Structural variants in the great tit genome and their effect on seasonal timing

Vinicius Henrique da Silva

Thesis committee

Promotors:
Prof. Dr Martien A. M. Groenen
Professor of Animal Breeding and Genomics
Wageningen University & Research

Prof. Dr Marcel E. Visser
Special Professor of Ecological Genomics
Wageningen University & Research

Co-promotors:
Dr Richard P.M.A Crooijmans
Assistant Professor, Animal Breeding and Genomics
Wageningen University & Research

Dr Anna M. Johansson
Researcher at Animal Breeding and Genetics
Swedish University of Agricultural Sciences

Other members:
Prof Dr Geert Wiegertjes, Wageningen University & Research
Dr Evelien Verhulst, Wageningen University & Research
Dr Ben Wielstra, Leiden University
Prof. Dr Linda Keeling, Swedish University of Agricultural Sciences

This research was conducted under the joint auspices of Swedish University of Agricultural Sciences and the Graduate School Wageningen Institute of Animal Sciences of Wageningen University & Research and is part of the Erasmus Mundus Joint Doctorate program "EGS-ABG"

Structural variants in the great tit genome and their effect on seasonal timing

Vinicius Henrique da Silva

ACTA UNIVERSITATIS AGRICULTURAE SUECIAE DOCTORAL THESIS Nº
2020:1

Thesis
submitted in fulfillment of the requirements for the joint degree of doctor between
Swedish University of Agricultural Sciences
by the authority of the Board of the Faculty of Veterinary Medicine and Animal Science and
Wageningen University
by the authority of the Rector Magnificus
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board and the Board of the Faculty of Veterinary Medicine and Animal Science
to be defended in public
on Friday, January 24, 2020
at 4 p.m. in the Aula of Wageningen University.

Vinicius Henrique da Silva
Structural variants in the great tit genome and their effect on seasonal timing
173 pages.

Joint PhD thesis, Swedish University of Agricultural Sciences, Uppsala, Sweden and Wageningen University, Wageningen, the Netherlands (2019)
With references, with a summary in English and a summary in Dutch (samenvatting)

ISBN (print version): 978-91-7760-520-1
ISBN (electronic version): 978-91-7760-521-8
ISSN: 1652-6880

Abstract

da Silva, V.H. (2019). Structural variants in the great tit genome and their effect on seasonal timing. Joint PhD thesis between Wageningen University & Research, the Netherlands and Swedish University of Agricultural Sciences, Sweden

The biodiversity of our planet has been increasingly endangered by human actions. This nature biodiversity is strictly correlated with genomic diversity of all the species in the ecosystem. Thus, a broader understanding on the genome of wild species may be extremely useful to understand selection and plasticity in the natural species of our changing world. The great tit (*Parus major*) is a songbird that has been extensively explored in ecological and evolutionary studies, shedding light on the effects of the global warming on nature. The seasonal timing of the great tit has been shifting under the global warming, but the knowledge on particular genes associated with timing is still limited. Although the effect of single nucleotide changes on the breeding timing of great tits has been investigated, the effect of more complex structural variants is largely unknown. In fact, the genomic structural variability was never explored in detail in these species. The aim of this thesis was to detect, map, characterize and associate, with seasonal timing, structural variants that are present in the great tit genome such as copy number variations (CNVs) and inversions. First, this thesis presents a genome-wide map of CNV regions in the great tit genome, showing how these variants are associated with genomic architecture that underlies their molecular formation. Great tit CNVs, in accordance to reported in several mammalian species, are enriched at evolutionary breakpoints. Although it supports the importance of CNVs during speciation like is described in mammals, a remarkable difference is that neuronal related genes may play a central role on the great tit speciation. Second, CNVs were associated with breeding timing. Although no strong association was found, suggestive associations such as a copy number gain in a gene related to circadian clock deserves further investigation. Finally, this thesis investigate in detail the genomic complexity of a large (≈64 Mb) and widespread (≈5%) inversion in the Chromosome 1A. Interestingly, this inversion is a recessive lethal selfish structural rearrangement (i.e. breaks the Mendel's law). The inversion is inherited twice more than expected from male carriers but are normally inherited from female carriers, suggesting that a meiotic drive mechanism during spermatogenesis maintains this large inversion in great tit populations.

Contents

Chapter 1

Introduction

1.1 Genetic diversity as the pillar of species conservation

1.1.1 Biodiversity and climate change

One of the most important challenges for humankind is the maintenance of biodiversity on our planet, given that species are disappearing at an alarming rate and may need intervention to guarantee their survival Frankham et al. (2009). There are a number of negative interactions between humans and the environment such as pollution and deforestation, which can harm an ecosystem and consequently the ecology of species. Ecology can be defined as the interaction between organisms and their environment whereas evolution is the heritable change in populations of organisms over generations. Ecology and evolution are strictly related themes and the majority of the scientific questions in one area to some extent will touch another one.

As genetic diversity is the substratum for evolution, diversity is an essential pillar in conservation genetics. Changes in the environment are the main driver of natural selection, where individuals with higher chance to reproduce have a higher fitness. Consequently, specific genetic variants from adapted animals will increase in future generations which can lead to a lower amount of genetic diversity. Therefore, species may start to disappear through changing ecosystems as a consequence of this damaged biodiversity.

The environment is constantly changing due to natural ecological processes. However, in the last decades many human activities such as deforestation (Zemanova et al., 2017), gas emission (Meinshausen et al., 2009); in great part coming from animal production (Koneswaran & Nierenberg, 2008) and industrialization (Mgbemene et al., 2016); caused fast and profound shifts in natural habitats. These human activities lead to a phenomenon that is increasingly studied, climate change. The effects of climate change on natural populations has been extensively studied in a wide range of species, which usually have their phenology affected by these environmental changes. The phenology of several species has been shifting and resulting in a mismatch between interconnected species belonging to the same ecosystem (Visser & Both, 2005). Therefore, a deeper understanding of the genetic variability, which directly reflects the biodiversity, may assist in future efforts to prevent ecological imbalance or even species extinction. In fact, the resettlement of individuals increases the genetic diversity and adaptive potential in species with a disrupted ecosystem, and may be a crucial step for their conservation (Coates et al., 2018).

1.1.2 Ecology and evolution of great tits

Box 1. Great tit: the model species

The great tit (*Parus major*) is a territorial songbird that occupies a wide range of habitats (van Balen, 2002) being found from North Africa across temperate Eurasia as well as into large parts of tropical South East Asia (Portenko & Wunderlich 1984, Figure 1.1). The great tit is a widely studied species in ecology and evolution that has been used as a model species to understand reproduction (Smith et al., 1989), learning/cognition (Cauchoix et al., 2017) and the effects of human activities on their behaviour (Corsini et al., 2017).

