also decreases the background noise as every sequence read can in principle be matched to a specific genomic region, avoiding the cross-hybridization problems that lead to background signals in the DNA microarray method.

However, all sequencing methods have different biases when sequencing mRNA. Most problematic is sequence characteristics that increased the risk of interrupting the polymerase reads . This is of less consequence when sequencing the genome or looking for SNP differences because such data errors can be compensated for by increased read depth. But in RNAseq protocols the quantity of reads is more important than the exact sequence (Mardis, 2013). Problems sequencing C/G rich regions tend to disrupt coverage severely for all methods and also in PCR amplification methods (Sims et al., 2014). Thus G/C rich regions

are generally under-represented in NGS data. pH change sequencing is most sensitive to such problems but both pyrosequencing and Illumina are also less efficient at sequencing G/C rich regions compared to regions of normal base composition (Ross et al., 2013). The length of genes also tend to bias the samples, longer transcripts has a higher probability of artifacts due to interruption during sequencing (Oshlack and Wakefield, 2009).

Sequence read quality

The read quality of NGS sequencers has been lower than for the established Sanger method, both with regard to read length and accuracy. With a sequence error rate of around 1/100 for NGS methods (known as Q29 quality) there will be lots of errors over a genome sequence. This can in many cases be managed by a high read coverage. 30 reads per sequence is considered the gold standard (Hayden, 2014) but as few as 8 reads is most often enough to secure a reasonable accuracy (Sims et al., 2014). In cases were allele frequency is uncertain, for example in mixed samples, high read coverage might not be enough to solve the accuracy problem.

SNP detection

Detecting and analyzing mutations is important when trying to determine evolutionary adaptations. When studying two closely related species, mutation event cannot individually be tracked down to a specific time or population because the original sequence is most often not available. Instead one needs to analyze the frequency of single nucleotide polymorphism (SNP) inside each population and between them. When strong or prolonged selection pressure promotes a rare SNP variant, for example a new mutation, that evolutionary process leaves traces in the genome of that population. As the frequency of that one SNP is increased in the population, SNPs in close proximity to the promoted one will gain the same, or a partial, increase in frequency. This is called a linkage disequilibrium. The effect can be seen as a decrease in SNP variants in an area and it decrease with distance from the promoted SNP. The effect also decrease with time from the event as chromosomal crossover scrambles the sequences variance. When considering population frequencies, all individuals that share a base variant at a single polymorphic positions is referred to as having the same allele variant. To determine the total allele variance between two populations one can calculate the differentiation index (DI). The DI represents the magnitude of allele variation between two populations and is calculated separately for every detected SNP position. 0 is representing no observed statistical difference in allelic frequency and 1 a perfect separation with no shared allels.

Gene ontology terms

To further understand gene expression data one can use Gene ontology methods to categorize groups of genes as belonging to particular biological processes, molecular functions or

cellular components. Gene ontology builds on a categorization of genes as belonging to a hierarchy of categories and sub-categories. Gene ontology can be used to identify gene ontology categories that are significantly enriched in differentially regulated genes (Schlicker et al., 2006). In this way it is possible to interpret lists of differentially regulated genes in terms of the processes, mechanisms and structures to which they contribute. Gene ontology is thus a tool for understanding differences between analyzed samples at the level of structure and function.

the Willow Warbler

In this study we use a small songbird, the Willow Warbler, to investigate the genetics and genomics of migration. The willow warbler (Phylloscopus trochilus) breeds in central and northern Europe as well as throughout northern Asia. There are two subspecies found in Europe, P.trochilus trochilus and the P.trochilus acredula. Phenotypically and genetically the birds are very similar ((Bensch et al., 1999; Bensch et al., 2009; Naurin et al., 2008) except for their wintering areas and migration routes. By using feather isotope measurements and ring recoveries the approximate wintering areas of the willow warbler have been established. P. t. trochilus migrates through western Europe down to western Africa while P.t. acredula travels across eastern Europe down to their wintering habitats in southern Africa (Bensch et al., 2006). They also occupy different breeding areas in Europe. P.t trochilus spreads from Britain through northern and central France to Germany and the southern half of Scandinavia. P.t acredula occupies the northern half of Scandinavia and all the way to Russia. The subspecies meet each other and coexist within hybrid zones through Poland/Lithuania and in the middle of Sweden (figure 2). Genetic markers determine the Swedish hybrid zone to be comparably narrow in regards to what would be expected for two successfully interbreeding populations. This indicates that the hybrid offspring are selected against or that there is a mating bias in favor of within subspecies mating (Chamberlain et al., 2000). One possible theory suggests that offspring might inherent a mixed migration behavior, leading them on an uncompetitive or even lethal flight path. This makes the system is ideal for studying genes involved in migration as there should be few other genetic adaptations between the subpopulations.

