Clock genes and their role in migratory phenotype among *Passer* sparrows

Caroline Øien Guldvog

Master of Science Thesis

2015

Centre for Ecological and Evolutionary Synthesis
Department of Biosciences
Faculty of Mathematics and Natural Sciences

University of Oslo, Norway

2015

Clock genes and their role in migratory phenotype among *Passer* sparrows

Caroline Øien Guldvog

http://www.duo.uio.no

Print: Reprosentralen, University of Oslo

Table of contents

Acknowledgements

This thesis was written at the Center for Ecological and Evolutionary Synthesis (CEES) at the Department of Biosciences, University of Oslo, under the supervision of Glenn-Peter Sætre and Cassandra Nicole Trier.

Firstly, I want to thank my supervisors. I feel very lucky to have had the two of you guiding me through this process. You have always taken the time to help me and find ways to make my thesis better. Glenn, I am so glad I decided to join your research group! Thank you for providing me with such a cool project. Not only has it been incredibly interesting, I have learnt so many valuable skills that I am very grateful for. To the magical Cassie, what would I have done without you? From day one, you have been a constant source of knowledge and support. Because I had you, I always knew everything was going to be all right. You truly are an inspiration and a role model to me.

To the sparrow group, I am so proud to be part of this amazing group of scientists! Thank you for all the support. Anna, thank you for being such warm and welcoming person and for miscellaneous assistance during this time. Lastly, I want to give a big thanks to Tore. You have been like a third supervisor to me, always so kind, encouraging and solution oriented.

Thank you to my friends and to my fellow residents at the 4th floor study room for most welcome distractions and reassuring talks. Lars, you have provided invaluable encouragement and laughs during all five (!) years we have shared together as students. I'm so glad we have had each other!

My wonderful family, without you I would not have been where I am today. I want to give a very special thanks to Mamma and Maggie, your love and support mean more to me than words can describe.

And to Jan-Erik, who has been there for me for me, holding my hand every day for the past year. I can't believe I finally have someone I can share my weirdness with. Thank you for all the pizza, nerdy conversations, R lessons, for helping me peak out of my shell once in a wile and most of all for all the love.

Abstract

Across the avian lineage, closely related species, or even populations within the same species may display differing migratory phenotypes. The many components of migratory behaviour are inherited as one migratory gene package leaving most, if not all, birds with the necessary machinery to migrate. The appropriate selection pressure can thus cause a switch from migratory to sedentary behaviour over relatively few generations. My study species, the hybrid Italian sparrow and its parental species, the house sparrow and the Spanish sparrow serve as an example of this. The house sparrow and the Italian sparrow are both predominantly sedentary species, while the Spanish sparrow differs by being migratory throughout most of its native habitat. Interestingly, different patterns of migration are observed even within species. Among Spanish sparrows, certain island-living populations have ceased to migrate while a subspecies of the house sparrow living in Central Asia has retained the putatively ancestral migratory phenotype. Although migratory behaviour is known to be under genetic control, less is known regarding which genes are involved. In this study, I attempted to shed light on associations between migratory phenotype and genotype in my focal species by employing a candidate gene approach combined with data from whole genome resequencing. As migration is a seasonal behaviour it is likely to be under the influence of circadian rhythm. I therefore selected 20 candidate genes reported to have a function influencing the circadian clock. Within these genetic regions, I performed a search for fixed differences between migratory and sedentary populations as well as population genetics analyses. However, these tests did not reveal any signs of selection acting on any of the candidate genes. My results are nevertheless consistent with much of the research to date. Any attempts at uncovering the genetics of migration thus far have failed to reveal associations that can be generalized across taxa, highlighting the need for further research.

Introduction

Bird migration is highly diverse, displaying a wide spectrum of phenotypes. Avian populations can differ in many aspects of this behaviour, including their tendency to migrate, timing, duration and migratory route. In addition, studies have shown that this trait can change rapidly over a short evolutionary timescale; migration as well as reversal to sedentary behaviour has evolved repeatedly and independently in many different avian lineages (Joseph *et al.* 2003; Outlaw *et al.* 2003; Davis *et al.* 2006). Together, these findings provide strong evidence for the highly flexible nature of avian migration as well as its ability to respond quickly to selection (Berthold *et al.* 1992; Berthold 1999; Pulido 2007).

A large number of experimental studies, performed primarily on blackcaps (*Sylvia atricipilla)*, have shown that a great deal of the variation observed in migratory traits is due to genetic differences, proving that migration has a strong genetic basis (Berthold & Helbig 1992; Berthold *et al.* 1992; Rolshausen *et al.* 2009; Pulido & Berthold 2010). These findings have led to the hypothesis that the genetic mechanism behind migration is an innate property inherited from a common ancestor. All birds appear to possess the genetic features necessary for migration, their activity being modified depending on migratory state (Berthold 1999; Pulido 2007).

Although we know migration to be genetically based, less is known about which genes are involved. However, as migration is a regular seasonal movement, it can be assumed that the genes involved in regulating this behaviour are related to the ability to respond to periodic changes in the environment. The circadian clock does precisely this. It plays a crucial role in the interpretation of seasonal changes in day length (photoperiod) and by this means, it is a great influencer of behaviour in both plants and animals known to affect behaviours such as flowering in plants, and timing of breeding and migration in animals (Suárez-López *et al.* 2001; Kumar *et al.* 2014). The molecular mechanism behind this circadian oscillator is well known and genetic polymorphisms in circadian clock genes have been associated with variation in behaviour in many organisms

(Tauber & Kyriacou 2005; Johnsen *et al.* 2007; Liedvogel *et al.* 2009; O'Malley *et al.* 2010).

In birds, the circadian clock consists of three independent systems. These clocks are located in the retina of the eyes, the pineal gland and the hypothalamus, each of which consist of an input component, a pacemaker and an output component. For example, the avian pineal gland has photoreceptor cells that detect light, generating circadian oscillations and in turn, rhythmic production and release of melatonin. Interactions between these clocks form a centralized clock system (Kumar *et al.* 2004, 2014). The circadian clock is an endogenous and self-sustained system. However, the circannual clock, controlling timing of annual events, seems to a greater extent to be influenced by environmental cues such as photoperiod. This applies to the timing of migration, however the daily manifestation, known as *Zugunruhe* (nocturnal migratory restlessness), is under control of the circadian clock (Bartell & Gwinner 2005; Rani *et al.* 2006). This biological process controls many components of migratory behaviour, including timing, extent and duration.

Circadian oscillations are generated by interconnected autoregulatory feedback loops. These loops involve transcription of clock genes, followed by translation and activity of clock proteins, each consisting of positive and negative elements. The positive element involves transcriptional activation of clock genes, creating mRNA followed by their translation into clock proteins. These proteins then provide the negative element of the feedback loop by blocking the transcription of the clock genes. This block causes a decline in mRNA, consequently followed by a reduction in clock proteins (Dunlap 1999).

The molecular basis of the circadian clock has been studied extensively in *Drosophila* and mice (Dunlap 1999; Young & Kay 2001), and many orthologues of the mammalian genes have also been discovered in birds (Table 1). Through these studies, important molecular components and workings of the vertebrate circadian clock have been uncovered. The positive limb of the core vertebrate oscillator most importantly consists of the genes *Clock* and *bmal1*. They code for proteins which heterodimerize and form a transcription-acting complex. This complex binds to the promoter regions of the genes of the negative limb, the *period* genes and *cryptochrome* genes, activating their transcription. As the proteins of the negative limb accumulate during the subjective night, they heterodimerize and interact with the CLOCK/BMAL1 complex, effectively blocking their own transcription. Both heterodimerized complexes are successively broken down, and the cycle starts over again. These fluctuations in mRNA and protein concentrations are characteristic of the periodic oscillations of circadian systems, illustrating an important example of how clock genes aid in the timing of circadian behaviours.

The gene *Clock* contains a microsatellite length polymorphism that has been associated with variation in timing of breeding and migration in several species (e.g. monarch butterfly (Merlin *et al.* 2009), Atlantic salmon (O'Malley *et al.* 2014), Chinook salmon (O'Malley *et al.* 2007, 2013; O'Malley & Banks 2008) and Pacific salmon (O'Malley *et al.* 2010)). In avian species, two studies conducted on blue tit (*Cyanistes caeruleus*) revealed some interesting results. One study of 14 blue tit populations spread across Europe and western Asia demonstrated that longer alleles are more prevalent at higher latitudes, possibly reflecting an adaptation to variation in photoperiod (Johnsen *et al.* 2007). Another study following a single blue tit population over two successive breeding seasons showed a trend in female blue tits for individuals with fewer repeats to breed earlier in the season. Moreover, selection for fewer repeats among females was found to be operating by producing a higher number of fledged offspring (Liedvogel *et al.* 2009).

A second gene, *ADCYAP1*, has been associated with migratory restlessness in the European blackcap (*Sylvia atricapilla*)(Mueller *et al.* 2011). This species has been used extensively in the study of avian migration including experiments which have been essential in demonstrating the genetic basis of migratory behaviour (Berthold *et al.* 1990, 1992; Berthold & Helbig 1992; Pulido & Berthold 1998, 2010; Rolshausen *et al.* 2009). Like *Clock*, alleles of the *ADCYAP1* gene differ in length based a repeat sequence. Allele length has been found to consistently associate with migratory restlessness in two independent populations. Long alleles are associated with high migratory activity, both in terms of migratory restlessness and proportion of migrants and migration distance.

The link between migratory behaviour, clock genes and divergence in birds has been further strengthened in a study of genomic divergence in Swainson's thrush (*Catharus ustulatus*) (Ruegg et al. 2014). Two genetically distinct subspecies of this bird differ in traits known to be instrumental in reproductive isolation, including timing of breeding as a consequence of migratory behaviour. Looking at several genes associated with circadian rhythm, some of these were found to be significantly more divergent in the two subspecies than expected by chance. This includes two genes previously linked to migration, *CPNE4* (Jones *et al.* 2008) and *ADCYAP1* (Mueller *et al.* 2011), as well as *CREB1* and *NPAS2*. The last two genes have no previous association with migration, but have been linked to circadian rhythm in other species (Steinmeyer *et al.* 2009). These results give further support to the notion that genes linked to migration can play an important role in divergence among avian species.

The aim of this study is to investigate associations between migratory phenotype and genotype in a study system consisting of three closely related *Passer* sparrows. The house sparrow (*Passer domesticus*) is one of the world's most widely distributed species. Its habitat extends most of Europe, the Mediterranean region and Asia with recent introductions to America, Africa, Australia and New Zealand (Summers-Smith 1988; Anderson 2006). The Spanish sparrow (*Passer hispaniolensis*) is native to the Mediterranean region, southwest and central Asia, sharing most of its breeding range with the house sparrow (Summers-Smith 1988). Their close relation, coupled with sympatric existence has enabled hybridization. Although these occurrences are normally rare, they have resulted in the formation of a reproductively isolated hybrid species, the Italian sparrow (*Passer italiae*) (Elgvin *et al.* 2011; Hermansen *et al.* 2011). This bird is distributed over most of Italy and some surrounding islands, replacing the nearly ubiquitous house sparrow in these areas (Summers-Smith 1988).

Within this system, migratory phenotypes vary both between species and among populations of the same species, making it a great candidate for studying phenotype/genotype association in migratory behaviour. The Spanish sparrow is migratory throughout most of its native environment, but non-migratory island-living populations can be found on Sardinia, the Canary Islands and Malta. The story of the house sparrow is reversed; with the exception of one migratory subspecies (*P.d.*

bactrianus) in Central Asia, the house sparrow is otherwise a resident species. As for the Italian sparrow, it shares the sedentary phenotype of the house sparrow and island-living Spanish sparrows.

These birds not only differ in migratory behaviour, but also in their ecological niches which migration is a reflection of. Sedentary house sparrows are commensal, living in close proximity to human activity (Summers-Smith 1988; Anderson 2006), while the non-commensal *bactrianus* subspecies is migratory and resides in natural grassland (Summers-Smith 1988). Commensalism in house sparrows is a relatively recent invention, following the advent of agriculture in early human civilisation (Summers-Smith 1988; Ericson *et al.* 1997). The *bactrianus* subspecies likely represents a relict population that has maintained the ancestral ecology of the house sparrow (Sætre *et al.* 2012). This indicates that migration is ancestral, while sedentary behaviour is the derived character state.

Migratory phenotype further affects timing of breeding as resident birds start breeding earlier in the season than the migratory birds (A. Gavrilov pers. comm). Avoidance of each other's breeding grounds coupled with differences in timing of breeding creates both spatial and temporal barriers respectively. These discrepancies might constitute premating barriers that can help explain why little interbreeding is observed between sympatric populations of *bactrianus* and other house sparrow subspecies.

As described previously, the great majority of Spanish sparrows are migratory, the only exception being certain island-living populations. A pattern of otherwise migratory birds shifting to sedentary behaviour on islands has been observed among large birds with deferred sexuality (Ferrer *et al.* 2011). However, this trend does not extend to passerines, indicating that the change in phenotype among Spanish sparrows does not necessarily have to do with an effect of islands versus mainland life. An alternative hypothesis could be that the absence of house sparrows on the relevant islands has allowed Spanish sparrows to take over the commensal niche normally occupied by the house sparrow. If the Spanish sparrow's migratory behaviour is connected to their affinity for rural environments as it is for the *bactrianus*, the shift in migratory phenotype could be an example of character release (Grant 1972). Association with human habitation offers a year-round supply of food, making migration an unnecessary cost.

The Italian sparrow, being a hybrid species, could not have arisen without the development of reproductive barriers hindering backcrossing with the parental taxa (Elgvin *et al.* 2011; Hermansen *et al.* 2011). Its sedentary behaviour might constitute a significant premating barrier on the Gargano peninsula. In this area, it lives in sympatry with migrating Spanish sparrow without any signs of interbreeding (Hermansen *et al.* 2011). In this way, selection for differing migration phenotypes, and the genes associated with them, could lead to adaptive divergence and potentially be instrumental in speciation.

The aim of this study was to investigate associations between migratory phenotype and genotype in *Passer* sparrows by employing a candidate gene approach. Candidate clock genes were selected from those reported in other species to have a function affecting circadian rhythm and therefore might be influencing timing of breeding. The candidate gene approach was combined with data from whole genome re-sequencing. My dataset included genomic sequences from both migratory and non-migratory house sparrow and Spanish sparrows, as well as (sedentary) Italian sparrows. For population pairs of con-specific migratory and non-migratory birds, I performed a search for fixed differences as well as an investigation of patterns of divergence. If variations in any of my candidate genes are causal in the change of phenotype, divergence is expected to be relatively high when comparing population pairs. I also performed population genetics analyses based on genetic diversity within populations. This statistic is expected to be lower at loci under the influence of strong selection. Therefore, performing these tests on my candidate genes may uncover possible signs of selection due to adaptive divergence of migratory phenotype.

Materials and methods

The overall aim of this study is twofold: i) to compare candidate circadian clock genes in migratory and sedentary populations of house sparrow and Spanish sparrow; ii) to investigate whether sedentary behaviour has been inherited from house sparrow or Spanish sparrow in the hybrid Italian sparrow.