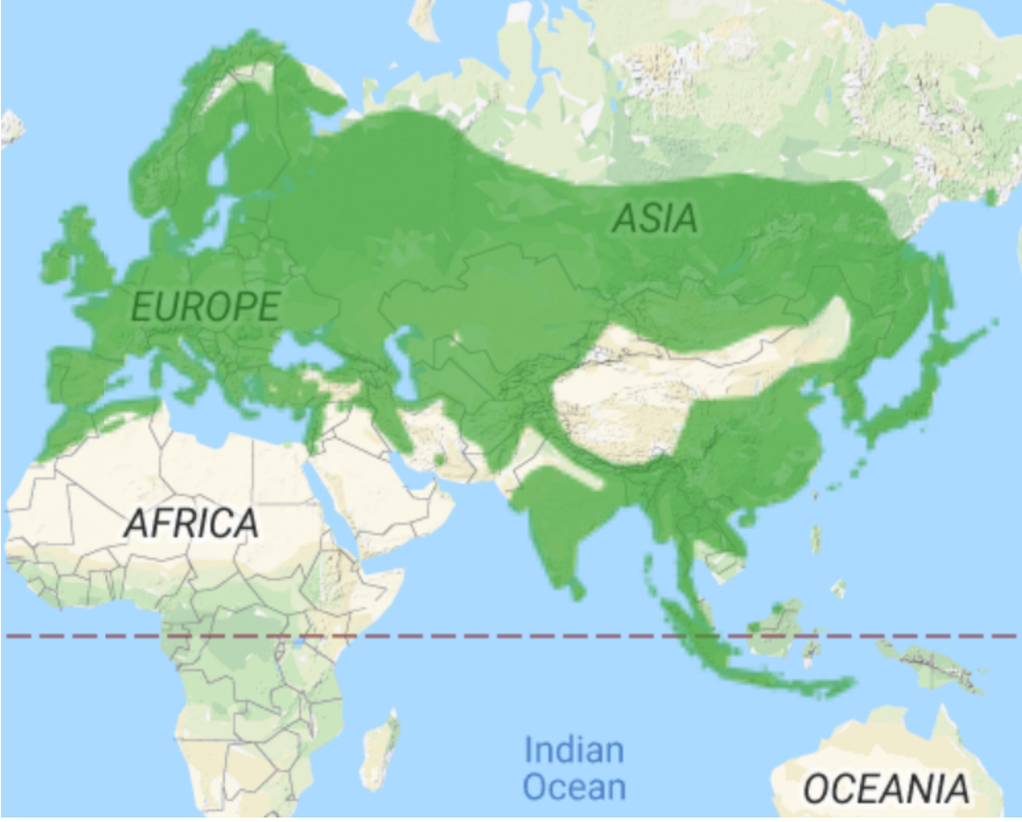


Figure 1.1: Distribution of *Parus major* species around the globe. Adapted from (BirdLife, 2019).

Studies on the great tit shed light on how the life cycle of natural species has been shifting under climate change (Visser & Both, 2005). For example, seasonal phenotypes, like e.g. egg-laying date during a breeding season, have been used to understand the relationship between warmer/colder seasons and breeding timing

(Schaper et al., 2011). However, the pace of change in phenology is clearly different in species that present trophic interactions with great tits, such as the caterpillar peak biomass date (Visser et al., 1998). This mismatch between newborn chicks and the date of the biomass peak of the caterpillars, which is the main food for the chicks, has generate questions about the effects of climate change in ecosystems.

Given the importance of great tit as a model species in ecology and evolution, more advanced molecular techniques have been developed and implemented to study this species. An important advancement was the publication describing a reference genome to the great tit, which in addition explored evolution of cognition by examining the species genome and methylome (Laine et al., 2016). The reference genome for the great tit allowed gene annotation and consequently evolutionary studies with genomic information. The great tit genome has a total number of 33 chromosomes, which harbors more than 4 millions SNPs. The knowledge on the great tit genome and the SNPs across the chromosomes was crucial to the development of a custom high density SNP array (Kim et al., 2018), which is able to successfully genotype more than 500 thousand single nucleotide polymorphisms (SNPs) per sample. It allowed genome-wide association studies (GWAS) to clarify the genetic basis of breeding timing (Gienapp et al., 2017) and beak size in great tits (Bosse et al., 2017). The breeding timing in birds is a seasonal trait that is reflected by the laying date of the first egg in a breeding season (i.e. egg-laying dates). Therefore, Gienapp et al. 2017 performed an environment-dependent SNP based GWAS to capture genes underlying variation in breeding timing. However, they found no genes that are strongly associated with egg-laying date in great tits, evidencing the polygenic and plastic nature of timing. On the other hand, Bosse et al. 2017 showed by selective sweep analysis, that the longer beaks are associated with a specific haplotype of the *COL4A5* gene, which is also positively associated with fledgling production (i.e. proxy for fitness). Interestingly, great tits from UK have longer beaks than those from the Netherlands, which suggests a recent human-driven selection for longer beaks in this species caused by more artificial feeding in UK than in the rest of Europe.

The recent effort to better understand the genetic and epigenetic variation in great tits is an important next step to comprehend how this species is responding to our changing world and how their populations may increase or decrease on the years to come. Moreover, molecular studies performed in great tit can assist similar efforts on other wild species. However, apart from the considerable advancements on the understanding of the great tit genome using SNPs and their respective haplotypes, structural variants (SVs) such as translocations, duplications/deletions and inversions have been poorly explored in this species. Fortunately, with the release of the great tit reference genome (Laine et al., 2016), the use of sequencing and genotyping (i.e. high density SNP array, (Kim et al., 2018)) to the identification of SVs was facilitated. There are an increasing number of software available to detect SVs of which can use more than one algorithm in order to improve specificity and sensitivity (Ye et al., 2016). On the other hand, by using SNP arrays, one of the SV types which focuses on genome duplications and deletions named copy number variation (CNVs) can be uncovered by signal intensity and heterozygosity level of their overlapping SNP probes. Also, different SNP array based algorithms are available to the identification of CNVs, which show different success rate, average stability rate, sensitivity, consistence and reproducibility (Zhang et al., 2014b).

1.2 Genomic structural variants and biodiversity

1.2.1 Biological effects and evolutionary footprints of structural variants

Research on genomic variants usually focuses on single nucleotide changes (Casci, 2010), but recently it has become clear that the complexity of the genome goes much further. Apart from single nucleotides, variants in the genome structure also underlie an important part of the evolutionary history (Katju & Bergthorsson, 2013) and are associated with a wide range of phenotypes (Weischenfeldt et al., 2013) in humans, livestock and wild species.

In humans, structural variants such as CNVs have been linked to different kinds of mental disability by causing disorders in the nervous system (Lee & Lupski, 2006), with obesity (D'Angelo & Koiffmann, 2012), cancer predisposition (Shlien & Malkin, 2010), hemophilia (Antonarakis et al., 1995) and several other diseases and syndromes (for a review see Zhang et al. (2009)). Studying CNVs is also important to understand the evolutionary history of humans as CNVs in genes underpinning inflammatory response and cell proliferation may underlie phenotypic differentiation of humans and chimpanzees (Perry et al., 2008). Susceptibility to diseases that are still not curable, such as the acquired immunodeficiency syndrome (AIDS), rely on CNVs. The importance of CNVs to understand AIDS was shown by a meta-analysis that included more than nine independent studies that indicated that an increase in the number of copies of the *CCL3L1* gene decrease the risk of a HIV-1 infection (Liu et al., 2010).

In livestock, CNVs have also been associated with different diseases, syndromes and morphological phenotypes (Clop et al., 2012) such as e.g. the pea-comb phenotype in chicken (Wright et al., 2009). Moreover, quality-related production traits such as meat tenderness have been associated with CNVs (da Silva et al., 2016), which in is known to underlie a widespread effect on gene expression in muscle (Geistlinger et al., 2018). Mainly in cattle, several studies have shown how CNVs have shaped the current breeds through natural and artificial selection (Keel et al., 2016; Upadhyay et al., 2017). CNVs are also important to the mutation dynamics of CpG dinucleotides leading to a higher genomic ‘flexibility’ in the evolution of chickens (Pértille et al., 2019). In fact, CNVs overlap CpG sites more than expected than change in other birds, such as the great tit (Chapter 2 of this thesis).

There is a growing effort to explore the evolutionary importance of CNVs in natural populations. For example, in house-mouse three conserved genes endured major population-specific duplications (Pezer et al., 2015). Other studies also exist in plasmodium (Simam et al., 2018), stickleback (Chain et al., 2014) and pine (Prunier et al., 2017) in which CNVs confer adaptability to a highly diverse/novel ecological environments that are rapidly changing. However, albeit some studies explored the role of CNVs to adaptation under fast environmental changes, the direct association of CNVs with intraspecific phenotypes and fitness components is poorly explored in the literature. Apart from CNVs, the fitness effect of the inversions have been increasingly explored in different species.

In human evolution, inversions had a fundamental role as more than 1,000 inversions diverge between human and chimpanzee genomes (Hellen, 2015). Moreover, the history of different human civilization is partially reflected by inversions. For example, different human populations show a distinct frequency for a pericentric inversion in chromosome 9 (Hsu et al., 1987). Although, the effect of inversion on human diseases is still limited (Puig et al., 2015), neurodegenerative diseases have associated with polymorphic inversions (Pittman et al., 2006), which in turn can cause a predisposition to other disease-related structural rearrangements (Puig et al., 2015).