Advances into Migration research

Research into how birds migrate has been going on for centuries but only a few genes has so far been suggested to guide the behavior. Limiting factors for investigative migration has been threefold:

1. Tracking birds through migration flight has been hard. Positioning trackers has until just recently been too big and heavy for small species.

2. No existing model organisms such as mice, fruit flies, zebra finch or Chickens exhibits migration behavior. This means that no gene candidates to test and regular methods of gene description can't be relied upon.

3. The high cost of first generation sequencing methods limited quantity and quality of genomic data.

Thanks to modern chip manufacturing and sequencing methods these limitations are no longer a problem. Positioning devices are now getting down to a weight that enables tracking of even small songbirds. By logging sunrise and sunset time the approximate GPS location of a bird can be calculated. More species gets sequenced every year and today many non model organisms have available reference genomes. And decreasing sequencing costs enables whole genome sequencing projects to be founded for a fraction of the cost from a decade ago.

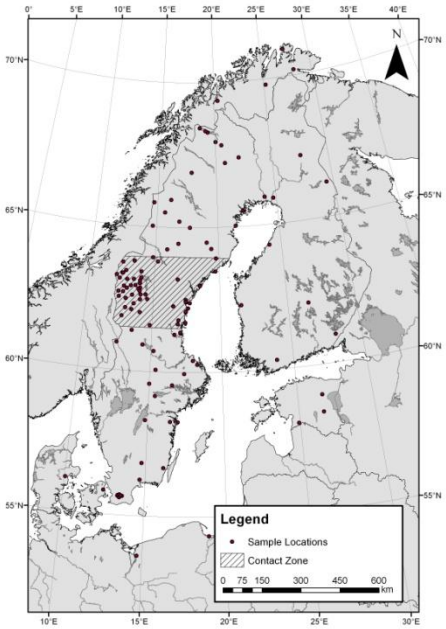

Figure 2: Showing the hybrid zone between the P.trochilus trochilus and the P.trochilus acredula.

3 COMMENTS ON METHODOLOGY, CHAPTER 3

To investigate gene expression and genetic variance with importance for migration we sampled the two subspecies at two different periods during the year. One sampling took place as the birds began to migrate, during early autumn. The other in the middle of summer during the birds' breading season. At each site and period we sampled 8 birds by way of decapitation and extracted the whole brain. The option to separate the brains into their main anatomical regions were considered but rejected in light of the difficulty of performing an accurate dissection in the field.

Paper I

Study design

With 8 samples for every population and season the total amounted to 32 samples. Half of these were selected to be sequenced. A high number of samples is required to estimate SNP and allele frequencies within populations. The number of samples required depends on the number of alleles and their distribution. Under unbiased conditions a read depth of 8 will give a 99.5% certainty of reading two equally expressed alleles. This certainty shrinks with increased allele number, thus requiring more samples, but an high number of samples also increases the amount of genome data to be sequenced. We choose cDNA sequencing rather than whole genome sequencing in order to decrease the amount of targeted sequence and thereby increase read coverage over genes.

Library preparation

We purified polyadenylated mRNA from the samples and transcribed it to cDNA using reverse transcriptase. mRNA levels can vary greatly between genes which results in an overrepresentation of abundant transcripts at the expense of rare transcripts. To provide a more equal coverage of transcripts we choose to have the cDNA normalized. Normalization is a technically complex and expensive treatment and in order to decrease cost the transcripts samples were first pooled together by subspecies into two cDNA library pools, representing birds of different subspecies. This decreased library preparation cost but made SNP detection more difficult. This happens because the sequence contribution of every individual bird cannot be controlled for. We could only partly compensate for this by increasing sequence depth required for SNP detection to a minimum of 8 reads per subspecies library.