Identification of candidate genes

Candidate genes

Appropriate candidate genes were identified using two different approaches. Firstly, I searched through available literature collecting examples of genes reported in other organisms to possess functions influencing circadian rhythm. Secondly, I utilized the National Center for Biotechnology Information (NCBI) database on the web (http://www.ncbi.nlm.nih.gov). I queried the Gene database using the key words circadian rhythm coupled with closest available model organism (chicken (*Gallus gallus*) or zebra finch (*Taeniopygia guttata*)). Using these two approaches, I uncovered a total of 20 candidate genes (Table 1).

Non-candidate genes

For use in population genetics analyses, I added a number of randomly selected non-candidate genes to my data set. These were to be used as a basis of comparison with the candidate genes to be able to determine the significance of any potential divergence outliers. I chose one gene for every candidate gene based on them being located on the same scaffold as the given candidate gene.

My procedure for selection of non-candidate genes started with extracting scaffold FASTA files from the house sparrow genome assembly using a custom perl script. Secondly, the zebra finch genome in its entirety was retrieved from Ensembl and used as a reference for blast searches (BLAST+ v2.2.26) with the scaffold FASTA files as query sequences. The region with the highest hit from each blast search was located in the zebra finch genome using Ensembl's search function (Fernández-Suárez & Schuster 2010). From within each of these areas, I randomly selected a gene to represent a non-candidate in my analyses, yielding in total 20 non-candidate genes (Supplementary table 1).

Table 1. List of clock genes included in this study.

Gene	PUBMED ID	Citations
AANAT	NC_011482	(Chong *et al.* 2000; Leder *et al.* 2006; Steinmeyer *et al.* 2009)
ADCYAP1 (=PACAP)	NC_006089	(Steinmeyer *et al.* 2009; Mueller *et al.* 2011)
BHLHE40	NC_011476	(Honma *et al.* 2002)
Bmal1/ARNTL/MOP3	NC_006092	(Dunlap 1999; King & Takahashi 2000; Okano *et al.* 2001; Yasuo *et al.* 2003; Cassone & Westneat 2012)
Bmal2/ARNTL2	NC_006088	(Okano *et al.* 2001)
Clock	NC_011467	(Dunlap 1999; King & Takahashi 2000; Yoshimura *et al.* 2000; Yasuo *et al.* 2003; Leder *et al.* 2006; Johnsen *et al.* 2007; O'Malley & Banks 2008; Liedvogel *et al.* 2009; O'Malley *et al.* 2010; Mueller *et al.* 2011; Cassone & Westneat 2012)
CKI δ/CSNK1D	NC_006127.3	(Steinmeyer *et al.* 2009)
CKI ε/CSNK1E	NC_011463	(Steinmeyer *et al.* 2009)
CPNE4	NC_011465	(Jones *et al.* 2008; Ruegg *et al.* 2014)
CREB1	NC_011471	(Steinmeyer *et al.* 2009; Mueller *et al.* 2011)
Cry1	NC_011463	(Kume *et al.* 1999; King & Takahashi 2000; Yasuo *et al.* 2003)
Cry2	NC_011469	(Kume *et al.* 1999; King & Takahashi 2000; Yasuo *et al.* 2003)
CSNK1A1	NC_011477.1	(Steinmeyer *et al.* 2009)
KIAA1737 (=CIPC)	NC_011469	(Zhao *et al.* 2007)
MTNR1A	NC_011467	(Ebisawa *et al.* 1999)
NFIL3/ E4bp4	NC_011493	(Yasuo *et al.* 2003)
NPAS2/MOP4	NC_011462	(Steinmeyer *et al.* 2009; Mueller *et al.* 2011)
Per2	NC_006096	(Dunlap 1999; King & Takahashi 2000; Yoshimura *et al.* 2000; Yasuo *et al.* 2003; Cassone & Westneat 2012)
Per 3	NC_011485	(King & Takahashi 2000; Yasuo *et al.* 2003; Cassone & Westneat 2012)
Timeless	NW_003778991	(Dunlap 1999; King & Takahashi 2000)

Sampling

Blood samples were collected from a total of 52 sparrows including all three *Passer* species, as well as different populations. Italian sparrows (N=10) were sampled in Guglionesi, Italy; house sparrows in northern Norway (*P. domesticus domesticus*, N=13); Golestan and Kardeh in Iran (*P. domesticus bactrianus,* N=9) and Spanish sparrows from Lesina on the Gargano peninsula in Italy (N=10) as well as the island of Sardinia (N=10). All populations were visited in springtime; Lesina and Guglionesi in 2008, Kardeh in 2009, Golestan in 2010 and Sardinia and northern Norway in 2013. The birds were caught using mist nets and blood samples were extracted by puncturing the brachial vein. 20-50 μl of blood was extracted from each bird and stored in 1 ml of Queen lysis buffer. For catching and sampling, permissions were obtained from the appropriate authorities in all locations. DNA isolation was performed using the Qiagen DNeasy Blood and Tissue Kits (Valencia, CA) according to the manufacturer's protocol. The only adjustments include adding 180 μl of blood/buffer mixture in the initial step as well as incubating the samples overnight at 56 C° after step 1.

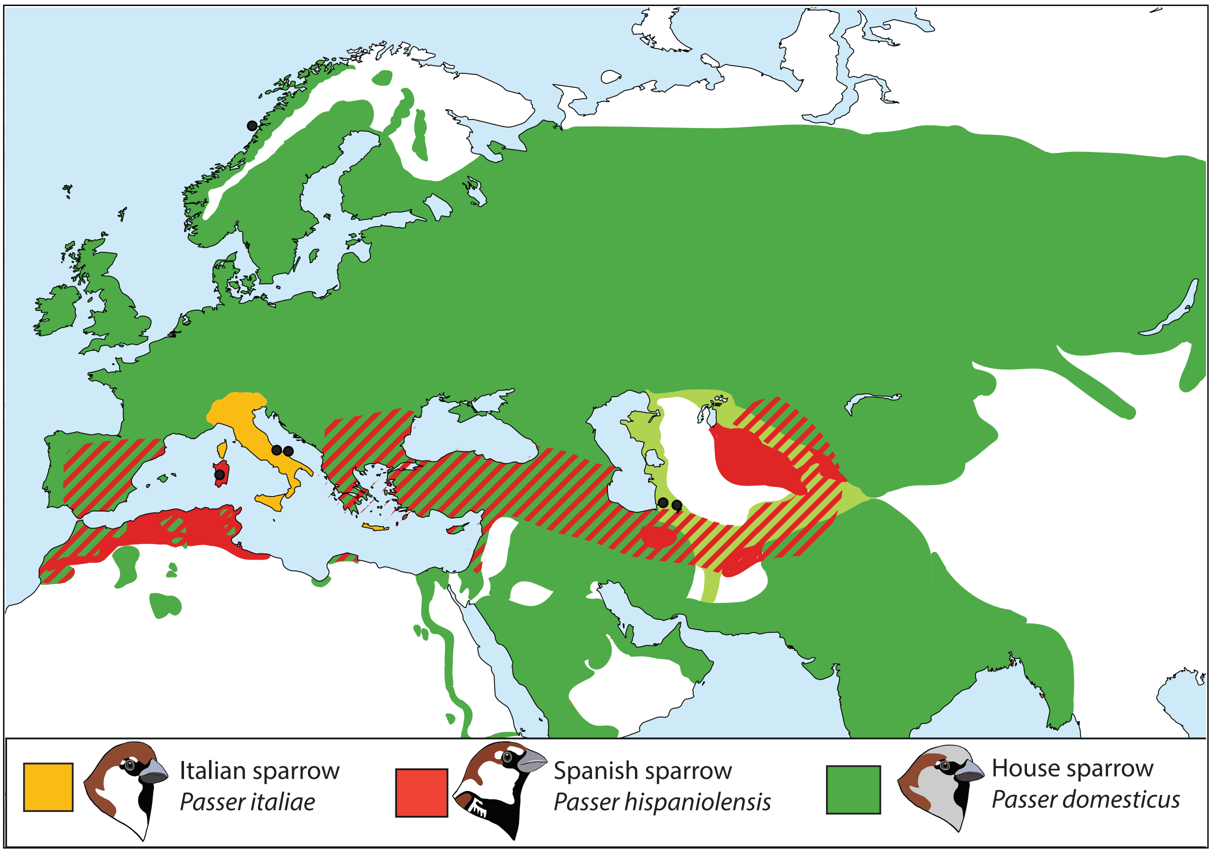


Figure 1. Distribution map of the *Passer* study system in Europe, Asia and North Africa: house sparrow (green), Spanish sparrow (red) and Italian sparrow (yellow). The black dots indicate sampling locations of house sparrows in northern Norway as well as Kardeh and Golestan in Iran; Spanish sparrows on the Gargano peninsula, Italy and the island of Sardiania as well as Italian sparrow in Guglionesi, Italy.

Variant calling pipeline

Pre-processing of resequencing data

The genomes of the 52 sparrows involved in this study were resequenced using Illumina 330-350 bp pair end libraries with coverage of ~8-13X at McGill University Innovation Centre in Montreal Canada.

Using this re-sequencing data, I performed a variant calling pipeline starting with aligning the raw sequence reads to a reference genome. This step is necessary in order to determine the location of each of these reads in the genome. As the reference genome for this process, I employed a genome assembly generated by (Elgvin & Trier *et al.* in prep.) using the DNA of a female house sparrow from Nordland, Norway. Sequencing of this house sparrow genome was performed on an Illumina platform with 180 bp pair end; 3, 8 and 10kb mate pair libraries. An ALLPATHS-LG (Gnerre *et al.* 2011) strategy was used for the assembly. The draft genome utilized has ~130X coverage and a size of 1.05Gbp.

Raw sequence data for each individual bird was mapped to the genome assembly using BWA v0.7.5a (Li & Durbin 2009). Default settings were applied, with the minor exception of the addition of an –M parameter, making the files compatible with Picard for downstream analyses. Following mapping, a coordinate header was added to the files using Picard tools v1.72 (http://broadinstitute.github.io/picard).

Read duplicates are created during the sequencing process when certain pieces of DNA are sequenced multiple times. As they are uninformative and should not be used as evidence during variant calling, it is important to identify and mask these before continuing with further analyses. I labeled these PCR duplicates among my mapped sequencing reads using Picard tools v1.72. Default settings were applied, with the exception of setting validation stringency to being equal to lenient.

The preceding steps are likely to have fabricated artifacts, which falsely appear to be SNPs, due to misalignments. This is particularly true of reads mapped towards the ends of indels. The first step in correcting this is to identify areas in need of realignment. Subsequently, an optimal consensus sequence is found by comparing a best alternate consensus alignment to the original alignment. Local realignment has then identified the most parsimonious placement of the reads relative to the indel, and the actual realignment is performed. I performed local realignment using GATK v3.1.1 (McKenna *et al.* 2010) IndelRealigner with default settings.

Variant calling

Next up in a variant calling pipeline is the process of variant discovery. This step entails determining sites where the sample conflicts with the reference as well as calculating genotypes. I initially used two different modules to achieve this: Unified Genotyper from GATK v3.1.1 (DePristo *et al.* 2011; Van der Auwera *et al.* 2013) and mpileup from SAMtools v0.1.19 (Li *et al.* 2009). My reason for using two variant callers was to increase the chance of identifying true SNPs; a variant recognized by both callers is more likely to be correct than if called only once. In both cases, I applied default settings and the individuals of a given population were called together. Following preliminary data analyses, the data set created using SAMtools mpileup was discarded from future analyses due to technical difficulties.

The GATK v3.2.2 HaplotypeCaller was used to variant call a third time for use in population genetics analyses. The HaplotypeCaller is preferable over the Unified Genotyper or mpileup as it is more accurate and does a better job at calling indels. This program could not be employed in the earlier rounds because at the time it was too computationally demanding. However, updates to the software cut down run time substantially, making it possible to adopt at a later stage to increase the accuracy of my data. Samples were called individually and subsequently merged and genotyped using GATK v3.2.2 GenotypeGVCFs.

Following variant calling, I implemented a round of hard filtering on the data sets created using the Unified Genotyper as well as the HaplotypeCaller. These callers have a high sensitivity when calling SNPs, and are therefore likely to call a fair amount of false positives. Filtering was performed using VCFtools v0.1.12b (Danecek *et al.* 2011). I filtered based on read coverage, which meant omitting every SNP with coverage of less than 5 reads. The re-sequenced genomes all had a read depth of >8x, but filtering at this depth left behind a worryingly low amount of SNPs. Consequently, I chose to filter based on a read depth of 5 to increase the stringency of variant calling, yet lower the concern of disregarding valuable data. In addition, all data sets had SNPs with mapping or genotype quality under 30 filtered out. For VCF files resulting from UnifiedGenotyper, I also removed variants with a minor allele frequency under 20%, whereas for files created using HaplotypeCaller, I excluded sites with more than 0.7 missing data.

Candidate genes in Passer sparrows

Determination of location

The first step in locating my candidate genes in the genomes of my focal species was to determine their structure in the birds' closest possible relative. The exonic sequences of each gene were retrieved from GenBank (Benson *et al.* 2013). In most cases, the appropriate information was available in the zebra finch, but in instances where it was not, I used data from the chicken. For the gene *Period 1*, the sequence employed was retrieved from the house mouse (*Mus musculus*), as no homologous genes has been discovered in any avian genome.

The relevant exonic sequences were compiled into one FASTA file for each gene. These files were used as query sequences in a BLAST v2.2.26 search against the house sparrow genome assembly. The generated e-value for each hit was used as a measure of significance. A value of e^{-10} was used as threshold (the same was used for all other BLAST searches); lower values were seen as reliable hits. As nearly all genes contained several exons, each blast generated multiple hits. This was used as a second measure of hit significance. Consequently, I chose to assign the location of the gene in the house sparrow genome to the genetic region of the highest scoring hit found among multiple exons. In the case of *Period 1,* no hit scored above my designated threshold. In addition, the exons hit different scaffolds; there was no consensus among the highest hits as to the location of the gene. All together, this made it unlikely that *Period1* had actually been identified, and it was consequently excluded from further analyses.

For use in population genetics analyses, I also needed to determine the location of my candidate and non-candidate genes in the data set created by GATK HaplotypeCaller. This time, exonic sequences from the closest available relative was fetched from Ensembl (Fernández-Suárez & Schuster 2010). The resulting FASTA files were blasted

(BLAST+ v2.2.26) against the house sparrow genome assembly to determine their location in my focal species. I used start of highest hit for the first exon and end of highest hit for the last exon as assumed position for each given gene.

Extraction of candidate genes

My search for fixed differences among sedentary and migratory populations of birds required a FASTA file for each genetic sequence in all individual genomes. The first step in realizing this was to convert VCF files resulting from variant calling to individual consensus FASTA files. Files generated using GATK UnifiedGenotyper and SAMtools mpileup, were converted to FASTA using GATK v3.2.2 FastaAlternateReferenceMaker and SAMtools v1.0 respectively.

From each FASTA file containing one of the 52 birds' resequenced genomes, I extracted all 20 genes using SAMtools v1.0. This was done twice; one extraction for each dataset resulting from the two variant callers. Assumed starting point of a given gene was retrieved from the start location specified by the top hit from the BLAST search. Gene length in closest available relative was retrieved from NCBI. This value was added/subtracted (depending on the orientation of the gene in the house sparrow genome assembly) to starting position, specifying the end of extraction.