Polymorphic inversions have been associated with a number of traits in *Drosophila*, ranging from body size to male mating success (reviewed in (Hoffmann & Rieseberg, 2008)), which can considered as a proxy for fitness. Moreover, the speciation in a major human malaria vector *(Anopheles funestus*) is associated with inversions (Ayala et al., 2011), evidencing the importance of inversions to better understand the recent evolution of widespread disease vectors. Moreover, the mating strategy in different wild birds is associated with inversions, such as the male morphs in ruff (*Philomachus pugnax*) (Lamichhaney et al., 2016) or the disassortative mating in white-throated sparrow (*Zonotrichia albicollis*) (Tuttle et al., 2016).

Although the inherent role of different SVs has been increasingly explored, the strategy used for detection and classification of SVs is not trivial. The methods to detect CNVs are still evolving and need to be interpreted carefully. Moreover, even ignoring the technical challenges, the biological variability among structural variants is stunning by itself. Different classes of structural variants can share definitions and mechanisms of formation (Carvalho & Lupski, 2016), which confer another layer of complexity to their study.

1.2.2 Methods to detect structural variants

Several methods have been used to discover structural variants in the genome. These greatly differ in resolution and false negative-positive rates (Alkan et al., 2011). The three methods are fluorescent *in situ* hybridization (FISH), different array types and Next generation sequencing (NGS). FISH was a pioneer method that is able to karyotype large structural variants (≈500 kb to 5 Mb, Trask (1998)). However, for the discovery of shorter variants the development of microarrays was crucial. There are two types of microarrays primarily represented by array comparative genomic hybridization (CGH) and SNP arrays. CGH compares the hybridization of two labelled samples (i.e. test and reference) to a set of hybridization targets, which are typically long oligonucleotides or bacterial artificial chromosome (BAC) clones. SNP array platforms are also based on hybridization, but the hybridization is performed per sample and intensities measured in several samples are clustered to detect signal deviations in each sample (Alkan et al., 2011). Most of the SNP array based software use the relative probe intensity signal (log R ratio - LRR) from each probe to estimate deviations in the number of copies. The interpretation and filtering of these signals have been evolving and more recently the frequency of the B allele (BAF) has been also integrated in some algorithms in order to improve sensitivity and activity of the CNV calls (Yau & Holmes, 2008). One of the most widely used algorithms that considers both LRR and BAF is implemented in the PennCNV (Wang et al., 2007) software, which has been pointed to have the best consistency with a CGH gold standard (≈24 million probes per sample, Zhang et al. 2014b).

The use of next-generation sequencing (NGS) technologies opened new possibilities to study structural variants. NGS technologies are able to produce millions of reads that can be used to construct a de-novo reference genome or be mapped onto an existent reference genome. Algorithms that use NGS read information to identify structural variants can be generally classified into read-pair (RP), split-read (SR), read-depth (RD) and assembly (review in Ye et al. 2016). RP is based on the fact that mapping distance between two reads in a pair will differ if a deletion/insertion is present. Moreover, some RP based software such as Break Dancer (Chen et al., 2009) can gather reads with abnormal insert size and orientation to uncover possible inversions and translocations. Otherwise, SR method uses the information of reads that split at the breakpoint of a structural variant. These split reads map separately, and/or in a reverse orientation, to the reference genome, which provides location, size and assist in the classification of the identified variants. RD is not based on the genomic location of the read pairs or split reads but otherwise on the number of reads overlapping certain genomic regions. Therefore, duplicated or deleted regions can be identified due to their significantly higher or lower read coverage. Finally, assembly based methods usually perform a local assembly on the missing read-pairs and therefore variants are called from the assembled contigs. However, although an increasing number of software to detect structural variants from sequencing data has been described in the literature, several computational and bioinformatics challenges remain (Tattini et al., 2015). Moreover, the underlying costs in NGS can be still prohibitive for large populations.

1.2.3 Genomic architecture underlies structural variant formation

The understanding of the molecular basis of a wide range of phenotypes, across several species, has evolved quickly. An increasing number of studies has shown the tremendous plasticity and dynamic nature of the genome. However, genomic variability can implicate in complex gene structures that are challenging to fully expose. The high complexity of a genome is usually linked with structural variants which sometimes can be confusing in their definitions, e.g. limit length to distinguish insertions/deletions (INDELs) and copy number variations (CNVs) or length, repetitive nature and mobility of a translocation to be considered a transposon (denominated transposition instead). In general, translocations, changes in copy number and inversions overlap, to a reasonable extent, the majority of the classes of structural variants that are reported in the literature.

Translocations are chunks of the genome moved from one genomic location to another, which can be balanced or unbalanced depending whether genetic material is lost or added at the translocated region (Harewood et al., 2017). Thus, an unbalanced translocation is followed by a copy number change. Formally, changes in copy number may be generally classified as CNVs if they encompass more than 1kb (or >50 bp in some definitions (Clop et al., 2012), which usually can be identified by NGS but not by SNP-arrays) or as INDELs if shorter than 50 bp in size (or <50 bp in some definitions (Sehn, 2015)). In fact, INDELs might be not even generally classified as SVs (Ye et al., 2016). In turn, CNVs that are located in reverse orientation can underlie the formation of inversions (Palacios et al., 2017) by providing substrate to non-allelic homologous recombinations (NAHR, Hoffmann & Rieseberg (2008); Carvalho & Lupski (2016)). There are also some evidence that small inversions and nonrecurrent CNVs can be also formed by microhomology-mediated break-induced replication (MMBIR) (Hastings et al., 2009) and fork stalling and template switching (FoSTeS) (Zhang et al., 2009).

Nomenclature in structural variants also encompass terms such as segmental duplications (SDs, also known as low copy repeats - LCRs), which represent the homologous regions in the genome; or transposable elements which account for a substantial fraction of copy number changes and are also known as 'jumping genes'. Segmental duplications, in essence, are CNVs that were fixed in a given species and may collaborate to the expansion of gene families. Otherwise, transposable elements can insertionally mutate the genes in which they land (Chen et al., 2005; Batzer & Deininger, 2002) and underlie the formation of additional variants as deletions, duplications, inversions, or translocations (Sen et al., 2006; Bailey et al., 2003). Given the interdependence among all different structural variant classes and their sharing mechanisms of formation, it may be informative to explore different classes of structural variants jointly also because one class can be intrinsically associated to another. The same group of replication-based mechanisms (RBMs) can produce different SVs classes of which in turn can be part of a specific genomic architecture, which endures a specific or multiple RBMs (Figure 1.2). For example, repetitive elements in the genome can be rich in adenine-timine (AT-rich intervals) or in CpG sites (i.e. which can be methylated), which are associated with regions prone to break (Franchitto, 2013) and with a high recombination rate (Singhal et al., 2015), respectively. It is known that AT-rich intervals are enriched for rare variants (Carvalho & Lupski, 2016) (multiple origins), likely formed by break-induced replication (BIR) mechanisms such as non-homologous end joining (NHEJ), whereas CpG to more common CNVs (Chapter 2 of this thesis) which tend to be formed by homologous recombination (e.g. NAHR). Thus, different RBMs are more prevalent at certain genomic architecture leading to a higher incidence of a specific SV. However, even considering the genomic architecture behind complex genomic rearrangements they can be mistaken for simple rearrangements, such as changes in copy number, due to technical challenges and the limited resolution capabilities of the methods used in structural variation detection (Carvalho & Lupski, 2016).