Sequencing

With no previous sequence genome published for the Willow Warbler we aimed at performing novo sequence assembly. Difficulties associated with a de novo sequence assembly include correctly calling homogenous genes but avoiding the mistake of detecting

allelic variants as different genes. We sequenced the cDNA librarys using 454 Pyrosequencing due to its long reads, which decrease assembly time and increase assembly accuracy. However the amount of successfully aligned sequences using de novo assembly turned out low, around 50%, compared to aligning direct to the closely related Zebra finch genome (84%). Gene calling using de novo assembly was also evaluated to have been too sensitive with many splice variants being falsely identified as separate genes. In light of these problems we instead aligned our sequences against the Zebra finch genome in an reference based assembly. This worked very well but at the outset it was unknown whether the Zebra finch was a close enough relative to the Willow Warbler to be used as a reference. The Zebra finch has well described genome and also belongs to the order of Passeriformes birds but it lacks behavioral traits associated with long-distance migrant birds (seasonally dependent restlessness and increased appetite)(Liedvogel et al., 2011).

Results

Using reference based assembly we could match 84% of the sequenced reads to the Zebra finch genome. We detected 84,847 SNP variants within 2,469 predicted genes. 55 of these SNPs were fixed between the two populations, giving them a differentiation index (DI) of 1. The SNPs with the clearest difference in frequency between the two subspecies were clustered around three regions in chromosome 1, 3 and 5 on the Zebra finch genome. Out of the highly differentiated SNP positions 14 were tested in other individual samples using Sanger sequencing. Eight of these could be confirmed while the other 6 had low DI score proving them to be false positives. The low SNP calling accuracy is likely caused by the few number birds sampled. A small sampling will never perfectly represent a bigger population, and the huge number of transcripts will multiply the chances of observing apparently fixed SNP frequencies. These problems are made worse by the use of sequence pools since the assumption that all individuals in the pool contribute equally may not always be valid. However the SNP calling accuracy can be improved by using more samples together with deeper sequence depth or by individually sequencing samples.

Paper II

Study design

The Gene expression analysis was performed using Affymetrix microarray chips based on Zebra finch expressed sequence tags (EST). Using this type of cross-species arrays might introduce problems regarding the specificity of the probes and the risk that the model species might lack expression of some genes. Another cross species hybridisation experiment ((Naurin. et. al 2008, 2011)) using the same chip design and phylogenic distance were performed on the Whitethroat (Sylvia communis). This experiment showed strong conservation between species and a high proportion of probes showed hybridization to the

Whitethroat cDNA. The chip design detects approximately 15,800 genes, each gene is detected by 11 different probes with an individual length of 25 bps (Naurin et al., 2008). 29 brain samples were examined, of those 3 were rejected, two samples did not have the expected age or gender and one microarray was marked as and outlier as it varied too greatly in expression levels indicating problems with detection or sample quality. The microarray method were chosen over RNAseq as chip designs were readily available.

Results

Individual expression data allowed us to freely compare sample groups according to our needs: differences within subspecies, between season or combinations of the two. Individual population comparisons, for example: acredula breeding vs acredula migrating, gave few significantly differentiated genes. This compelled us to combine populations from different sample sites and times in the subspecies and seasonal comparisons in order to use all available data. This gave us 3045 seasonally and 14 subspecies differentiated EST. One gene the PPP3CA, changed in regards to both seasons and subspecies. The seasonally changed genes include the previously described ADCYAP1 gene involved in migration restlessness in European blackcap. Although multiple mRNA expression level differences were identified the need to rely on combined population comparisons reveal a lack of samples for statistical purposes and prevented us from effectively identifying subspecies differences during migration. A bigger sample size will be needed in future studies to provide more statistical power.