Ensembl's (http://www.ensembl.org/Multi/Tools/Blast, Birney *et al*., 2004) and NCBI's BLAST tools (http://blast.ncbi.nlm.nih.gov/Blast.cgi) were used to verify the identity of the extracted genes. All sequences were found to hit to avian versions of the gene (or areas within the gene) in question, confirming that the correct genomic sequences had been extracted. However, one possible error in the extraction might be the exact stop and starting positions. It is possible that the ends of the genes were cut either too short, or too long. This could have resulted in genetic areas either being missing or I could have included parts of the genome that does not in reality belong to the gene in question.

Data analyses

Fixed differences between migratory and sedentary populations

In searching for fixed differences between migratory and sedentary populations, I performed three rounds of analyses for each gene: (i) comparing Spanish sparrow from Lesina (migratory) to Spanish sparrow from Sardinia (sedentary), (ii) comparing house sparrow of the *P. d. domesticus* subspecies (sedentary) to house sparrow of the *P. d. bactrianus* subspecies (migratory) and, lastly, (iii) comparing Italian sparrow (sedentary) to both of the sedentary parental populations (*P. d. domesticus* and Sardinian Spanish sparrow). Comparisons between migratory and sedentary conspecifics were performed in order to uncover potential causative SNPs behind the alternative phenotypes. Italian sparrows were compared to sedentary populations of their parental species in the hopes of uncovering from which of these it has inherited its sedentary phenotype.

In the first round of investigations, referred to as preliminary analyses, I explored all 20 candidate genes twice; once for each genetic sequence stemming from variant called data by UnifiedGenotyper and mpileup. In this round, I uncovered a small number of genes displaying fixed differences between certain population pairs. I later followed up these analyses by taking a second look at these particular genes of interest using only filtered data from Unified Genotyper.

Alignments and detection of SNPs/variants were made using Geneious v7.1.7 (http://www.geneious.com, Kearse *et al*., 2012). For each gene, one alignment was created for each of the three populations of *P. d. bactrianus*, Sardinian and Italian sparrows. Alignments were performed using MUSCLE with default settings applied. The respective consensus sequences resulting from the alignments were used as reference sequences, to which the individuals of the corresponding population were mapped. All SNPs or variations between each pair of populations were detected using the 'Find variations/SNPs…' option. When I found a SNP to be fixed in one population relative to the consensus sequence of the corresponding population, the query population's individual sequences would all need to posses this SNP/variation in order to be classified as a fixed. In such a case, the individuals of the population from which the consensus was created would be investigated in order to determine whether both populations were fixed for alternate SNPs/variations.

The method used so far for detection of variation is only suitable in the case of point mutations. However, for some of my candidate genes where polymorphic regions have previously been discovered in other avian species, the allelic differences consist of a variation in length caused by different copy numbers of repeat elements. This applies to the genes *Clock*, *ADCYAP1*, *CREB1* and *NPAS2* (Steinmeyer *et al.* 2009; Mueller *et al.* 2011; Peterson *et al.* 2013). For *ADCYAP1* and *CREB1* these repeat elements are situated in the 3'UTR. I have not been able to include UTRs in my genetic sequences, as they are not incorporated in the annotated version of either the zebra finch or chicken genome. However, for *NPAS2* and *Clock* the region containing the length polymorphisms in question is exon 20, meaning I was able to investigate the relevant areas. To locate exon 20 in my sparrow individuals, I first aligned the entire putative gene from one of them to the annotated version in zebra finch (downloaded from: http://www.ncbi.nlm.nih.gov). Using SAMtools v1.0, the area corresponding to exon 20 in zebra finch was extracted from all 52 individual FASTA files created based on filtered data from GATK UnifiedGenotyper. To locate the repeat regions in question, the exonic sequences were imported into Geneious v7.17 where I performed the same process of aligning and comparing sequences as previously described for detection of point mutation variation.

Population genetics

Estimates of population genetic parameters were calculated to look for signs of selection and patterns of divergence associated with my candidate genes. F_{st} (Weir & Cockerham 1984), Tajima's D (Tajima 1989) and nucleotide diversity (π) (Nei & Li 1979) were computed using VCFtools v0.1.12b. I performed tests for both candidate and non-candidate genes, and in the case of F_{st} and nucleotide diversity, for all scaffolds where one or more candidate gene(s) were located. Tests were done on a windowed or per-site basis, generating multiple values which where later averaged using a custom python script to give one number per gene/scaffold. The gene *Timeless* was excluded from population genetics analyses due to low quality sequence information at this location.

F_{st} was calculated to give an estimate measure of genetic divergence between taxa (Weir & Cockerham 1984). I performed this test pairwise between populations using the same population pairs as in the search for fixed differences between migratory and sedentary populations. Tests were performed for each gene and for the scaffold on which the gene was present. The genes and scaffolds were individually extracted to their own VCF files from the whole genome VCF file using GATK v3.2.2 SelectVariants. Values were calculated on a windowed basis. For scaffolds, I used a window size of 100 kb with overlapping steps of 25 kb. For genes, I adjusted the window size to 10 bp with overlapping steps of 5 bp. Choice of window size was made based on a compromise between the desire to keep window small enough to ensure good resolution, while avoiding large amounts of windows lacking SNPs altogether. Calculating F_{st} for genetic regions should give an indication of differentiation between migratory and sedentary sparrows within these areas. Adding scaffolds to these tests was done in order to give an idea of how divergent the putative gene is in context of the larger genomic region where it is situated.

Tajima's D was implemented in the attempt to separate sequences evolving according to the neutral theory model, from those under the influence of natural selection (Tajima 1989). A negative D value could indicate positive selection (causing a selective sweep), whereas a positive value could point towards balancing selection shaping the region in question. Other non-random demographic processes, such as population expansion or contraction, may also affect the value of Tajima's D. I ran this test for all five sparrow populations based on VCF files created using VCFtools v0.1.12b. For this test, values were calculated using non-overlapping windows with a size of 150 bp based on the average size of exons. I calculated Tajima's D for each gene, but not for scaffolds. The large differences in size of coding and non-coding regions, coupled with their different modes of evolution would have made these values relatively uninformative. The most appropriate way of calculating Tajima's D would have been based on exonic regions alone, but calculations based on whole gene sequences may also be informative.

Nucleotide diversity (π) is used to quantify the extent of polymorphism within a population (Nei & Li 1979). Here, low π values can be used to verify any potential outlier genes from the previous tests, as this is consistent with an area being under the influence of selection. However, low values could also stem from a gene being situated in a region of low diversity such as near a centromere. Consequently, nucleotide diversity was estimated both for genes as well as scaffolds to be able to examine the surrounding area of the gene. Within population tests were run using the same VCF files as mentioned previously. For scaffolds, I used sliding window sizes of 100 kb with overlapping steps of 25 kb. For genes, I calculated π on a site-by-site basis. Calculating nucleotide diversity on a per site basis gave informative computations, so in order to achieve as good of a resolution as possible, I saw no need to increase window size to that used for F_{st} or Tajima's D.

Finally, potential candidate loci under natural selection were identified using F_{st}-outlier analysis with BayeScan v2.1 (Foll & Gaggiotti 2008), a software that uses a Bayesian approach for detecting non-neutral loci. Analyses were performed for all SNPs located within a candidate or non-candidate gene. For this purpose, I created one VCF file for each population pair where each file contained the genetic sequences for all individuals of the respective population pairs. To generate these files, I started with the gene VCF files containing all individuals created using GATK SelectVariants. For each gene, I used VCFtools to include only the individuals of each given population pair as well as excluding sites with more than 0.9 missing data. Finally, all genes were merged using GATK CombineVariants in order to create one VCF file per population pair. Conversion to GESTE/BayeScan format was performed using PGDSpider v2.0.8.1 (Lischer & Excoffier 2012).

Results

Fixed differences between migratory and sedentary populations

Among my candidate genes, no fixed differences between migratory and sedentary populations of house sparrow were revealed. Likewise, I could not detect any fixed differences between the migratory and sedentary population of Spanish sparrow, or between Italian sparrow and any of the sedentary representatives of its parental species. As neither sedentary house sparrow nor Spanish sparrow was characterized by any distinct genotypes, I was not able to assess from which of these the Italian sparrow has inherited its sedentary phenotype.

The genes *Clock* and *NPAS2* were in addition evaluated for variation in repeat length mutations. In *Clock*, I was able to identify the relevant poly-Q repeat element (Steinmeyer *et al.* 2009), but detected no variation among populations. The trinucleotide element was repeated 11 times in all individuals. However, the reported repeat element found by Steinmeyer *et al.* (2009) in *NPAS2* appeared not be present in my data.

Population genetics

F_{st}-tests were performed as a measure of divergence between populations. The values calculated when looking at my candidate genes (Table 2) revealed in general low levels of differentiation. As grounds of comparison, I executed the same test for the non-candidate genes (Table 3) as well as for each scaffold containing one or more candidate genes (Supplementary table 2). The resulting values for both of these were overall lower than those found for the candidate genes. In relation to these low numbers, a small selection of candidate genes displayed comparatively high levels of divergence between certain population pairs. Standout genes include *bmal2* and *NFIL3* comparing sedentary and migratory Spanish sparrows as well as *CSNK1D* comparing the two house sparrow subspecies. In order to visualize divergence in candidate genes in the context of the scaffold where they are located, I produced plots outlining F_{st}-values scaffold wide for all population pairs. Most of these can be found in Supplementary figure 1 except the plots including the standout genes, which can be seen in Figure 3.

Table 2. F_{st}-values for all candidate genes in four population pairs of *Passer* sparrows.

Gene	House sparrow: sedentary - migratory	Italian sparrow - sedentary house sparrow	Italian sparrow - sedentary Spanish sparrow	Spanish sparrow: sedentary - migratory
AANAT	0.0202639	0.0300124	-0.0227147	-0.00331638
ADCYAP1	0.150149	0.0147082	-0.0238878	-0.0183404
BHLHE40	0.0320913	0.0104301	0.0113505	-0.0120498
Bma1	0.0402195	0.00756043	-0.0150665	0.0678491
Bmal2	0.00887357	0.387614	-0.0602276	0.249847
Clock	0.0274332	0.0172684	-0.0160303	0.056523
CREB1	0.0168264	0.000519789	0.000207421	0.0256779
CSNK1A1	0.0344921	0.00264335	0.000198332	0.0115074
CSNK1D	0.204361	-0.00354186	-0.0449297	0.0
CSNK1E	0.0344804	0.0238245	-0.0261513	-0.0253905
NFIL3	nan	0.38077	0.043216	0.212768
NPAS2	0.0511552	0.00637329	0.000602987	0.0314213
Per2	0.0394631	0.337232	-0.0146811	0.0316667
Per3	0.0623653	0.02999	-0.00186797	-0.00990179
Cry1	0.0500097	0.0183836	0.000857106	0.0264704
Cry2	0.0638519	0.00330194	-0.00425636	-0.00267074
KIAA1737	0.0559113	-0.000469675	0.0257169	-0.0337889
MTNR1A	0.163322	0.024096	-0.109635	0.0709536
CPNE4	0.0736491	0.016707	-0.00072662	0.0150999

Table 3. F_{st}-values for all non-candidate genes in four population pairs of *Passer* sparrows.

Gene	House sparrow: sedentary - migratory	Italian sparrow - sedentary house sparrow	Italian sparrow - sedentary Spanish sparrow	Spanish sparrow: sedentary - migratory
AGO2	0.0652865	0.0669575	0.025434	0.00250538
HELZ	0.0290919	0.011652	-0.0143113	0.0116918
RHOT1	0.032872	0.0290672	-0.00404266	0.00770681
MXD4	0.0342831	0.0235923	-0.000815782	0.00460663
ZNF521	0.0566608	0.0092038	0.0017907	0.0530931
SLC39A10	0.0346218	0.0124393	-0.00365314	-0.00132021
PCM1	0.0413128	0.0246142	0.000856177	0.0994383
LTK	0.0610393	0.0286487	-0.00193013	-0.00414776
FOXP1	0.0520556	0.023752	0.000246742	0.00395765
SYK	0.051578	0.0283605	0.0627809	0.183874
ORAOV1	0.0505964	0.0491682	-0.012942	-0.00491546
NCF4	0.0491401	0.036034	0.01191	0.00760987
TOM1	0.0768838	0.0234685	-0.00833788	0.0241197
MACC1	0.0166526	0.0337139	0.00517527	0.163147
GALNT10	0.0316323	0.026301	0.00848025	0.00195342
AGK	0.0848931	0.0431043	-0.00602907	0.0265014
PLCH2	0.0255802	0.0197797	-0.00827888	0.00536999
UGGT1	0.0488937	0.00504513	0.0111346	0.0173915
MRPS9	0.0846598	0.030501	-0.0136404	0.0631965

Finally, I employed an F_{st}-outlier test using BayeScan to assess the significance of the standout values. The results of this analysis revealed no significant F_{st}-outliers (Figure 2) providing no evidence for selection at work on the candidate genes. Results from Tajima's D as well as nucleotide diversity tests confirm the neutrality of my candidate genes; all values were close to zero with candidate genes differing little from non-candidate genes (Table 4 and 5 for candidate genes, Supplementary tables 3 and 4 for non-candidates). Scaffold wide nucleotide diversity plots for the standout genes can be seen in Figure 3 and for all other genes in Supplementary figure 2.