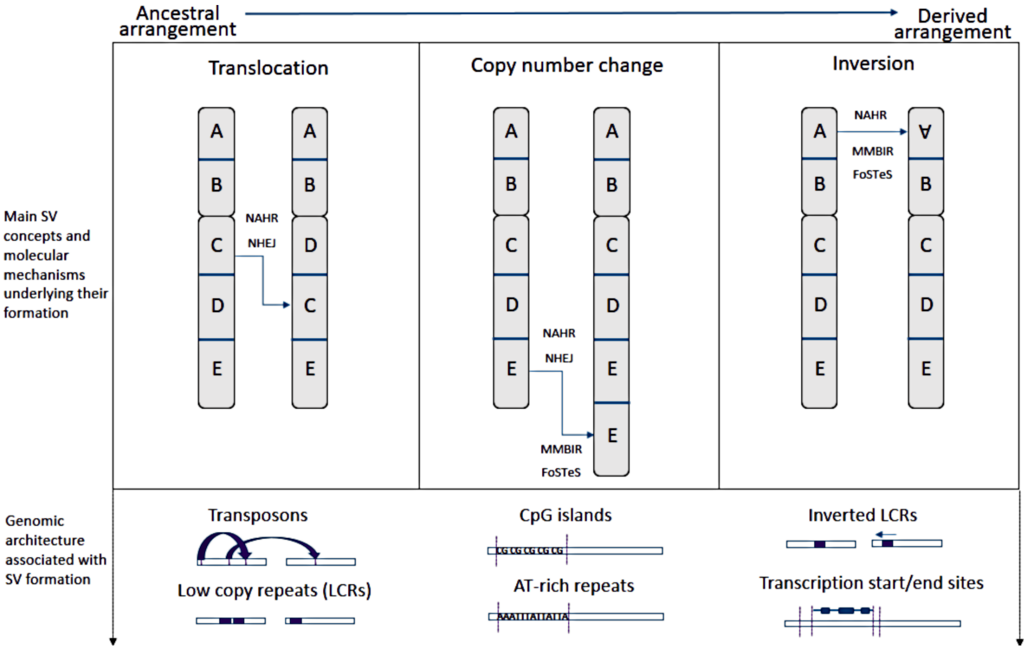


Figure 1.2: Representation of the main structural variant (SV) concepts in the genome. SVs can be formed through replication-based mechanisms (RBMs) such as non-allelic homologous recombination (NAHR) and non-homologous end joining (NHEJ). These two repair mechanisms can generate different kinds of structural variants during replication due to their instability in more complex regions (e.g. in low-copy repeats - LCRs) (Carvalho & Lupski, 2016). There are evidence that CNVs and inversions may be also formed by microhomology-mediated break-induced replication (MMBIR) and fork stalling and template switching (FoSTeS) (Hastings et al., 2009; Zhang et al., 2009). Transposons lead to increased template switching and can consequently promote SV formation (Mayle et al., 2015). AT-rich repeats may be prone to break during replication (Carvalho & Lupski, 2016; Zhang & Freudenreich, 2007; Fungtammasan et al., 2012; Franchitto, 2013), promoting SV formation (Carvalho et al., 2013; Deem et al., 2011). Transcription start and end sites are enriched with CpG islands and both features have been associated with recombination in birds (Singhal et al., 2015).

1.3 Thesis Overview

This thesis explores the structural variants in the genome of a well-studied songbird in ecology and genomics. Popularly known as the great tit, *Parus major* has been investigated for several decades at long-term study sites in the Netherlands and United-Kingdom. Here, using birds from these sites, I explore mainly two classes of structural variants in the great tit genome: CNVs and inversions. I first describe these structural variants, followed by exploring the possible associations with seasonal measurements such as egg-laying date. In **chapter 2** I detected CNVs in a great tit population from the Netherlands and performed a detailed characterization of the genomic architecture, including other structural variant classes such as SDs and transposons, which might underlie CNVs in great tits. Although the biological and technical challenges were evident, it was possible to assess the CNV inheritance patterns and calling confidence in our data-set (e.g. the high number of false negatives calls). Moreover, CNVs were enriched at evolutionary breakpoints, which in turn are enriched for neuron and cardiac related genes. In **chapter 3** I performed a genome-wide association study (GWAS) with egg-laying dates as an individual trait and CNVs. For this, I used the populations from the Netherlands and United-Kingdom. For the population from the Netherlands I used the CNVs detected in **chapter 2** and for the population from United-Kingdom I used the same methods described in **chapter 2** to infer CNVs. CNVs within genes related to circadian clock and reproduction were identified, evidencing the possible effects of CNVs on breeding time. However, CNV-GWAS with quantitative phenotypes have a not well-defined 'gold standard' in the literature (e.g. strategy to define a 'CNV locus' when multiple overlapping CNV calls have distinct breakpoints), sometimes including studies that make use of commercial software (i.e. black boxes). Therefore, I incorporated the CNV-GWAS methodology, which was developed in **chapter 3**, into an open-source R/Bioconductor (Huber et al., 2015) package that is described in **chapter 4**. The package, called CNVRanger, will allow other researchers to perform a CNV-GWAS with a digestible and clear methodology. Moreover, the CNVRanger package includes additional features to deal with downstream analysis of CNVs including methods for summarization (e.g. concatenation of CNV calls into regions) and association with gene expression. To go beyond CNVs, in the **chapter 5** I explored a very large inversion present in ≈5% of the Dutch population which encompasses 90% of Chromosome 1A. The inversion harbors complex breakpoints and evidences a possible gene flux in the center. In the **chapter 6** I show that this large inversion is lethal in homozygotes but it is on balancing selection by a meiotic drive mechanism (i.e. a 'selfish gene').

Chapter 2

CNVs are associated with genomic architecture in a songbird

Vinicius H. da Silva[1,2,3], Veronika N. Laine[2], Mirte Bosse[1], Kees van Oers[2], Bert Dibbits[1], Marcel E. Visser[1,2], Richard P.M.A Crooijmans[1] and Martien A. M. Groenen[1]

[1]Animal Breeding and Genomics, Wageningen University & Research, Wageningen, The Netherlands
[2]Department of Animal Ecology, Netherlands Institute of Ecology (NIOO-KNAW), Wageningen, The Netherlands
[3]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden

BMC Genomics 2018 19:195

Abstract

Understanding variation in genome structure is essential to understand phenotypic differences within populations and the evolutionary history of species. A promising form of this structural variation is copy number variation (CNV). CNVs can be generated by different recombination mechanisms, such as non-allelic homologous recombination, that rely on specific characteristics of the genome architecture. These structural variants can therefore be more abundant at particular genes ultimately leading to variation in phenotypes under selection. Detailed characterization of CNVs therefore can reveal evolutionary footprints of selection and provide insight in their contribution to phenotypic variation in wild populations. Here we use genotypic data from a long-term population of great tits (*Parus major*), a widely studied passerine bird in ecology and evolution, to detect CNVs and identify genomic features prevailing within these regions. We used allele intensities and frequencies from high-density SNP array data from 2,175 birds. We detected 41,029 CNVs concatenated into 8,008 distinct CNV regions (CNVRs). We successfully validated 93.75% of the CNVs tested by qPCR, which were sampled at different frequencies and sizes. A mother-daughter family structure allowed for the evaluation of the inheritance of a number of these CNVs. Thereby, only CNVs with 40 probes or more display segregation in accordance with Mendelian inheritance, suggesting a high rate of false negative calls for smaller CNVs. As CNVRs are a coarse-grained map of CNV loci, we also inferred the frequency of coincident CNV start and end breakpoints. We observed frequency-dependent enrichment of these breakpoints at homologous regions, CpG sites and AT-rich intervals. A gene ontology enrichment analyses showed that CNVs are enriched in genes underpinning neural, cardiac and ion transport pathways. Great tit CNVs are present in almost half of the genes and prominent at repetitive-homologous and regulatory regions. Although overlapping genes under selection, the high number of false negatives make neutrality or association tests on CNVs detected here difficult. Therefore, CNVs should be further addressed in the light of their false negative rate and architecture to improve the comprehension of their association with phenotypes and evolutionary history.