Future perspective

Dissection of brain samples and gene expression profiling of different brain regions separately would probably give stronger and anatomically more detailed fold change values. Physiological changes occur in specific regions of the brain during migration (Healy et al., 1996) and a focus on the these regions might provide better gene expression data. Other study design biases should also be eliminated as far as possible. There is a critical difference involved in catching birds during migration and breeding seasons. Breading males can easily be lured into portable mist-nets using a recorded song of a challenging male. This method induces a territorial behavior which likely activates its own set of gene responses. The behavior is natural and not unrepresentative of breeding birds but still a direct manipulation of the test subjects. The migrating birds cannot be captured using the same method. Birds don't express territorial aggression wile migrating and have to be captured using other methods. Bird stations facilitate huge immobile nets at natural migration paths to passively capture birds. The best approach would be to also capture breeding birds using passive methods.

4 ACKNOWLEDGEMENTS

I would like give thanks and extend my gratitude to all people who supported me through this work.

My supervicor, Anhony Wright, who with endless patience and great wisdom helped me through this years.

My co-supervisor Mats Grahn who enormous enthusiasm and great ideas makes everything a little easier.

Past and present members of the AWR-group: Helmi, Yongtao, Amir, Gustav and Chiou-Nan. Every day is better with you guys, thanks for all.

My coworkers in Lund: Max, Staffan and Keith.

Inger Porsch-Hällström, for your support and care.

My friends at Södertörn: Kristina, Nasim and Stafan

5 REFERENCES

Bensch, S., G. Bengtsson, and S. Akesson, 2006, Patterns of stable isotope signatures in willow warbler Phylloscopus trochilus feathers collected in Africa: Journal of Avian Biology, v. 37, p. 323-330.

Berthold, P., 1991, Orientation in birds. Spatiotemporal programmes and genetics of orientation: Exs, v. 60, p. 86-105.

Berthold, P., A. J. Helbig, G. Mohr, and U. Querner, 1992, RAPID MICROEVOLUTION OF MIGRATORY BEHAVIOR IN A WILD BIRD SPECIES: Nature, v. 360, p. 668-670.

Chamberlain, C. P., S. Bensch, X. Feng, S. Akesson, and T. Andersson, 2000, Stable isotopes examined across a migratory divide in Scandinavian willow warblers (Phylloscopus trochilus trochilus and Phylloscopus trochilus acredula) reflect their African winter quarters: Proceedings of the Royal Society B-Biological Sciences, v. 267, p. 43-48.

Djebali, S., C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. H. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Roder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, N. S. Bar, P. Batut, K. Bell, I. Bell, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, E. Falconnet, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, O. J. Luo, E. Park, K. Persaud, J. B. Preall, P. Ribeca, B. Risk, D. Robyr, M. Sammeth, L. Schaffer, L. H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. B. Yu, X. A. Ruan, Y. Hayashizaki, J. Harrow, M. Gerstein, T. Hubbard, A. Reymond, S. E. Antonarakis, G. Hannon, M. C. Giddings, Y. J. Ruan, B. Wold, P. Carninci, R. Guigo, and T. R. Gingeras, 2012, Landscape of transcription in human cells: Nature, v. 489, p. 101-108.

Duggan, D. J., M. Bittner, Y. D. Chen, P. Meltzer, and J. M. Trent, 1999, Expression profiling using cDNA microarrays: Nature Genetics, v. 21, p. 10-14.

Hacia, J. G., L. C. Brody, and F. S. Collins, 1998, Applications of DNA chips for genomic analysis: Molecular Psychiatry, v. 3, p. 483-492.

Hayden, E., Check, 2014, Is the $1,000 genome for real?, Nature.

Healy, S. D., E. Gwinner, and J. R. Krebs, 1996, Hippocampal volume in migratory and non-migratory warblers: Effects of age and experience: Behavioural Brain Research, v. 81, p. 61-68.

Horton, T. W., R. O. Bierregaard, P. Zawar-Reza, R. N. Holdaway, and P. Sagar, 2014, Juvenile Osprey Navigation during Trans-Oceanic Migration: PLoS One, v. 9, p. e114557.

Liedvogel, M., S. Akesson, and S. Bensch, 2011, The genetics of migration on the move: Trends in Ecology & Evolution, v. 26, p. 561-569.