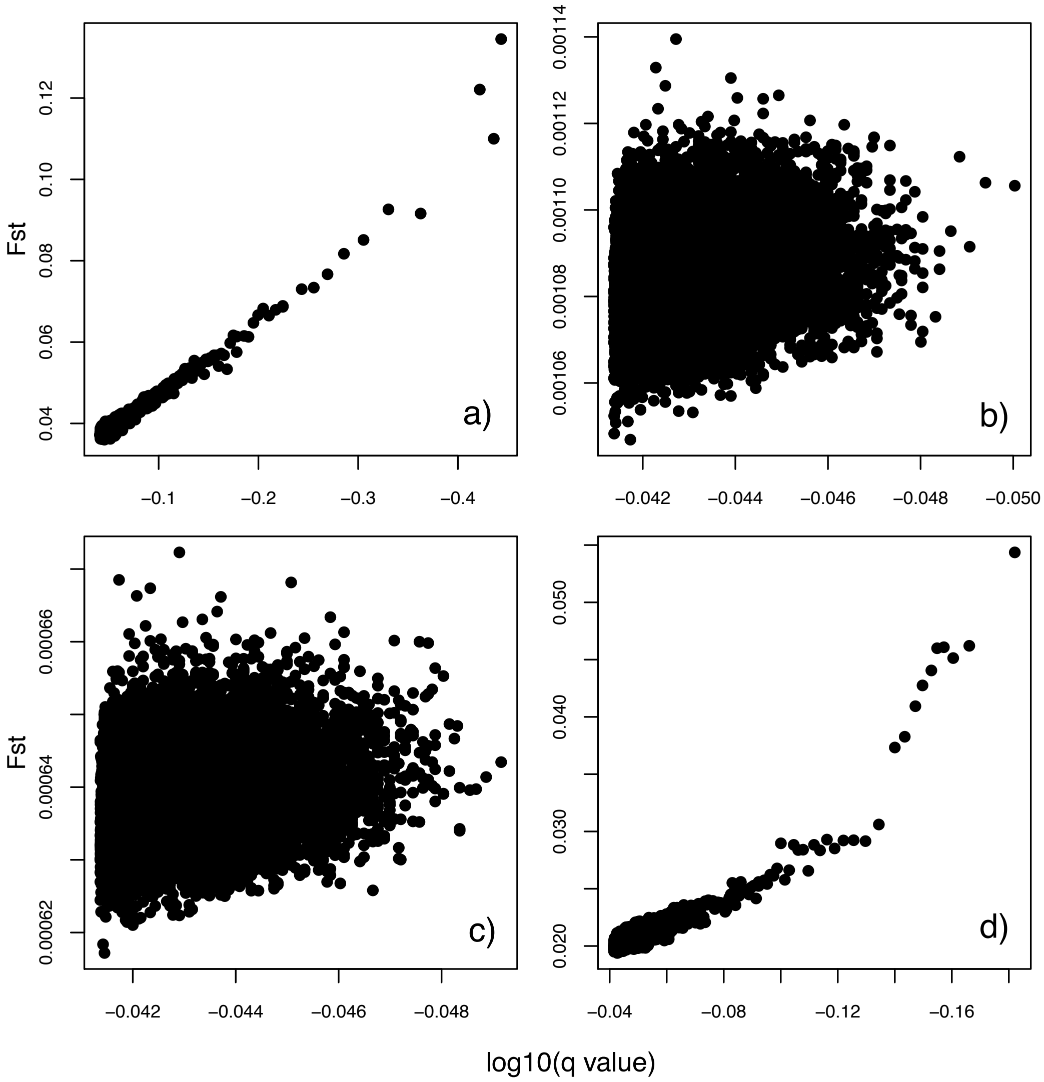


Figure 2. F_{st} results obtained from the BayeScan outlier test. Each point corresponds to one SNP located within a genetic area, either candidate or non-candidate. The y-axis presents F_{st}-values for four populations pairs: a) Migratory versus sedentary house sparrows, b) sedentary house sparrows versus Italian sparrows, c) sedentary Spanish sparrows versus Italian sparrows and d) migratory versus sedentary Spanish sparrows. The x-axis presents the q-value of a given locus, i.e. the minimum FDR value at which it would have become significant (FDR = 0.05). No outlier markers are identified, as these would have been marked with the SNP number.

Table 4. Nucleotide diversity (π) for candidate genes in all populations.

Gene	Migratory house sparrow	Sedentary house sparrow	Italian sparrow	Migratory Spanish sparrow	Sedentary Spanish sparrow
AANAT	0.441027	0.601552	0.475571	0.409461	0.488921
ADCYAP1	0.359423	0.651963	0.470791	0.48591	0.517088
BHLHE40	0.518492	0.533938	0.480796	0.426048	0.480298
Bmal1	0.360792	0.440141	0.434023	0.220364	0.406189
Bmal2	0.0277337	0.0632159	0.483702	0.0862594	0.462817
Clock	0.386455	0.407134	0.42267	0.351903	0.614174
CREB1	0.431729	0.462896	0.469597	0.439835	0.544763
CSNK1A1	0.429306	0.431733	0.448151	0.407568	0.491344
CSNK1D	0.325	0.460634	0.512546	0.667619	0.722222
CSNK1E	0.404583	0.502733	0.415014	0.33202	0.403428
NFIL3	0.0	0.0	0.559524	0.161616	0.487879
NPAS2	0.408839	0.467426	0.453822	0.411162	0.499566
Period2	0.197143	0.189371	0.472975	0.353085	0.49883
Period3	0.385331	0.440751	0.384348	0.369135	0.448117
Cry1	0.39225	0.419554	0.431567	0.35538	0.489115
Cry2	0.377027	0.504857	0.439573	0.397347	0.462156
KIAA1737	0.402303	0.41946	0.40535	0.409879	0.369191
MTNR1A	0.297271	0.482623	0.472249	0.303611	0.65539
CPNE4	0.361659	0.363637	0.334051	0.273686	0.361461

Table 5. Tajima's D for candidate genes in all populations.

Gene	Migratory house sparrow	Sedentary house sparrow	Italian sparrow	Migratory Spanish sparrow	Sedentary Spanish sparrow
AANAT	0.985045	2.01854	1.37002	0.996305	1.64432
ADCYAP1	0.183597	0.493104	0.385325	0.39143	0.427319
BHLHE40	0.6977	0.787489	0.746539	0.446434	0.615188
Bmal1	0.315935	0.553194	0.450486	0.15567	0.417267
Bmal2	0.0	-0.00506534	0.137999	0.00393445	0.105582
Clock	0.694849	1.00792	0.999826	0.863353	1.127
CREB1	0.528726	0.800067	0.66164	0.637101	0.756385
CSNK1A1	0.546855	0.70601	0.657973	0.61575	0.712199
CSNK1D	0.0	0.0	0.0	0.0	0.0
CSNK1E	0.590672	0.888534	0.58086	0.504869	0.558988
NFIL3	nan	nan	0.0	0.0	0.0
NPAS2	0.526659	0.786576	0.696275	0.602009	0.824114
Period2	0.0557176	0.0239242	0.240055	0.0782913	0.260096
Period3	0.540892	0.769552	0.534048	0.520889	0.661802
Cry1	0.382509	0.591476	0.488271	0.335075	0.629439
Cry2	0.31852	0.618312	0.498705	0.458804	0.529685
KIAA1737	0.338088	0.683138	0.636321	0.632788	0.480192
MTNR1A	0.0619864	0.618376	0.470115	0.136906	0.560074
CPNE4	0.592456	0.756338	0.618317	0.579098	0.747994

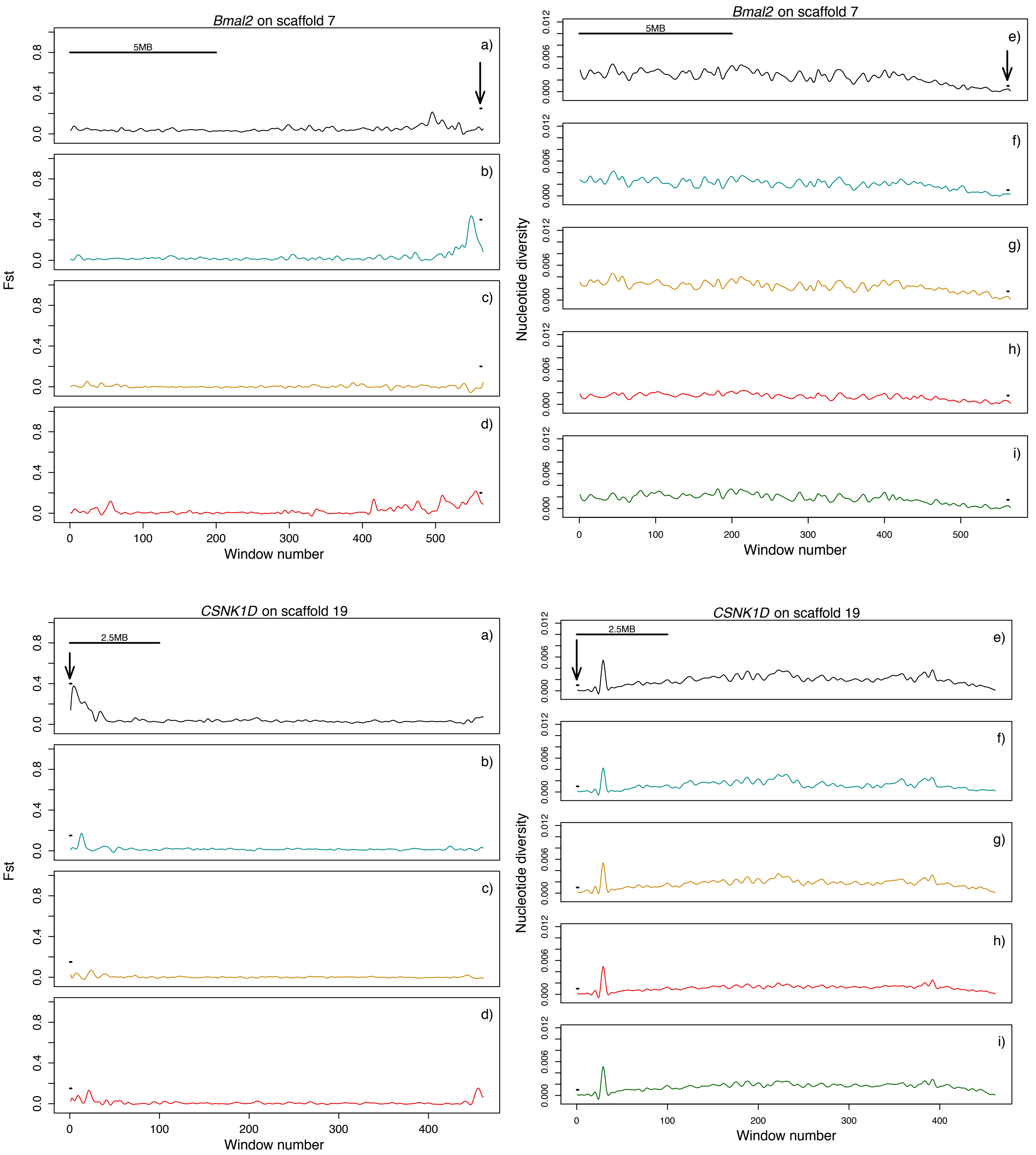


Figure 3. Scaffold wide F_{st}-values and nucleotide diversity (π) for those containing a candidate gene with a standout value. The y-axis represents F_{st}-values or nucleotide diversity respectively. The x-axis represents window number as both statistics were calculated on a windowed basis with a window size of 100kb in overlapping steps of 25kb. Arrow and point mark the location of the relevant candidate gene. F_{st}-values (plots a-d) were estimated between the following population pairs from top to bottom: a) migratory versus sedentary house sparrows, b) Italian sparrows versus sedentary house sparrows, c) Italian sparrows versus sedentary Spanish sparrows and d) migratory versus sedentary Spanish sparrows. Nucleotide diversity (plots e-i) was calculated for all populations: e) migratory house sparrows, f) sedentary house sparrows, g) Italian sparrows, h) sedentary Spanish sparrows and i) migratory Spanish sparrows.

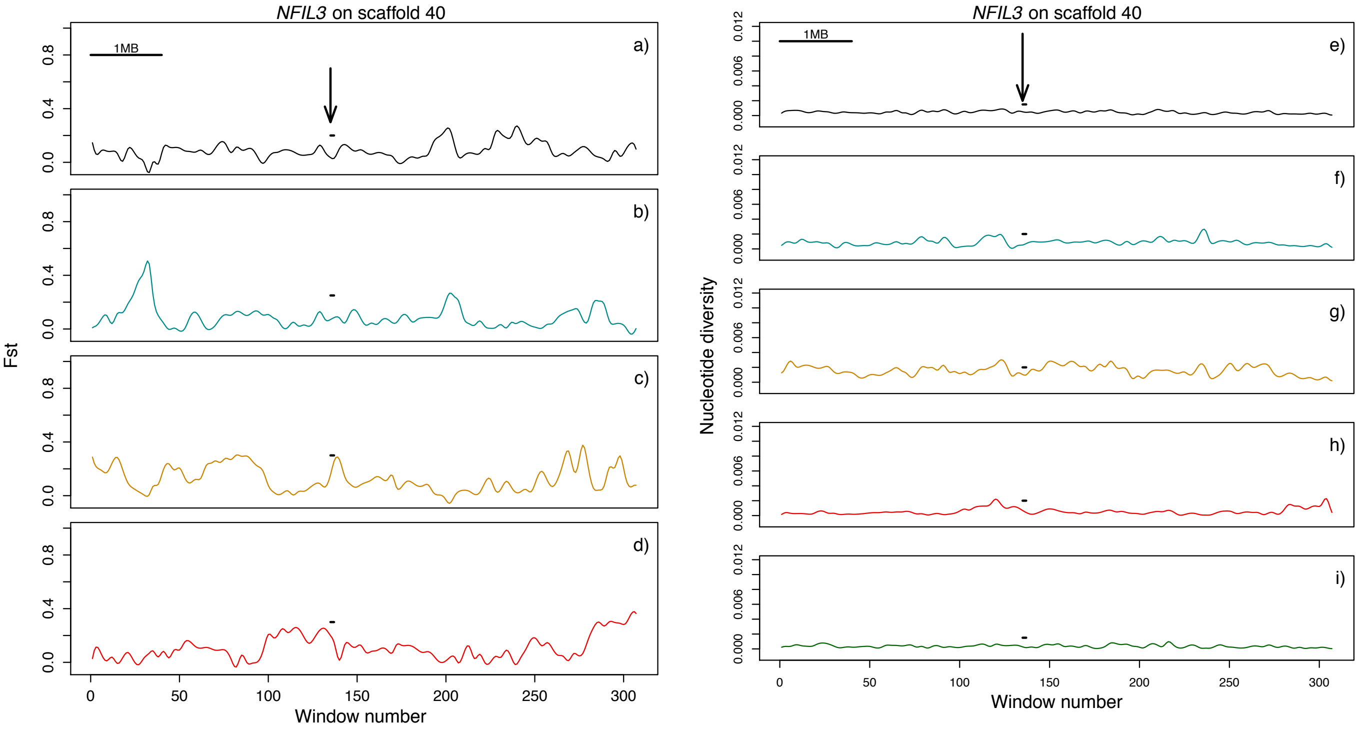


Figure 3. Continued.

Discussion

Avian migration exhibits large amounts of variation both across, and within taxa. Although a large degree of this diversity is genetically based (Berthold *et al.* 1992; Rolshausen *et al.* 2009; Pulido & Berthold 2010), the underlying genetic mechanism is largely unknown.

With this study, I have attempted to shed light on the genetics behind differences in migratory behaviour among three non-model species of the genus *Passer*. I have investigated associations between migratory phenotype and genotype by employing a candidate gene approach. I selected genes based on their putative role in circadian rhythm and procured these genetic sequences in my focal species by using data from whole genome re-sequencing. Within my study system, differences in migratory phenotype exist both between and within species. However, my results were overall negative. No fixed differences were detected between migratory and sedentary populations of birds. Likewise, population genetics analyses revealed no clear signs of selection. Although F_{st}-values for certain candidate genes seemed particularly high for subspecies/population comparisons, the F_{st}-outlier test did not identify any loci as non-neutral. Furthermore, the standout values were neither mirrored in levels of nucleotide diversity, nor by Tajima's D.

The lack of positive findings is consistent with much of the available research on genes involved in migratory and breeding behaviour. The gene *Clock* is among the most extensively investigated genes in this respect. *Clock* has been connected to migratory and breeding behaviour in blue tits (Liedvogel *et al.* 2009) and barn swallows (Caprioli *et al.* 2012), but failed to be so in a number of other avian species including blue throats (Johnsen *et al.* 2007), black caps (Mueller *et al.* 2011), great tits (Liedvogel & Sheldon 2010), Swainson's thrush (Ruegg *et al.* 2014) and *Tachycineta* swallows (Dor *et al.* 2012). *ADCYAP1* is another gene associated with migratory behaviour in black caps (Mueller *et al.* 2011) and Swainson's thrush (Ruegg *et al.* 2014) that together with *Clock* has been linked to different aspects of migratory behaviour in a study of the avian genus *Junco*. However, the associations were inconsistent across different species of the genus

(Peterson *et al.* 2013). All together, these findings suggest the previously discovered relationships between genetic variation and migratory behaviour cannot be extrapolated to a general pattern across taxa.