2.1 Introduction

Genetic variants in the genome have been selected over the course of evolution based on their adaptive value under changing environmental conditions but are also affected by random drift (Lynch et al., 2016). These variants range from single nucleotide changes to complex rearrangements in structure (Vitti et al., 2013), which modulate gene expression (Pastinen, 2006; Williams et al., 2007; Bryois et al., 2014) leading to ample phenotypic variation in wild populations (Šťovíček et al., 2014; Vu et al., 2015; Conover et al., 2016). Structural variants show different degrees of complexity, and include copy number variations (CNVs), inversions, insertions, translocations, fissions and fusions (Yalcin et al., 2012; Zhao et al., 2016). A better understanding of these structural variants is essential for detecting important genomic features under selection and their association with phenotypes. In fact, CNVs are known to be major mutations that encompasses more nucleotides than single nucleotide polymorphisms (SNPs) (Redon et al., 2006b) and underlie differences within populations and between closely related species such as human and chimpanzee (Perry et al., 2008).

Although complex rearrangements in the genome which involves combined events like inversions and translocations can be technically challenging and costly to fully characterize (Alkan et al., 2011), CNVs are more easily assessed and be an indication of complex variants (Carvalho et al., 2013). Moreover, CNVs are the raw material for gene family expansion and diversification (Perry, 2008), which ultimately lead to repetitive regions that have an important role in evolutionary breakpoints (Sankoff, 2009). CNVs are usually defined as genomic intervals larger than 1 kilobase (kb) containing deletions or duplications, which can be studied using widely available SNP arrays (Yau & Holmes, 2008). Despite their limited resolution, these SNP arrays are cost effective for studies in large populations (Perkel, 2008) and CNVs can be uncovered by signal variability and heterozygosity level in overlapping SNP probes (Yau & Holmes, 2008).

Species-specific SNP arrays have been used extensively to study CNVs and their association with phenotypes and evolutionary history, in domestic animals (Clop et al., 2012; da Silva et al., 2016), humans (Perry et al., 2006, 2008) and natural populations (Prunier et al., 2017). In mammals, CNVs has been associated with production traits (Prinsen et al., 2017) and pathogen resistance (Liu et al., 2011). Deletions or duplications of genes underpinning inflammatory response and cell proliferation are involved in the phenotypic differentiation of humans and chimpanzees (Perry et al., 2008). An interesting example of phenotypic variation as a result of CNV is the pea-comb phenotype in chicken which is caused by a CNV in intron 1 of SRY-Box 5 (*SOX5*, (Wright et al., 2009)). Interestingly, the number of repeats quantitatively affects this phenotype when in heterozygous state (Moro et al., 2015). Although CNVs are increasingly recognized as source of phenotypic variation, other aspects of CNVs as their inheritance, genomic distribution and rate of false positive or negatives lacks further investigation in large populations.

CNVs usually follow a Mendelian inheritance pattern (Locke et al., 2006), but also de novo mutations have been shown to be functionally relevant and to be associated with a number of diseases (Veltman & Brunner, 2012). Structural rearrangements, like CNVs, result from a number of distinct recombination mechanisms (for a review see (Carvalho & Lupski, 2016)). Such mechanisms like non-allelic homologous recombination or break induced replication prevails at regions in the genome exhibiting specific architecture like segmental duplications and common fragile sites, respectively. Moreover, structural mutability is associated with hypomethylation (Li et al., 2012; Harris et al., 2013) and CpG islands and transcription start and end sites have been shown to be associated with high recombination rates in birds (Singhal et al., 2015).

We identified and studied CNVs in a natural population of great tits (*Parus major*). The great tit is a widely studied passerine bird species in ecology that, in the past decades, has provided important insights into speciation (Kvist et al., 2003), phenology (Perrins, 1970; Visser et al., 1998; Buse et al., 1999), behavior (van der Meer & van Oers, 2015; Fidler et al., 2007) and microevolution (Husby et al., 2011). After completion of the great tit genome sequence (Laine et al., 2016), a customized high density 650k SNP array was developed enabling more detailed genomic studies in this species. We present a CNV analysis in the great tit genome using intensities and allele frequencies from this SNP array. We annotated functional features, accessed mother-daughter inheritance and characterized the genomic architecture underlying different molecular mechanisms, which in turn are known to give rise to different CNV classes. Our study lays the foundations for future studies on complex genetic variants in this population, to better understand genetic variation under global warming and association with shifting seasonal phenotypes.

2.2 Material and methods

Genotype calling and population description

Blood samples of great tits (*Parus major*) were collected from our long-term study populations on the 'Veluwe' area near Arnhem (52°02' N, 5°50' E, the Netherlands). Whole blood samples were stored in either 1 ml Cell Lysis Solution (Gentra Puregene Kit, Qiagen, USA) or Queens buffer (Seutin et al., 1991). DNA was extracted by using the FavorPrep 96-Well Genomic DNA Extraction Kit (Favorgen Biotech corp.). DNA quality and DNA concentration were measured on a Nanodrop 2000 (Thermo Scientific).

A total of 2,648 great tits were genotyped using a custom made Affymetrix® great tit 650K SNP chip at Edinburgh Genomics (Edinburgh, United Kingdom). SNP calling was done following the Affymetrix® best practices workflow by using the Axiom® Analysis Suite 1.1. Nine individuals with dish quality control value of <0.82 were discarded. The length of the probes is 70 bp and more information is available in the raw data submitted to gene expression omnibus (GEO, GSE105131).

Input construction and individual CNV calling

We applied the files denominated 'summary', 'calls' and 'confidences', built during SNP genotyping, to obtain the inputs for CNV detection. These files were used to generate canonical clusters (Peiffer, 2006) by the PennCNV (version 08 Feb 2013) function '`generate_affy_ geno_cluster.pl`', which allowed the estimation of the relative signal intensities (i.e. LRR) and relative allele frequencies (B allele frequency, BAF) by the '`normalize_affy_geno_cluster.pl`' PennCNV function. Using individual BAF values we then estimated the population BAF for each SNP marker, with the `'compile_pfb.pl'` PennCNV function.

As the CG ratio content around each SNP marker is known to influence the signal strength (Diskin et al., 2008), their relative content (1 Mb window) was estimated using the `'nuc'` BEDTools function (Quinlan & Hall, 2010). Therefore, we used the '`genomic_wave.pl`' PennCNV function to adjust individual raw LRR signal values.

To identify the individual CNVs, we applied the '`detect_cnv.pl -test`' for all 31 autosomes. The raw CNVs were filtered out if smaller than 1 kb or supported by less than 3 SNPs. Birds with LRR standard deviation >0.30 or BAF drift >0.02 were also filtered out. A total of 2,175 birds had at least one CNV call after quality control.

Establishment of CNV hotspots and CNV frequency

The genomic regions with at least one individual CNV mapped were defined by the '`reduce`' function from GenomicRanges Bioconductor/R package (version 1.28, (Lawrence et al., 2013)) and then defined as CNVRs. The frequency of each CNVR was estimated based on the number of samples mapped at the genomic interval comprised by the CNVR.

We inferred the frequency of all CNV start and end positions and extend by 5 kb up and downstream these breakpoints. These genomic intervals are defined throughout the text as CNV breakpoint windows and their coordinates were compared with functional and repetitive intervals in the great tit genome.

CNV validation by quantitative PCR

Primers were designed using Primer3plus (Untergasser et al., 2007) and quality testing was performed with NetPrimer (`http://www.premierbiosoft.com/netprimer`).

Samples to be validated were checked for quality based on the amount of dsDNA, which was measured with Qubit® Fluorometer. Subsequently, in each sample we used four different concentrations to determine primer efficiency: 15ng, 7.5ng, 3.8ng and 1.9ng of DNA. Reactions were joined in a final volume of 12.5μl, containing 3.75μl DNA, 6.25μl 2X reaction buffer (MESA Blue from Invitrogen®), 1.25μl forward primer (2μM) and 1.25μl reverse primer (2μM). Samples with CNV and diploid (2n, reference samples) were tested with the designed primer sets. Measurements were performed with the Applied Biosystems® 7500 real-time PCR system. Cycle thresholds (log2 Ct) were corrected based on the efficiency of each primer. ΔCt was calculated as Ct from the sample with a specific CNV minus Ct of the diploid (2n) reference sample (D'haene et al., 2010). The reference sample was given by a random bird with 2n state on the tested region.