Mardis, E. R., 2013, Next-Generation Sequencing Platforms, *in* R. G. Cooks, and J. E. Pemberton, eds., Annual Review of Analytical Chemistry, Vol 6: Annual Review of Analytical Chemistry, v. 6: Palo Alto, Annual Reviews, p. 287-303.

McKinnon, E. A., K. C. Fraser, C. Q. Stanley, and B. J. M. Stutchbury, 2014, Tracking from the Tropics Reveals Behaviour of Juvenile Songbirds on Their First Spring Migration: Plos One, v. 9, p. 9.

Mueller, J. C., F. Pulido, and B. Kempenaers, 2011, Identification of a gene associated with avian migratory behaviour: Proceedings of the Royal Society B-Biological Sciences, v. 278, p. 2848-2856.

Naurin, S., S. Bensch, B. Hansson, T. Johansson, D. F. Clayton, A. S. Albrekt, T. Von Schantz, and D. Hasselquist, 2008, A microarray for large-scale genomic and transcriptional analyses of the zebra finch (Taeniopygia guttata) and other passerines: Molecular Ecology Resources, v. 8, p. 275-281.

Oshlack, A., and M. J. Wakefield, 2009, Transcript length bias in RNA-seq data confounds systems biology: Biology Direct, v. 4, p. 10.

Quackenbush, J., 2001, Computational analysis of microarray data: Nature Reviews Genetics, v. 2, p. 418-427.

Ross, M. G., C. Russ, M. Costello, A. Hollinger, N. J. Lennon, R. Hegarty, C. Nusbaum, and D. B. Jaffe, 2013, Characterizing and measuring bias in sequence data: Genome Biology, v. 14, p. 20.

Sanger, F., S. Nicklen, and A. R. Coulson, 1977, DNA SEQUENCING WITH CHAIN-TERMINATING INHIBITORS: Proceedings of the National Academy of Sciences of the United States of America, v. 74, p. 5463-5467.

Schlicker, A., F. S. Domingues, J. Rahnenfuhrer, and T. Lengauer, 2006, A new measure for functional similarity of gene products based on Gene Ontology: Bmc Bioinformatics, v. 7.

Schuster, S. C., 2008, Next-generation sequencing transforms today's biology: Nature Methods, v. 5, p. 16-18.

Sims, D., I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting, 2014, Sequencing depth and coverage: key considerations in genomic analyses: Nature Reviews Genetics, v. 15, p. 121-132.

Wang, Z., M. Gerstein, and M. Snyder, 2009, RNA-Seq: a revolutionary tool for transcriptomics: Nature Reviews Genetics, v. 10, p. 57-63.

Willemoes, M., R. Strandberg, R. H. G. Klaassen, A. P. Tottrup, Y. Vardanis, P. W. Howey, K. Thorup, M. Wikelski, and T. Alerstam, 2014, Narrow-Front Loop Migration in a Population of the Common Cuckoo Cuculus canorus, as Revealed by Satellite Telemetry: Plos One, v. 9, p. 9.

Wiltschko, R., and W. Wiltschko, 2013, The magnetite-based receptors in the beak of birds and their role in avian navigation: Journal of Comparative Physiology a-Neuroethology Sensory Neural and Behavioral Physiology, v. 199, p. 89-98.

Zhao, W., X. P. He, K. A. Hoadley, J. S. Parker, D. N. Hayes, and C. M. Perou, 2014, Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling: Bmc Genomics, v. 15, p. 11.

Zhu, H. S., R. J. Gegear, A. Casselman, S. Kanginakudru, and S. M. Reppert, 2009, Defining behavioral and molecular differences between summer and migratory monarch butterflies: Bmc Biology, v. 7.

Zhulidov, P. A., E. A. Bogdanova, A. S. Shcheglov, I. A. Shagina, L. L. Wagner, G. L. Khazpekov, V. V. Kozhemyako, S. A. Lukyanov, and D. A. Shagin, 2005, A method for the preparation of normalized cDNA libraries enriched with full-length sequences: Russian Journal of Bioorganic Chemistry, v. 31, p. 170-177.