There are several reasons why genetic associations to migratory behaviour identified in other species, may not be present in my focal species. The circadian clock can be affected by numerous environmental factors of which most research to date has focused on photoperiod. However, other non-photic environmental influences, including temperature, social cues and food availability, have proven to be instrumental in adjusting the timing of the central clock (Ramenofsky *et al.* 2008; Visser *et al.* 2010). Genetic differences between species or populations may therefore be reflective of the different cues they use to adjust their circadian clock. For example, *Clock* may be associated with photoperiod. Selection for this gene could consequently vary depending on whether photic or non-photic cues are essential in influencing the circadian behaviour. However, as migration is a complex trait, many cues may be instrumental in all species. As a result, different genotypes could emerge depending on a host of influences. Another alternative is that the switches between migratory and sedentary behaviour in my study system may not be related to circadian rhythm at all. If this is the case, we may need to search other yet unknown genes. Among house sparrows, the change in migratory phenotype may be related to food availability as sedentary behaviour has been selected for following the recent adaptation to commensalism (Sætre *et al.* 2012). I hypothesised that this could be the case for non-migratory Spanish sparrow as well. Thus, the lack of between-population variation in *Clock* and other potentially informative genes could reflect adaptations responding to different cues than those affecting other examined species.

Assuming the relevant genes shaping migratory phenotype have been included in this study, any causative loci may not be possible to identify with the methods I have employed. Migration entails changes in movement, orientation, feeding, fattening, and sleeping patterns. This complexity is likely to be modulated by many genes, each of which contribute a small effect to the change (Pulido 2007). In addition, phenotypic expression is probably influenced by both epistatic effects and gene-environment interactions (Ayroles *et al.* 2009). For example, even though polymorphisms in

ADCYAP1 was found to associate with migratory behaviour in black caps, a large portion of additive genetic variance was still left unexplained indicating that several additional unidentified loci contribute to the phenotype (Mueller *et al.* 2011). If each gene has such a small effect, more direct methods of estimating associations, such as gene expression studies, may be required in order to recognize their contribution. Expression patterns can vary seasonally, throughout the day and among different organs (Yoshimura *et al.* 2000). By this means, differential expression patterns of clock genes among migratory and sedentary sparrows could be associated with behavioural differences and may provide insight into other, as yet unidentified relationships.

Genomic areas essential to the modification of migratory behaviour may be missing in my study. I have attempted to include all exonic and intronic regions of my candidate genes, but there is no guarantee that I have been able to incorporate these in their entirety. Furthermore, an essential genetic component missing from my data set are the regulatory regions. Variations in regulatory regions of genes, such as UTRs, can be instrumental in shaping the phenotype by modifying the expression of a gene (Ayoubi & Van De Ven 1996). To date, only two polymorphisms have been associated with migratory behaviour, one of which is located in the 3’ UTR of *ADCYAP1* (Mueller *et al.* 2011; Peterson *et al.* 2013). Additionally, a potentially informative polymorphism has been located in the 3’ UTR of *CREB1,* further indicating the importance of these regions (Steinmeyer *et al.* 2009). As regulatory regions have not been included in my research, I am unfortunately unable to test the generality of these findings in my focal species.

A few of the potentially informative polymorphisms identified in avian clock genes have contained repeat elements where di- or trinucleotides are repeated a varying number of times creating different allele lengths (Steinmeyer *et al.* 2009). Of these, I only investigated the potential length polymorphisms of *Clock* and *NPAS2*. The repeat elements of the other genes are located in UTRs, which as previously explained, I have been unable to include in my data set. In *NPAS2* the putative repeat element appears not to be present in my focal species, as I was unable to locate it within the relevant genetic region. However, I was able to locate the relevant repetitive region of *Clock*, but found no variation in allele length. The absence of variation in repeat elements of *Clock* could be a reflection of natural selection shaping this locus by favouring a certain repeat length. Alternatively, the locus in question could be selectively neutral with a present day length inherited from a common ancestral sparrow. However, repetitive DNA has a high mutation rate, as slippage during DNA replication is very common in these areas (Wren *et al.* 2000). This means that low variation within such a region is unlikely to be maintained in the absence of stabilizing selection. Consequently, I count the notion that neutral processes could explain absence of variation as extremely unlikely.

To the best of my knowledge, the repetitive region of *Clock* is polymorphic in all other birds where this locus has been studied (Fidler & Gwinner 2003; Johnsen *et al.* 2007; Steinmeyer *et al.* 2009, 2012; Liedvogel *et al.* 2009, 2012; Liedvogel & Sheldon 2010; Dor *et al.* 2011, 2012; Mueller *et al.* 2011; Caprioli *et al.* 2012; Saino *et al.* 2013; Peterson *et al.* 2013). For this reason, the presence of only a single allele in all 54 sparrows included in my study makes me question whether this is in fact a true representation of the genetic variation available. The genomic sequences I used for my analyses are products of next-generation sequencing. These methods have short read lengths and produce enormous volumes of data, making repetitive sequences difficult to resolve. One problem is that repeat regions may be larger than read length, making them difficult to interpret even with the use of mate-pair reads. Further, reads mapping to multiple locations can cause ambiguities in both *de novo* assembly and mapping to a reference. For my data, uncertainties arise both from any mis-assemblies in the reference as well as by the low depth of coverage of re-sequencing data limiting the chance of properly covering the repeat area. Near identical tandem repeats, such as those found in the repetitive regions of *Clock* and *NPAS2*, are often collapsed into fewer repeats due to the difficulty in assigning the true copy number. Length polymorphisms that went undetected through the variant calling pipeline might therefore exist in any of my candidate genes.

In addition, the sample size I used is markedly smaller than for other such studies (see for example (Dor *et al.* 2011; Caprioli *et al.* 2012)). A potential drawback of this is that it may limit my ability to detect available genetic variation. However, most comparable studies investigate continuous traits whereas the sparrows included in my study display a discrete phenotype; all individuals of a given population either migrate or stay resident. Consequently, the number of individuals included here should be sufficient to identify broad scale patterns. However, any smaller trends and shifts in allele frequencies could be missing from my study.

Potential outliers, such as the loci located within the genes with standout F_{st}-values, may have been drowned out due to the scale at which I conducted my analyses. I performed population genetics tests using sliding windows on both exonic and intronic regions together as well as across scaffolds. Analyses restricted to exonic sequences such as the Hudson-Kreitman-Aguadé (HKA) test (Hudson *et al.* 1987), the McDonald-Kreitman test (McDonald & Kreitman 1991) and Tajima's D (Tajima 1989), could have provided more trustworthy results by eliminating the confounding contribution of introns present in my study.

Finally, the use of a candidate gene approach in searching for associations between migratory phenotype and genotype is greatly challenged by lack of known migration linked genes. The candidate gene approach is based on genetic structure and function exhibiting a high degree of conservation across lineages. Genes identified in one species as influencing the expression of the relevant phenotype, might therefore have a similar effect in other organisms (Tabor *et al.* 2002; Fitzpatrick *et al.* 2005; Piertney & Webster 2010). Such appropriate candidate genes are unknown for migratory behaviour, as none have ever been directly identified in a model species. However, as migration is a seasonal behaviour, studies addressing this issue have looked at genes associated with circadian rhythm; a biological mechanism aiding in interpretation of daily and seasonal variation. In this way, circadian rhythm is a candidate trait for migration. The absence of any clear candidate genes for migration is a weakness in this pursuit reducing the likelihood of positive results.

Conclusion

I was unable to detect any signs of selection at work on my candidate clock genes among migratory and non-migratory populations of house sparrow, Spanish sparrow and Italian sparrow. This is in keeping with much of the relevant research to date, as no genes have been found to consistently associate with migratory behaviour. A major challenge in this endeavour is the lack of clear candidate genes for migratory behaviour. Furthermore, the absence of general findings is likely related to this complex trait being regulated by a host of different genes, each of which contribute a small effect to the overall phenotype. However, further analyses such as expanding sequence data to include regulatory regions of candidate genes, performing analyses on a finer scale as well as gene expression studies all have the potential of revealing relationships I was unable to identify with the current methods employed.

References

Anderson T (2006) *Biology of the ubiquitous House Sparrow from genes to population*. Oxford University Press, Oxford.

Van der Auwera GA, Carneiro MO, Hartl C *et al.* (2013) From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, 1–33.

Ayoubi TA, Van De Ven WJ (1996) Regulation of gene expression by alternative promoters. *The FASEB journal*, **10**, 453–460.

Ayroles JF, Carbone MA, Stone EA *et al.* (2009) Systems genetics of complex traits in *Drosophila melanogaster*. *Nature genetics*, **41**, 299–307.

Bartell PA, Gwinner E (2005) A separate circadian oscillator controls nocturnal migratory restlessness in the songbird *Sylvia borin*. *Journal of biological rhythms*, **20**, 538–549.

Benson DA, Cavanaugh M, Clark K *et al.* (2013) GenBank. *Nucleic Acids Research*, **41**, D36–D42.

Berthold P (1999) A comprehensive theory for the evolution, control and adaptability of avian migration. *Ostrich*, **70**, 1–11.

Berthold P, Helbig A (1992) The genetics of bird migration: stimulus, timing, and direction. *Ibis*, **134**, 35–40.

Berthold P, Helbig A, Mohr G, Querner U (1992) Rapid microevolution of migratory behaviour in a wild bird species. *Nature*, **360**, 668–670.

Berthold P, Wiltschko W, Miltenberger H, Querner U (1990) Genetic transmission of migratory behavior into a nonmigratory bird population. *Experientia*, **46**, 107–108.

Birney E, Andrews TD, Bevan P *et al.* (2004) An Overview of Ensembl. *Genome Research*, **14**, 925–928.

Caprioli M, Ambrosini R, Boncoraglio G *et al.* (2012) Clock Gene Variation Is Associated with Breeding Phenology and Maybe under Directional Selection in the Migratory Barn Swallow. *PLoS ONE*, **7**, e35140.

Cassone VM, Westneat DF (2012) The bird of time: cognition and the avian biological clock. *Frontiers in Molecular Neuroscience*, **5**, 1–8.

Chong NW, Bernard M, Klein DC (2000) Characterization of the chicken serotonin N-acetyltransferase gene: Activation via clock gene heterodimer/E box interaction. *The Journal of biological chemistry*, **275**, 32991–32998.

Danecek P, Auton A, Abecasis G *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Davis LA, Roalson EH, Cornell KL, Mcclanahan KD, Webster MS (2006) Genetic divergence and migration patterns in a North American passerine bird: Implications for evolution and conservation. *Molecular Ecology*, **15**, 2141–2152.

DePristo MA, Banks E, Poplin R *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, **43**, 491–498.

Dor R, Cooper CB, Lovette IJ *et al.* (2012) Clock gene variation in *Tachycineta* swallows. *Ecology and evolution*, **2**, 95–105.

Dor R, Lovette IJ, Safran RJ *et al.* (2011) Low variation in the polymorphic Clock gene poly-Q region despite population genetic structure across barn swallow (*Hirundo rustica*) populations. *PloS one*, **6**, e28843.

Dunlap JC (1999) Molecular Bases for Circadian Clocks. *Cell*, **96**, 271–290.

Ebisawa T, Kajimura N, Uchiyama M *et al.* (1999) Alleic variants of human melatonin 1a receptor: function and prevalence in subjects with circadian rhythm sleep disorders. *Biochemical and biophysical research communications*, **262**, 832–837.

Elgvin TO, Hermansen JS, Fijarczyk A *et al.* (2011) Hybrid speciation in sparrows II: a role for sex chromosomes? *Molecular ecology*, **20**, 3823–3837.

Elgvin TO, Trier CN, Jensen H, Lien S, Sætre G-P (2015) A *de novo* genome assembly and linkage map of the house sparrow (*Passer domesticus*). (in preparation)

Ericson PG, Tyrberg T, Kjellberg A (1997) The earliest record of house sparrows (*Passer domesticus*) in northern Europe. *Journal of Archaeological Science*, **24**, 183–190.

Fernández-Suárez XM, Schuster MK (2010) Using the ensembl genome server to browse genomic sequence data. *Current Protocols in Bioinformatics*, **30**, 1–48.

Ferrer M, Bildstein K, Penteriani V, Casado E, de Lucas M (2011) Why birds with deferred sexual maturity are sedentary on islands: A systematic review. *PLoS ONE*, **6**, 1–7.

Fidler AE, Gwinner E (2003) Comparative analysis of Avian BMAL1 and CLOCK protein sequences: a search for features associated with owl nocturnal behaviour. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, **136**, 861–874.

Fitzpatrick MJ, Ben-Shahar Y, Smid HM *et al.* (2005) Candidate genes for behavioural ecology. *Trends in Ecology and Evolution*, **20**, 96–104.

Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics*, **180**, 977–993.

Gnerre S, MacCallum I, Przybylski D *et al.* (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences*, **108**, 1513–1518.

Grant PR (1972) Convergent and divergent character displacement. *Biological Journal of the Linnean Society*, **4**, 39–68.

Hermansen JS, Saether SA, Elgvin TO *et al.* (2011) Hybrid speciation in sparrows I: phenotypic intermediacy, genetic admixture and barriers to gene flow. *Molecular ecology*, **20**, 3812–3822.

Honma S, Kawamoto T, Takagi Y *et al.* (2002) *Dec1* and *Dec2* are regulators of the mammalian molecular clock. *Nature*, **419**, 841–844.

Hudson RR, Kreitman M, Aguadé M (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics*, **116**, 153–159.

Johnsen A, Fidler AE, Kuhn S *et al.* (2007) Avian Clock gene polymorphism: evidence for a latitudinal cline in allele frequencies. *Molecular ecology*, **16**, 4867–4880.

Jones S, Pfister-Genskow M, Cirelli C, Benca RM (2008) Changes in brain gene expression during migration in the white-crowned sparrow. *Brain Research Bulletin*, **76**, 536–544.

Joseph L, Wilke T, Alpers D (2003) Independent evolution of migration on the South American landscape in a long-distance temperate-tropical migratory bird, Swainson's flycatcher (*Myiarchus swainsoni*). *Journal of Biogeography*, **30**, 925–937.

Kearse M, Moir R, Wilson A *et al.* (2012) Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**, 1647–1649.

King DP, Takahashi JS (2000) Molecular genetics of circadian rhythms in mammals. *Annual Review of Neuroscience*, **23**, 713–742.

Kumar V, Singh BP, Rani S (2004) The Bird Clock: A Complex, Multi-Oscillatory and Highly Diversified System. *Biological Rhythm Research*, **35**, 121–144.

Kumar V, Wingfield JC, Dawson A *et al.* (2014) Biological clocks and regulation of seasonal reproduction and migration in birds. *Physiological and biochemical zoology*, **83**, 827–835.