Identification of repetitive regions in the great tit genome

To identify masked regions in the reference genome and their respective functionality we applied RepeatMasker (Smit et al., 2013-2015) version open-4.0.6 using the default mode run with cross match version 0.990329. The query species was assumed to be '`aves`'. The regions identified were classified as retroelements, RNA-related regions, DNA transposons and *in-tandem repeats*. Subclassification to define the families within each class was also described when available for a specific class. For simplification, we considered three general families in retrotransposons (short interspersed nuclear elements [SINEs], long interspersed nuclear elements [LINEs] and long terminal repeats [LTRs]) and *in-tandem* repeats (satellites, regions of low complexity and simple repeats). Uncertain family classification was neglected in DNA transposons (e.g. “hAT?” was considered “hAT”).

To identify homologous regions in the great tit genome we used a protocol described elsewhere (Khaja et al., 2006), which applied the megablast greedy algorithm (Zhang et al., 2000) on the great tit reference genome build 1.1 (Laine et al., 2016). We performed all possible comparisons among autosomes and each one against itself to identify inter and intra chromosomal duplications, respectively. We subset regions larger than 1 kb and >90% in sequence similarity, which suggest regions containing recent segmental duplications (Khaja et al., 2006). We filtered out all homologies with more than 10% of its size containing unknown nucleotides ("N") or/and with less than 1 kb of know nucleotides: adenine (A), cytosine (C), thymine (T) or guanine (G).

Functional features and patterns in great tit genome

Thus, we identified genomic intervals containing $[CG]_n$ ($_n = 1$) and TSSs (defined the gene promoters as regions starting 300 bp upstream and ending 50 bp downstream each gene start position, always considering the transcription orientation in each gene). We also identified regions rich in AT ($[AT/TA]_n$ or $[AA/TT]_n$, where $_n$ ≥ 4), due to their role on recombination by break induced replication (Franchitto, 2013). CpG sites and AT-rich intervals were converted into reference genomic ranges (build 1.1, Laine et al. 2016) by `‘vmatchPattern’` function in GenomicRanges Bioconductor/R package (version 1.28, Lawrence et al. 2013). The overlap expected by chance was obtained by simulating genomic tiles of 10 kb with `‘randomizeRegions’` function in regioneR Bioconductor/R package (version 1.80, Gel et al. 2015).

Gene annotation and enrichment analysis

We used gene annotation version 101 from the general feature format (GFF) file from National Center for Biotechnology Information (NCBI) great tit genome 1.1 (`https://www.ncbi.nlm.nih.gov/assembly/GCF_001522545.2`). From 17,545 unique gene names, 16,541 were assigned to autosomal chromosomes which were then used to the subsequent enrichment steps. Gene names were converted to Entrez Ids and subsequently enriched with ‘*enrichKEGG*’ function to identify KEGG pathways; and ‘*enrichGO*’ function to identify GO gene sets overrepresented in all CNVRs and in CNV breakpoint windows present in four birds or more. Both functions, implemented in the *ClusterProfiler* bioconductor R package (version 3.4.1, Yu et al. 2012), used human as the organism (*org.Hs.eg.db* bioconductor R package version 3.4.1, 2017-Mar29, Carlson 2017) due to high accuracy in gene and pathway annotation. The p-values were adjusted by Benjamini and Hochberg method (FDR, Benjamini & Hochberg 1995). The gene background to enrichment of CNV breakpoint windows included just genes up to 5 kb from SNPs (reflecting every 10 kb window around SNPs). To infer the enrichment expected by chance using the same number of genes, we randomly sampled 6,812 genes (total number of unique gene names overlapping CNVRs) 10,000 times and followed the same enrichment process.

Thus, for each significant KEGG pathway in CNVRs, we compared the number of protein/gene names in CNVRs with random enrichments. Therefore, the permutation p-value was based in the number of times that a random enrichment obtained equal more protein/gene names linked to a specific process (times/10,000).

Identification of Syntenic blocks and evolutionary breakpoints

We used the chicken (*Gallus gallus*, Gallus_gallus-5.0) and zebra finch (*Taeniopygia guttata*, taeGut3.2.4) genomes to find sequence synteny with the great tit genome build 1.1 (Laine et al., 2016). All FASTA files were used in the '*FindSynteny*' and '*AlignSynteny*' functions, which are both implemented in the R/Bioconductor package DECIPHER (Wright 2016, version 2.6.0). The synteny blocks were merged by overlap with '*reduce*' function (GenomicRanges Bioconductor/R package, version 1.28, Lawrence et al. 2013). We classified the resulting output into (i) syntenic blocks, (ii) evolutionary breakpoints and (iii) evolutionary breakpoint regions as described previously (Ruiz-Herrera et al., 2006).

2.3 Results

CNV identification, frequency assignment and inheritance

We performed a CNV analysis in great tit genomes using a high density SNP array intensities and allele frequencies from 2,077 females and 98 males. After quality control, 41,029 CNVs were identified which were subsequently merged into 8,008 distinct CNV regions (CNVRs).

The CNVRs cover 28.09% (259.50 millions of base pairs - Mb) of the great tit autosomes. The relative coverage by CNVRs for the different chromosomes ranged from 20.18% for chromosome 14 to 89.30% for chromosome 25LG2. The absolute genomic length overlapped by CNVRs varied from 0.36 Mb for chromosome LGE22 to 40.06 Mb for chromosome 2. The CNVRs had variable sizes ranging from 1.01 kb to 2.83 Mb with a mean size of 32.40 kb. The number of birds with CNVs mapped onto a given CNVR ranged from 1 (0.04%) to 623 (28.63%) of the 2,175 birds analyzed. We found 263 CNVRs to occur in more than 1% of the population ($\geq$ 21 birds) and denote them as 'polymorphic CNVRs' as previously suggested (Itsara et al., 2009).

To investigate CNV inheritance, we used a mother-daughter structure available for 381 mothers and their 625 daughters in this population. We found 460 CNV calls that overlap at least 1 base pair (bp) in the same state (gain or loss) between a mother and at least one of her respective daughters, representing only 6.83% of all 6,733 CNVs identified in the mothers. Thereafter, we classified all CNVs in mothers depending on the number of probes by CNV and found a positive correlation between probe number and inheritance ratio (Pearson's correlation coefficient = 0.62 and p-value $\approx$ 1.68e$-$7). Considering an expected Mendelian inheritance of 50% (all sires in normal state), only CNVs supported by 40 probes or more reach this Mendelian expectancy (for most of the probe groups, Figure 2.1a). Also, CNVs within polymorphic CNVRs showed higher inheritance ratios (367 out of 3,035, 12.09%) but comparable positive correlation with probe number (Pearson's correlation coefficient = 0.60 and p-value $\approx$ 4.254e-06, Figure 2.1b).

Breakpoint variability of overlapping CNVs can unravel molecular mechanisms in their formation and inheritance patterns, which in turn rely on specific patterns in genome architecture (Carvalho & Lupski, 2016). However, there is an unavoidable technical bias in genomic breakpoints of CNVs based on SNP probe intensities (Fadista et al., 2010; Redon et al., 2006b), making it challenging to estimate the frequency of CNV loci. To avoid coarse-grained CNVR breakpoints, which can harbor several CNVs with distinct breakpoints, we tried to improve our description of the breakpoint variability using the number of CNVs sharing the same start or end positions (Figure 2.2). We extended each of these breakpoints by 5 kb up and downstream to establish genomic windows of 10 kb (CNV breakpoint windows). This resulted in 45,372 breakpoint windows identified in 1 to 355 birds. The total of these windows represents 254.14 Mb of the genome, which the large majority (224.38 Mb) reflects rare events (frequency = 1).

Copy number inference by quantitative PCR

To obtain insight in the false discovery rate of our method to identify CNVs, we validated 16 CNVs in our great tit population using quantitative PCR (qPCR). We selected 6 rare and 10 frequent CNV calls based on CNV incidence, size and state. Concerning incidence, we selected CNVs identified in only one bird, those present in two and those present in four to five birds (all with exactly the same breakpoints). Within each frequency class we tried to choose different sizes of events. Concerning state, in each frequency class we ensured the inclusion of at least one CNV belonging to each of the most common states (1n and 3n). The size of the CNVs chosen for validation ranged from 1.06 to 77.12 kb, and are located within CNVRs ranging from 1.06 to 494.36 kb. The number of SNPs supporting these CNVs ranges from 3 to 19. The gain or loss of specific genomic intervals, detected by PennCNV, was confirmed by qPCR for 15 of these 16 CNVs (93.75%). However, we observed discrepancies in the copy number based on PennCNV and qPCR. Considering exactly the same state (i.e. copy number between one and four), 9 out of the 16 CNVs (56.25%) showed

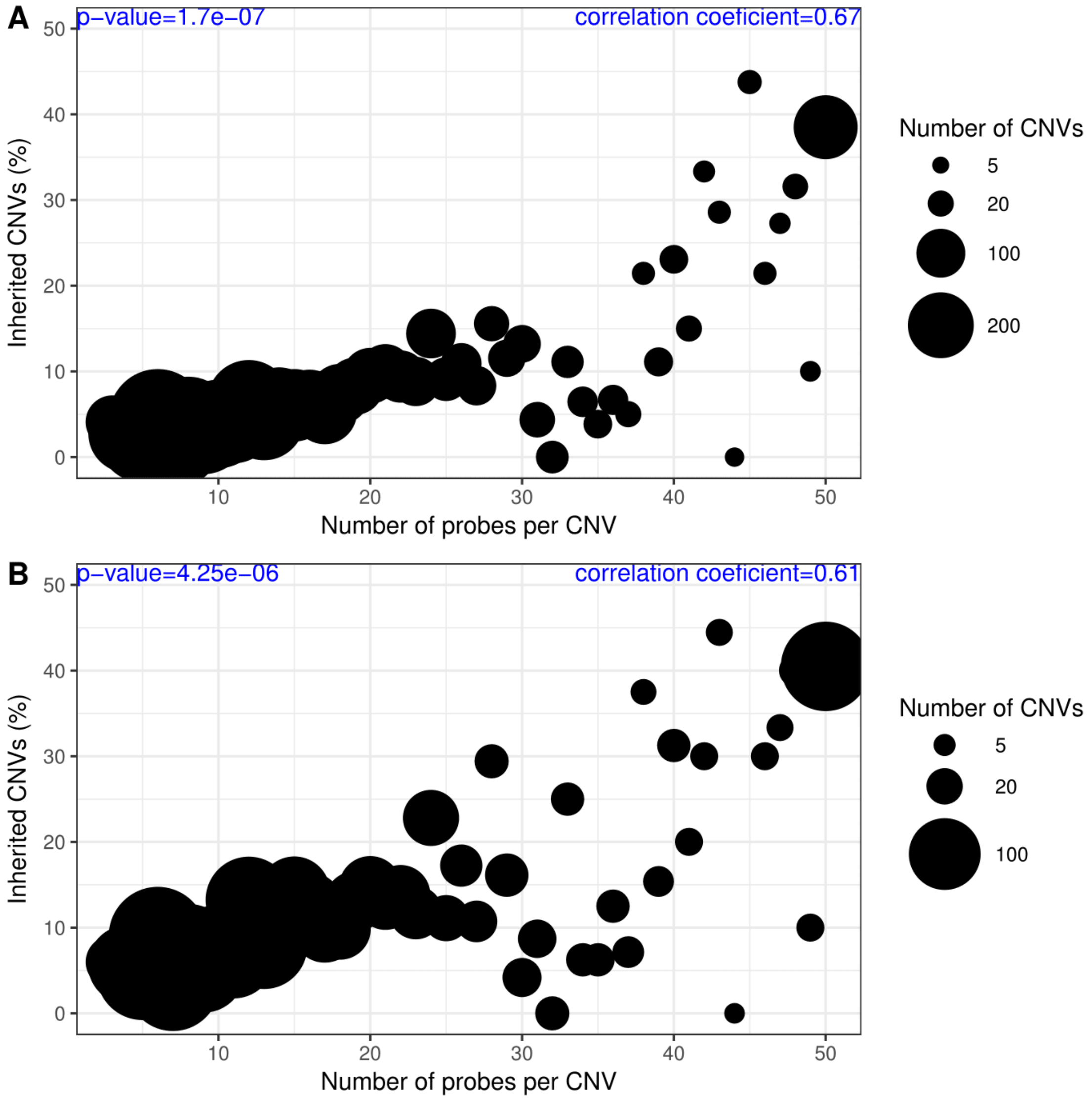


Figure 2.1: CNV inheritance in mother-daughter family structure. We inferred the percentage of CNVs in mothers overlapping CNVs at the same state (gain or loss) in their respective daughters. The x-axis indicates distinct groups of CNVs which were classified based on the number of SNP probes supporting each of them. CNVs supported by 50 SNP probes or more are grouped together. In the y-axis the percentage of inherited CNVs represents the ratio between all CNVs and inherited ones in each probe group. The number of CNVs per group is reflected by the dot size. A: All CNVRs. B: Polymorphic CNVRs ($\geq$ 21 birds, at least 1% of the population with CNVs identified).

the same number of copies using these two methods.

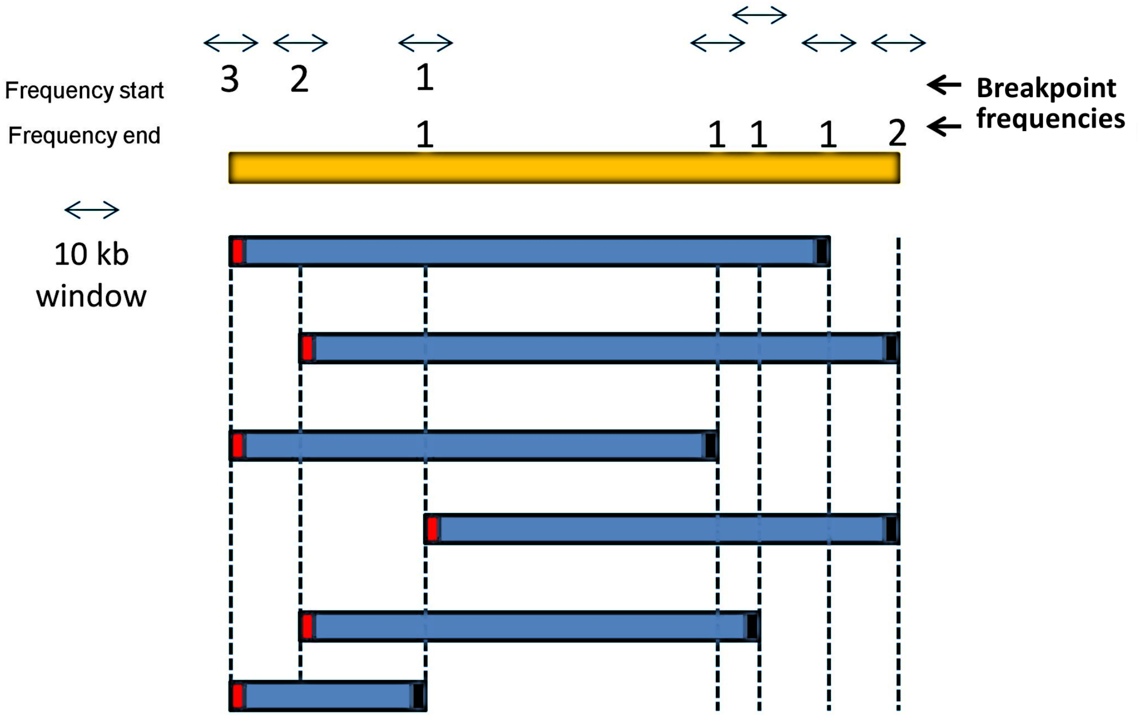


Figure 2.2: CNVR example and the strategy to estimate the frequency of CNVs which are sharing breakpoints. The frequency for a given genomic interval is given by the number of CNVs starting or ending at certain SNP probes. All the windows around the breakpoints have 10 kb and may have one frequency for the common start positions and one for the end positions.