Kume K, Zylka MJ, Sriram S *et al.* (1999) mCRY1 and mCRY2 Are Essential Components of the Negative Limb of the Circadian Clock Feedback Loop. *Cell*, **98**, 193–205.

Leder EH, Danzmann RG, Ferguson MM (2006) The candidate gene, Clock, localizes to a strong spawning time quantitative trait locus region in rainbow trout. *The Journal of heredity*, **97**, 74–80.

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

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Liedvogel M, Cornwallis CK, Sheldon BC (2012) Integrating candidate gene and quantitative genetic approaches to understand variation in timing of breeding in wild tit populations. *Journal of evolutionary biology*, **25**, 813–823.

Liedvogel M, Sheldon BC (2010) Low variability and absence of phenotypic correlates of *Clock* gene variation in a great tit *Parus major* population. *Journal of Avian Biology*, **41**, 543–550.

Liedvogel M, Szulkin M, Knowles SCL, Wood MJ, Sheldon BC (2009) Phenotypic correlates of *Clock gene* variation in a wild blue tit population: evidence for a role in seasonal timing of reproduction. *Molecular ecology*, **18**, 2444–2456.

Lischer HEL, Excoffier L (2012) PGDSpider: An automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, **28**, 298–299.

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature*, **351**, 652–654.

McKenna A, Hanna M, Banks E *et al.* (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.

Merlin C, Gegear RJ, Reppert SM (2009) Antennal circadian clocks coordinate sun compass orientation in migratory monarch butterflies. *Science*, **325**, 1700–1704.

Mueller JC, Pulido F, Kempenaers B (2011) Identification of a gene associated with avian migratory behaviour. *Proceedings of the Royal Society B: Biological sciences*, **278**, 2848–2856.

Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, **76**, 5269–5273.

O'Malley KG, Banks MA (2008) A latitudinal cline in the Chinook salmon (*Oncorhynchus tshawytscha*) *Clock* gene: evidence for selection on PolyQ length variants. *Proceedings of the Royal Society B: Biological sciences*, **275**, 2813–2821.

O'Malley KG, Camara MD, Banks MA (2007) Candidate loci reveal genetic differentiation between temporally divergent migratory runs of Chinook salmon (*Oncorhynchus tshawytscha*). *Molecular ecology*, **16**, 4930–4941.

O'Malley KG, Cross TF, Bailie D *et al.* (2014) Circadian clock gene (*OtsClock1b*) variation and time of ocean return in Atlantic salmon *Salmo salar*. *Fisheries Management and Ecology*, **21**, 82–87.

O'Malley KG, Ford MJ, Hard JJ (2010) Clock polymorphism in Pacific salmon: evidence for variable selection along a latitudinal gradient. *Proceedings of the Royal Society B: Biological sciences*, **277**, 3703–3714.

O'Malley KG, Jacobson DP, Kurth R, Dill AJ, Banks MA (2013) Adaptive genetic markers discriminate migratory runs of Chinook salmon (*Oncorhynchus tshawytscha*) amid continued gene flow. *Evolutionary applications*, **6**, 1184–1194.

Okano T, Yamamoto K, Okano K *et al.* (2001) Chicken pineal clock genes: implication of BMAL2 as a bidirectional regulator in circadian clock oscillation. *Genes to cells*, **6**, 825–836.

Outlaw DC, Voelker G, Mila B, Girman DJ (2003) Evolution of Long-Distance Migration in and Historical Biogeography of Catharus Thrushes: A Molecular Phylogenetic Approach. *The Auk*, **120**, 299–310.

Peterson MP, Abolins-Abols M, Atwell JW *et al.* (2013) Variation in candidate genes CLOCK and ADCYAP1 does not consistently predict differences in migratory behavior in the songbird genus Junco. *F1000Research*, **2**.

Piertney SB, Webster LMI (2010) Characterising functionally important and ecologically meaningful genetic diversity using a candidate gene approach. *Genetica*, **138**, 419–432.

Pulido F (2007) The Genetics and Evolution of Avian Migration. *BioScience*, **57**, 165–174.

Pulido F, Berthold P (1998) The microevolution of migratory behaviour in the blackcap: effects of genetic covariances on evolutionary trajectories. *Biologia e Conservazione della Fauna*, **102**, 206–211.

Pulido F, Berthold P (2010) Current selection for lower migratory activity will drive the evolution of residency in a migratory bird population. *Proceedings of the National Academy of Sciences*, **107**, 7341–7346.

Ramenofsky M, Agatsuma R, Ramfar T (2008) Environmental Conditions Affect the Behavior of Captive, Migratory White-Crowned Sparrows. *The Condor*, **110**, 658–671.

Rani S, Malik S, Trivedi AK, Singh S, Kumar V (2006) A circadian clock regulates migratory restlessness in the blackheaded bunting, *Emberiza melanocephala*. *Current Science*, **91**, 1093–1096.

Rolshausen G, Segelbacher G, Hobson KA, Schaefer HM (2009) Contemporary Evolution of Reproductive Isolation and Phenotypic Divergence in Sympatry along a Migratory Divide. *Current Biology*, **19**, 2097–2101.

Ruegg K, Anderson EC, Boone J *et al.* (2014) A role for migration-linked genes and genomic islands in divergence of a songbird. *Molecular Ecology*, **23**, 4757–4769.

Saino N, Romano M, Caprioli M *et al.* (2013) Timing of molt of barn swallows is delayed in a rare *Clock* genotype. *PeerJ*, **1**, e17.

Steinmeyer C, Kempenaers B, Mueller JC (2012) Testing for associations between candidate genes for circadian rhythms and individual variation in sleep behaviour in blue tits. *Genetica*, **140**, 219–28.

Steinmeyer C, Mueller JC, Kempenaers B (2009) Search for informative polymorphisms in candidate genes: clock genes and circadian behaviour in blue tits. *Genetica*, **136**, 109–117.

Suárez-López P, Wheatley K, Robson F *et al.* (2001) CONSTANS mediates between the circadian clock and the control of flowering in *Arabidopsis*. *Nature*, **410**, 1116–1120.

Summers-Smith JD (1988) *The Sparrows: A study of the genus Passer*. T & AD Poyser, Calton, Staffordshire, England.

Sætre G-P, Riyahi S, Aliabadian M *et al.* (2012) Single origin of human commensalism in the house sparrow. *Journal of evolutionary biology*, **25**, 788–96.

Tabor HK, Risch NJ, Myers RM (2002) Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nature Reviews Genetics*, **3**, 391–397.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Tauber E, Kyriacou CP (2005) Molecular Evolution and Population Genetics of Circadian Clock Genes. *Methods in Enzymology*, **393**, 797–817.

Visser ME, Caro SP, van Oers K, Schaper S V, Helm B (2010) Phenology, seasonal timing and circannual rhythms: towards a unified framework. *Philosophical transactions of the Royal Society of London B: Biological Sciences*, **365**, 3113–3127.

Weir B, Cockerham CC (1984) Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, **38**, 1358–1370.

Wren JD, Forgacs E, Fondon JW *et al.* (2000) Repeat polymorphisms within gene regions: phenotypic and evolutionary implications. *American Journal of Human Genetics*, **67**, 345–356.

Yasuo S, Watanabe M, Okabayashi N, Ebihara S, Yoshimura T (2003) Circadian clock genes and photoperiodism: Comprehensive analysis of clock gene expression in the mediobasal hypothalamus, the suprachiasmatic nucleus, and the pineal gland of Japanese quail under various light schedules. *Endocrinology*, **144**, 3742–3748.

Yoshimura T, Suzuki Y, Makino E *et al.* (2000) Molecular analysis of avian circadian clock genes. *Molecular brain research*, **78**, 207–215.

Young MW, Kay SA (2001) Time zones: a comparative genetics of circadian clocks. *Nature reviews Genetics*, **2**, 702–715.

Zhao W-N, Malinin N, Yang F-C *et al.* (2007) CIPC is a mammalian circadian clock protein without invertebrate homologues. *Nature cell biology*, **9**, 268–275.

Appendix

Supplementary table 1. Applicable scaffolds of the house sparrow genome assembly and the candidate/non-candidate genes they contain.

Scaffold	Canditate gene	Non-candidate gene
scaffold00001	*Clock*	*MXD4*
scaffold00003	*BHLHE40*	*FOXP1*
scaffold00005	*CPNE4*	*MACC1*
scaffold00006	*CSNK1A1*	*GALNT10*
scaffold00007	*Bmal2*	*AGK*
scaffold00009	*NPAS2*	*MRPS2*
scaffold00019	*AANAT*	*RHOT1*
	CSNK1D	*HELZ*
scaffold00021	*ADCYAP1*	*ZNF521*
scaffold00028	*CREB1*	*SLC39A10*
scaffold00031	*MTNR1A*	*PCM1*
scaffold00037	*Cry2*	*LTK*
scaffold00040	*NFIL3*	*SYK*
scaffold00050	*Bmal1*	*ORAOV1*
scaffold00054	*Cry1*	*NCF4*
	CSNK1E	*TOM1*
scaffold00055	*KIAA1737*	*AGO2*
scaffold00087	*Period3*	*PLCH2*
scaffold00089	*Period2*	*UGGT1*

Supplementary table 2. F_{st}-values for relevant scaffolds in four population pairs of *Passer* sparrows.

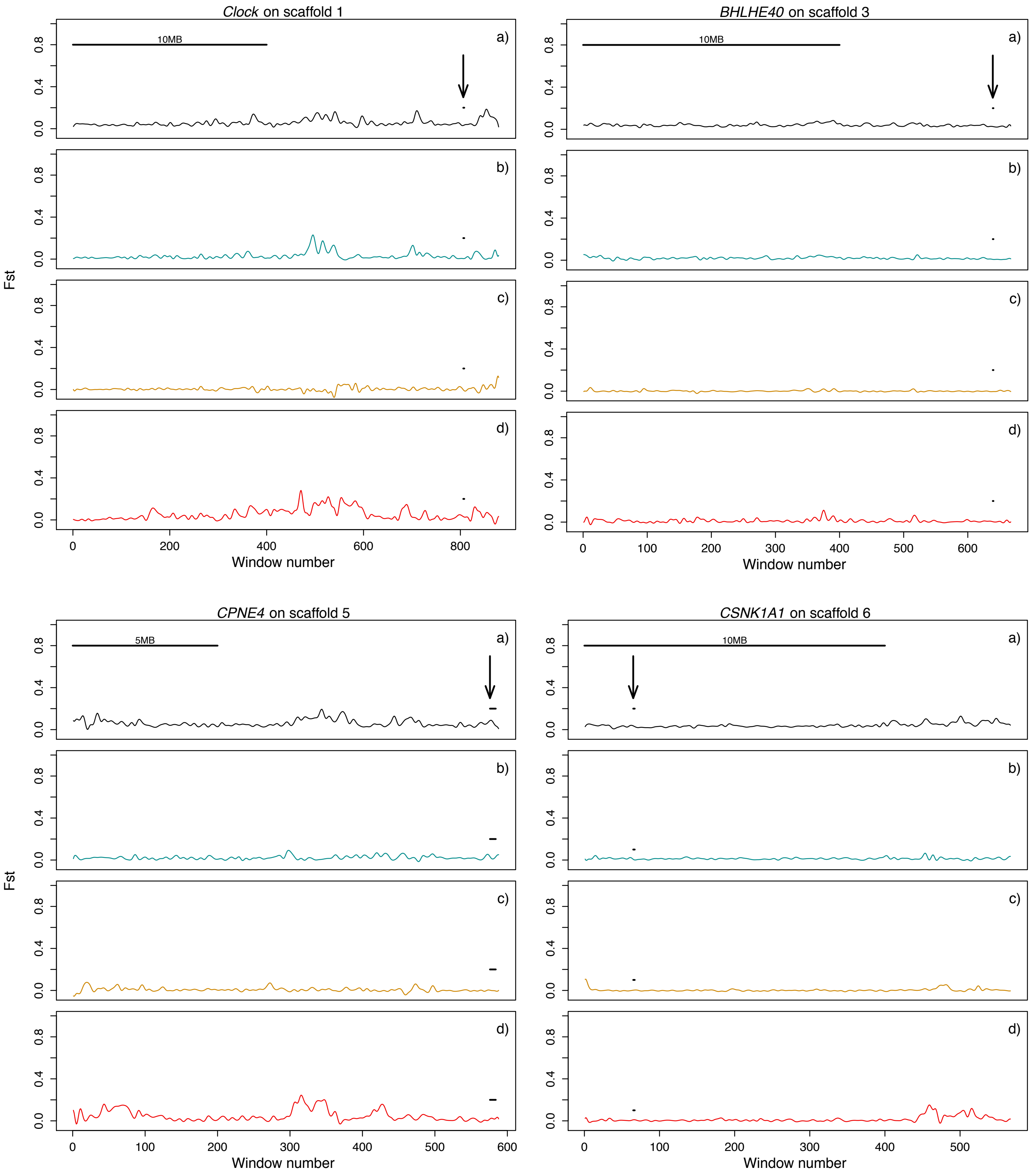
Scaffold	house sparrow: migratory - sedentary	Italian sparrow - sedentary house sparrow	Italian sparrow - sedentary Spanish sparrow	Spanish sparrow: migratory - sedentary
scaffold00001	0.0579759	0.0299707	0.00476039	0.0542608
scaffold00003	0.0394694	0.0195812	0.000780393	0.0114567
scaffold00005	0.0622684	0.0226133	0.00723906	0.0487495
scaffold00006	0.0418175	0.0136605	0.00287932	0.0147911
scaffold00007	0.0481057	0.0343813	0.00202554	0.0273731
scaffold00009	0.0627733	0.0222045	0.00416968	0.0360398
scaffold00019	0.0446541	0.0169151	0.00176553	0.0100832
scaffold00021	0.0660988	0.0254306	0.00864922	0.0505021
scaffold00028	0.0489985	0.0206178	0.00324594	0.0139377
scaffold00031	0.0538337	0.032581	0.00484825	0.0872387
scaffold00037	0.060982	0.0200299	0.0031494	0.0151563
scaffold00040	0.0873365	0.0859376	0.124593	0.106935
scaffold00050	0.0431271	0.0205848	0.000983561	0.0245876
scaffold00054	0.0508389	0.018941	0.00579453	0.0257943
scaffold00055	0.0698091	0.0247137	0.0179749	0.0415718
scaffold00087	0.0346598	0.0157012	-0.0022424	-0.000383631
scaffold00089	0.0587506	0.087105	0.0145791	0.0338458

Supplementary table 3. Nucleotide diversity (π) for non-candidate genes in all populations.

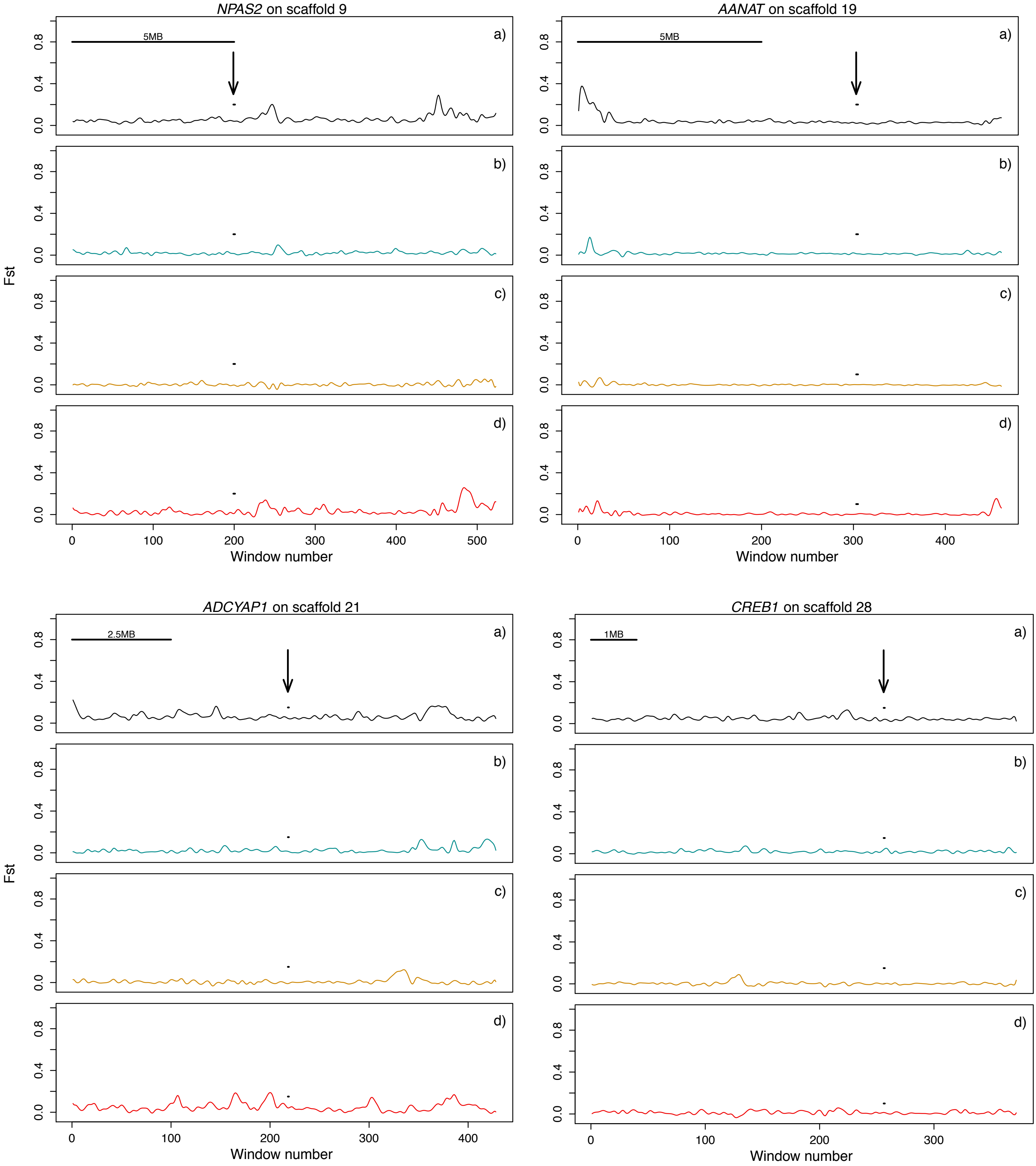
Gene	Migratory house sparrow	Sedentary house sparrow	Italian sparrow	Migratory Spanish sparrow	Sedentary Spanish sparrow
AGO2	0.317909	0.413905	0.317137	0.339728	0.410107
HELZ	0.380282	0.418692	0.407768	0.371934	0.46736
RHOT1	0.419794	0.461199	0.413103	0.406635	0.457981
MXD4	0.39893	0.518828	0.456936	0.440319	0.502289
ZNF521	0.327075	0.441923	0.445217	0.284416	0.520665
SLC39A10	0.388874	0.440199	0.41738	0.399341	0.490744
PCM1	0.325506	0.356814	0.338907	0.186854	0.41602
LTK	0.370586	0.463168	0.432453	0.365881	0.424793
FOXP1	0.40094	0.491008	0.444989	0.418041	0.484794
SYK	0.334621	0.392159	0.404277	0.187521	0.411417
ORAOV1	0.401827	0.452273	0.389863	0.405956	0.501121
NCF4	0.362114	0.469447	0.465137	0.438431	0.507038
TOM1	0.366416	0.376529	0.526745	0.452203	0.524739
MACC1	0.438602	0.38535	0.496287	0.325498	0.564163
GALNT10	0.432076	0.486832	0.42808	0.441375	0.468835
AGK	0.342074	0.466396	0.380236	0.359981	0.447251
PLCH2	0.414946	0.461027	0.427118	0.396919	0.486982
UGGT1	0.389269	0.394035	0.389308	0.376734	0.479013
MRPS9	0.32313	0.332063	0.322271	0.211566	0.427639

Supplementary table 4. Tajima's D for non-candidate genes in all populations.

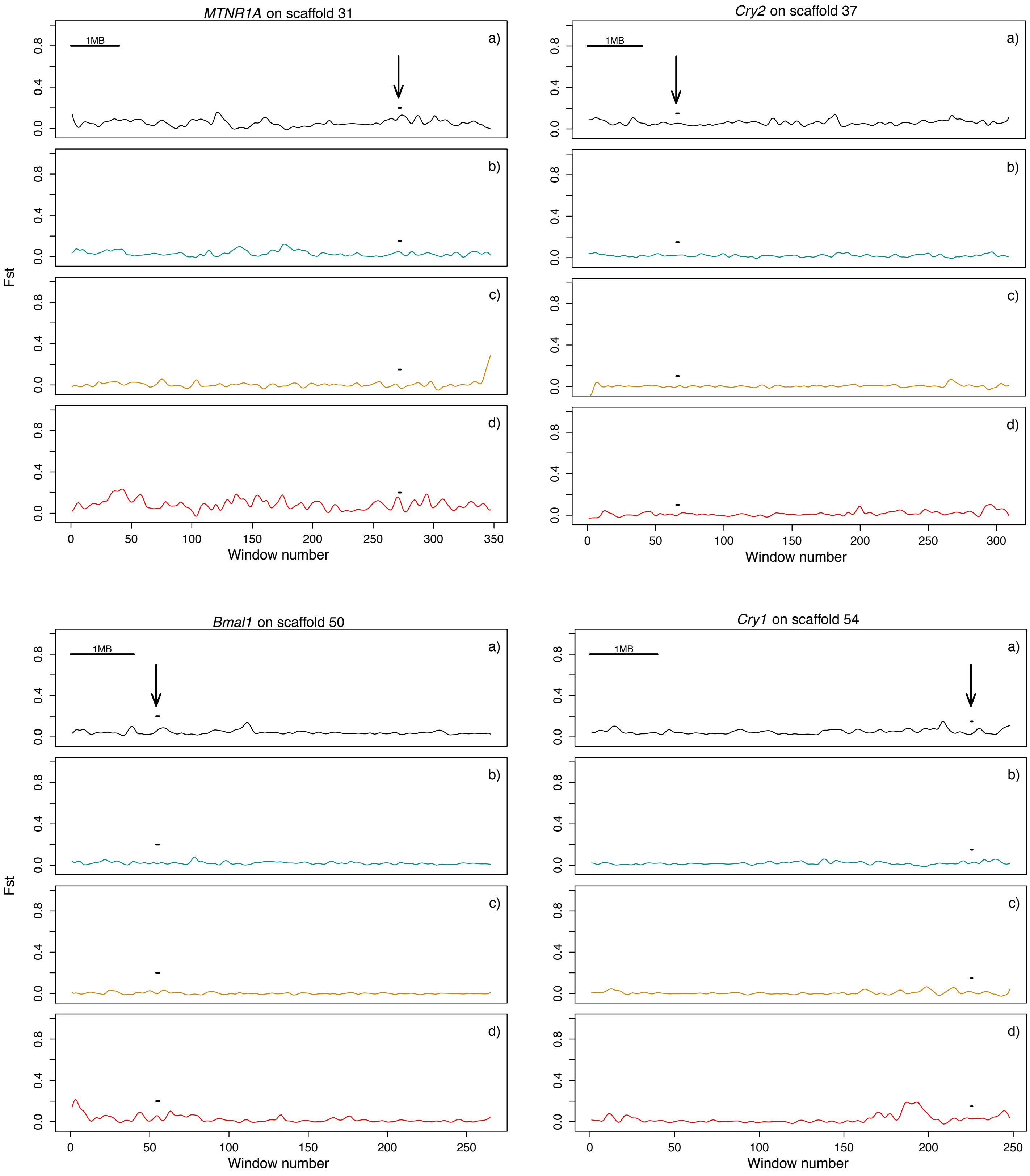
Gene	Migratory house sparrow	Sedentary house sparrow	Italian sparrow	Migratory Spanish sparrow	Sedentary Spanish sparrow
AGO2	0.707401	1.24392	0.88027	0.596071	0.818564
HELZ	0.57985	0.820185	0.688499	0.654299	0.820382
RHOT1	0.744582	1.11541	0.851921	0.851304	0.987923
MXD4	0.416744	0.738624	0.592102	0.584923	0.624032
ZNF521	0.213487	0.54925	0.501766	0.203544	0.443815
SLC39A10	0.665053	1.04524	0.824801	0.771999	0.892747
PCM1	0.611987	1.05219	0.89573	0.457959	0.976789
LTK	0.433882	0.786321	0.656794	0.538972	0.679073
FOXP1	0.237309	0.396916	0.289897	0.277365	0.323131
SYK	0.449942	0.64656	0.74392	0.141391	0.692247
ORAOV1	0.833666	0.910646	0.515862	0.662547	0.835553
NCF4	0.5335	1.11252	0.835403	0.766738	0.748178
TOM1	0.393582	0.496503	0.740504	0.630447	0.794847
MACC1	0.299219	0.542172	0.864678	0.268198	0.88253
GALNT10	0.472579	0.648336	0.503318	0.489827	0.570445
AGK	0.449117	0.85896	0.532879	0.495106	0.68336
PLCH2	0.671159	0.904066	0.792053	0.687367	0.886207
UGGT1	0.97156	0.14143	1.08073	1.07499	1.40159
MRPS9	0.549753	0.93893	0.650638	0.363677	0.789778



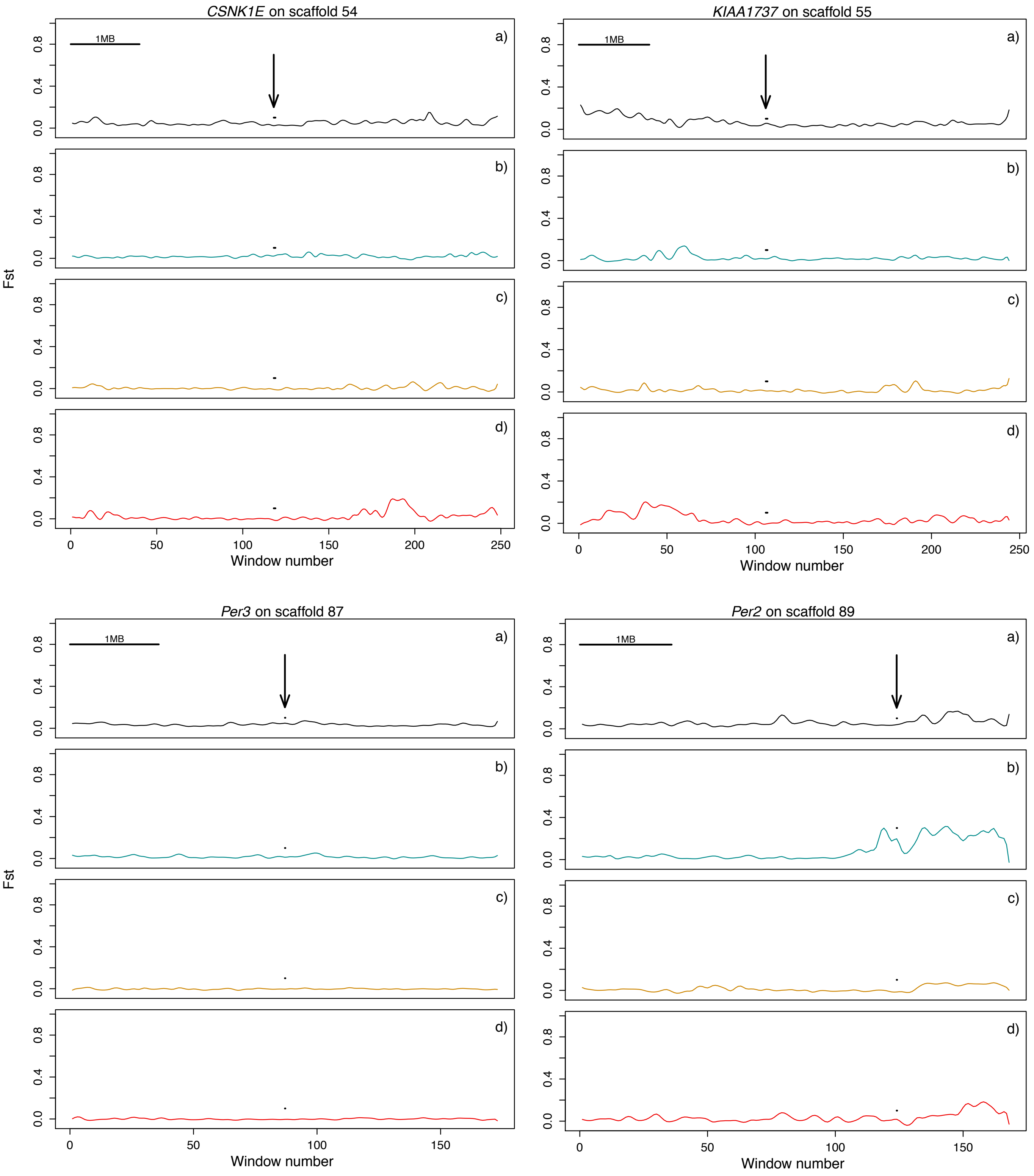
Supplementary figure 1. Scaffold wide F_{st}-values for those containing a candidate gene. The y-axis represents F_{st}-values. The x-axis represents window number as F_{st} was calculated on a windowed basis with a window size of 100kb in overlapping steps of 25kb. Arrow and point mark the location of the relevant candidate gene. Divergence was estimated between the following population pairs form top to bottom: a) migratory versus sedentary house sparrow, b) Italian sparrow versus sedentary house sparrow, c) Italian sparrow versus sedentary Spanish sparrow and d) migratory versus sedentary Spanish sparrow.



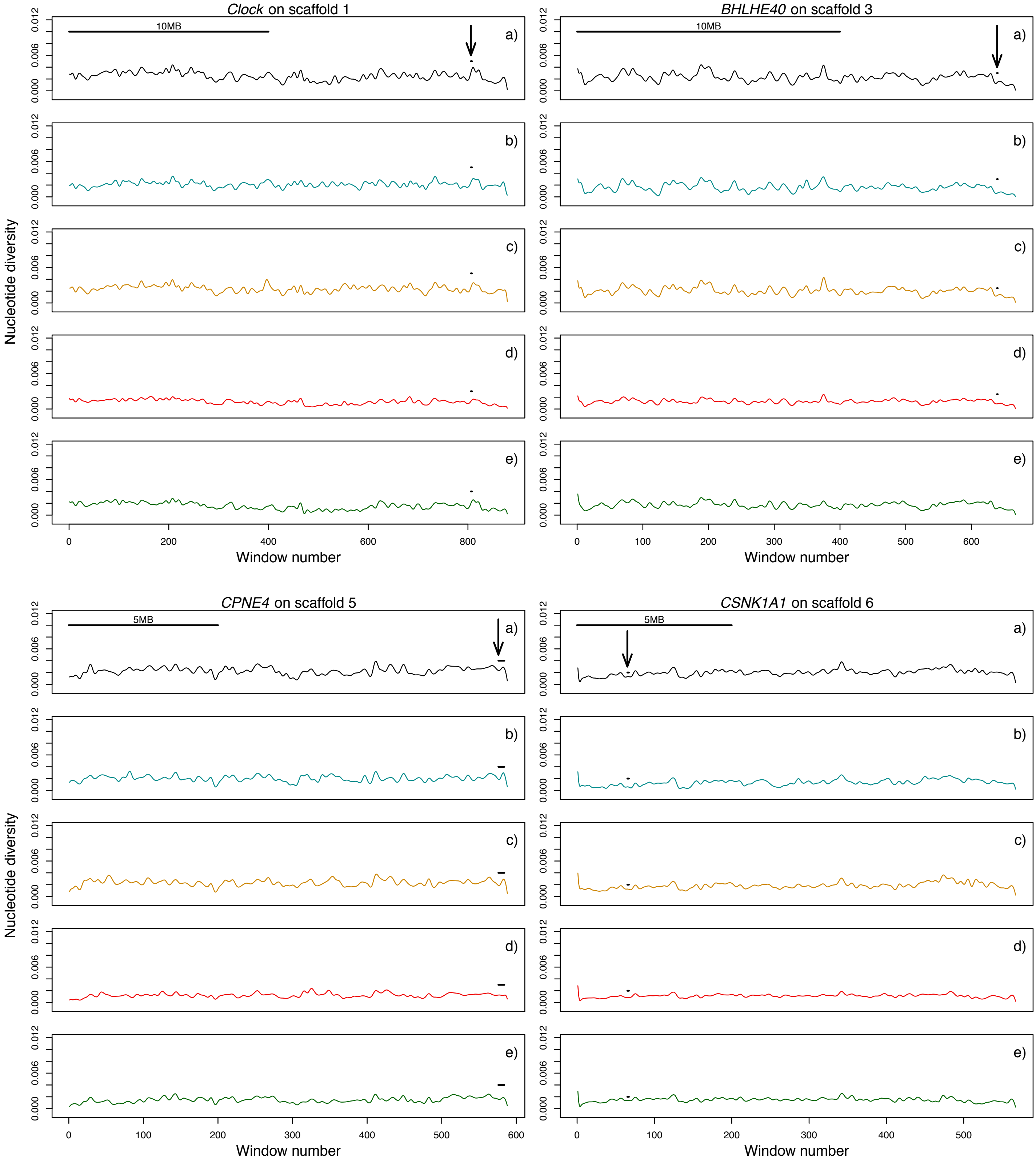
Supplementary figure 1. Continued.



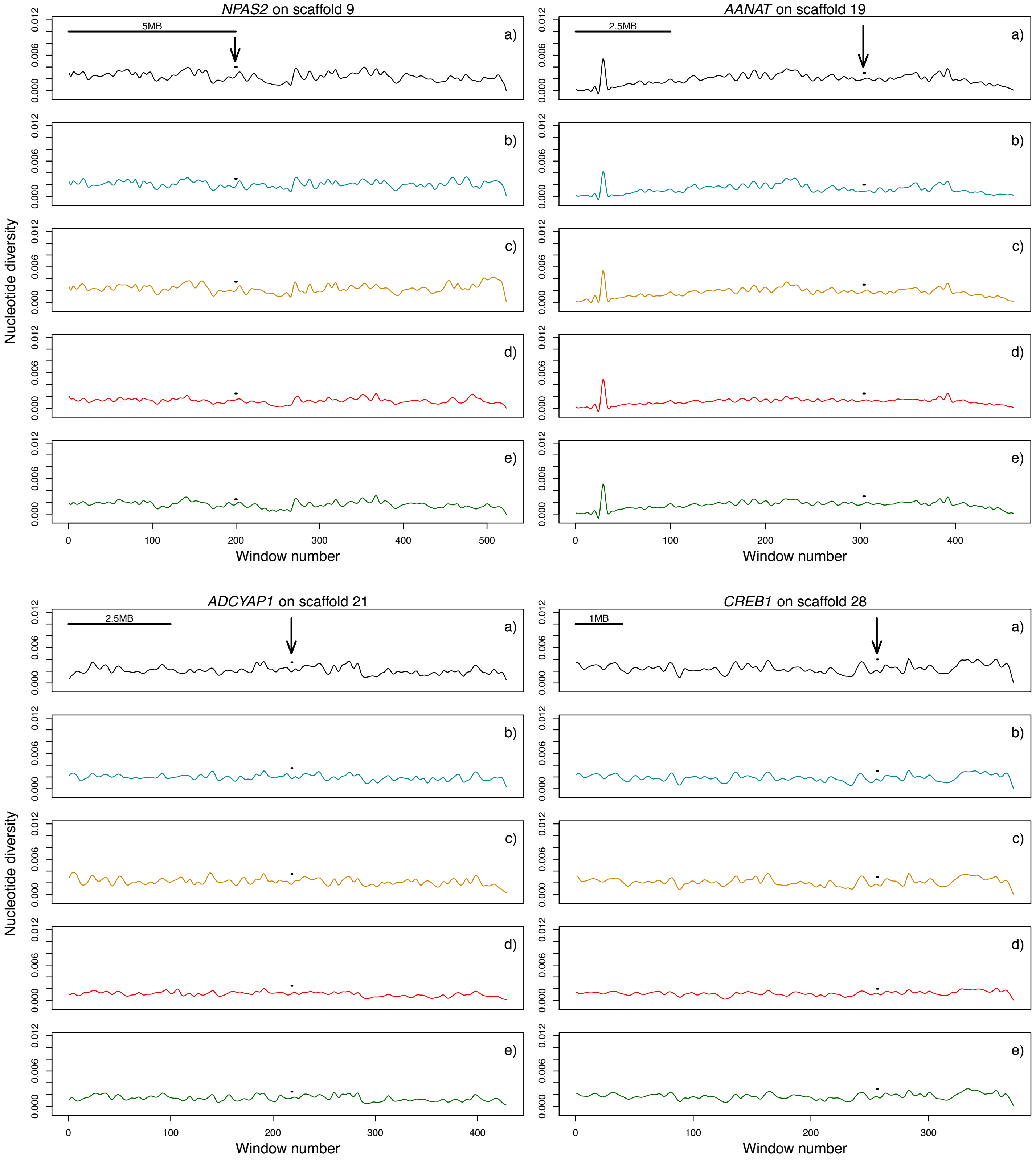
Supplementary figure 1. Continued.



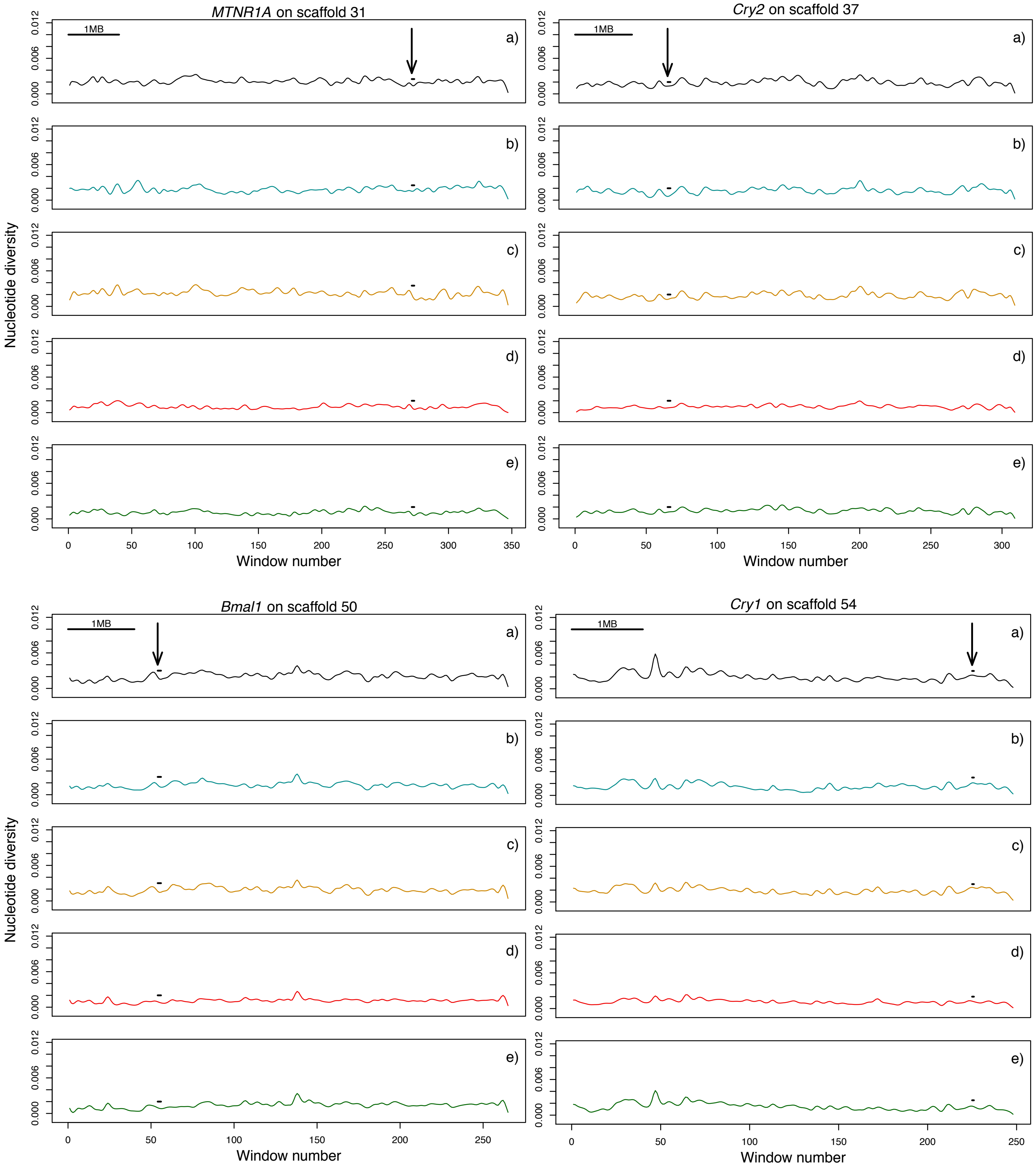
Supplementary figure 1. Continued.



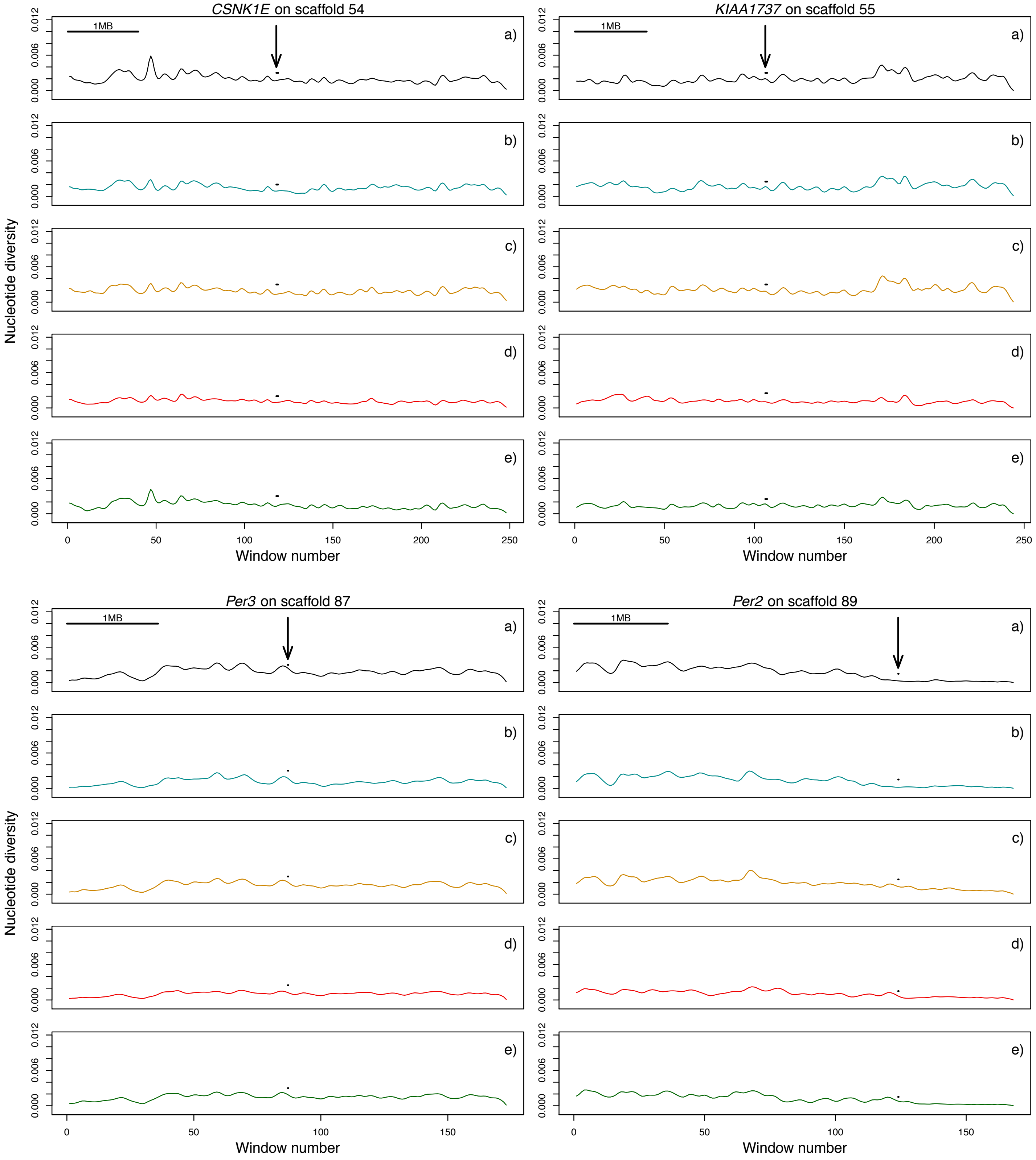
Supplementary figure 2. Scaffold wide nucleotide diversity (π) for those containing a candidate gene. The y-axis represents nucleotide diversity. The x-axis represents window number as values were calculated on a windowed basis with a window size of 100kb in overlapping steps of 25kb. Arrow and point mark the location of the relevant candidate gene. Nucleotide diversity was calculated for all populations: e) migratory house sparrows, f) sedentary house sparrows, g) Italian sparrows, h) sedentary Spanish sparrows and i) migratory Spanish sparrows.



Supplementary figure 2. Continued.



Supplementary figure 2. Continued.



Supplementary figure 2. Continued.