Repetitive and functional intervals in CNVs

We evaluated five different sequence features in the great tit genome for their overlap with CNV breakpoint windows: (I) Homologous regions, (II) Interspersed repeats and low complexity DNA sequences, (III) CpG sites, (IV) Transcription start sites (TSSs) and (V) AT-rich regions.

It has been shown that homologous regions reflect segmental duplications and promote CNV formation (Khurana et al., 2010). In order to study this in great tits we identified large homologous regions ($\geq$ 1 kb and at least 90% sequence identity) using megablast (Zhang et al., 2000). We identified 3.44Mb of the automosomes as homologous regions (0.37%), representing 1,111 intra- and 879 inter-chromosomal homologies respectively (Table 2.1). The breakpoints observed at very low frequency ($\leq$ 2) are not correlated with the occurrence of homologous sequences whereas the more frequent ones (>3) show progressively more overlap with homologous regions (Figure 2.3A). The sequence identity of the homologies is also correlated with breakpoint frequency. Homologous regions with higher sequences identity tend to overlap more with CNV breakpoints with a frequency equal or more than four (Figure 2.4), in agreement with previous studies in human and chimpanzee describing an excess of CNVs at regions with high sequence homologies (Perry et al., 2008).

Table 2.1: Homologous regions in the great tit genome with more than 90% of sequence identity and respective proportions of intra and interchromosomal homologies.

Homology	Number of regions	Total size (Mb)	Similarity (+-SD)
Intrachromosomal	1111	2.66	92.97+-2.26
Interchromosomal	879	1.58	92.78+-2.1
All	1512	3.44	92.89+-2.25

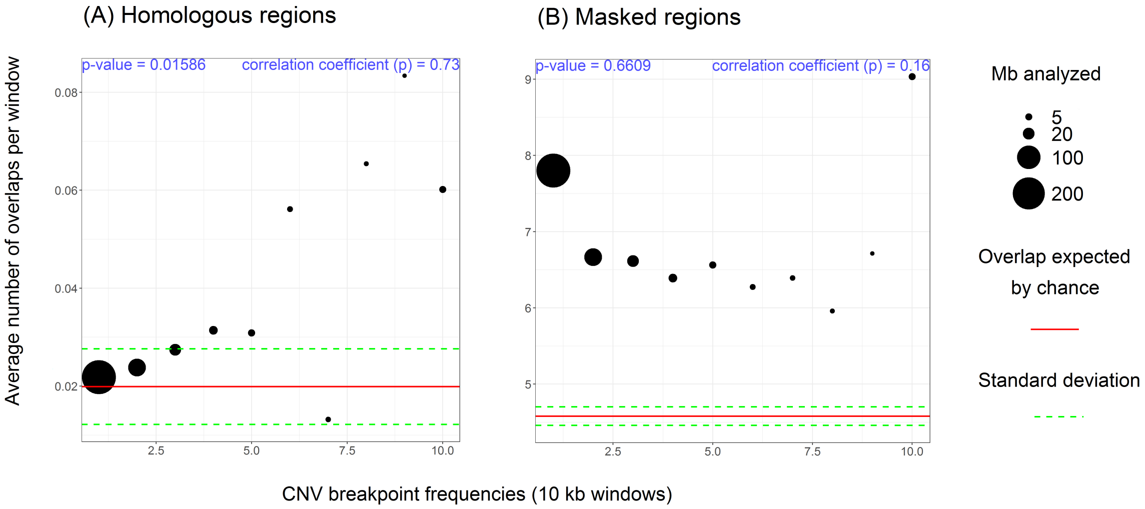


Figure 2.3: Overlap of CNV breakpoints with repetitive regions in the genome. CNV breakpoints with 10 in frequency or more are grouped together. A: Homologous regions with more than 90% in similarity and 1 kb. B: Masked regions as retroelements, RNA-related regions, DNA transposons and *in-tandem* repeats.

In addition to the homologous regions, we identified repetitive elements masked by RepeatMasker (Smit et al., 2013-2015). These elements represent 6.16% (56.92 Mb) of the total length of the great tit autosomes. We found 400,503 masked regions, representing mainly retroelements (145,689; 43.06 Mb), *in-tandem* repeats (240,115; 11.54Mb) and DNA transposons (13,374; 1.95 Mb). All frequencies of CNV breakpoints (Figure 2.2) overlap masked regions more than expected by chance, but there was no correlation between the overlap and frequency (correlation coefficient = 0.16, p-value = 0.66, Figure 2.3B).

Noteworthy is that although homologous and masked regions show substantial overlap, their distribution differs. Intervals covered by both features (i.e. intersection) are considerably smaller than the regions overlapped in each of them. From 1,512 homologous regions, 1,302 (3.13 Mb; 91%) overlap intervals masked by RepeatMasker

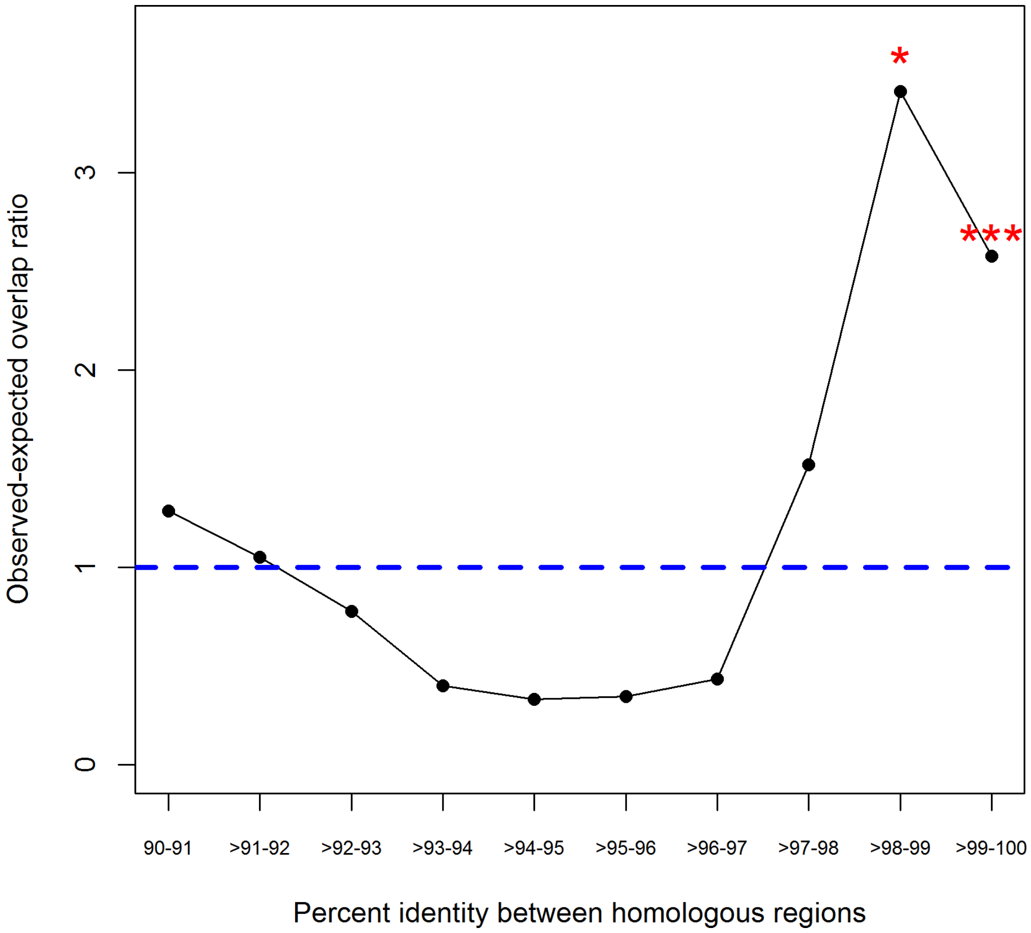


Figure 2.4: Colocalization of CNV breakpoints (10 kb windows with ≥4 in frequency) and homologous regions binned by sequence identity. The y-axis depicts the ratio between observed and expected number of overlaps (based on 10,000 randomic simulations) between CNV breakpoints and homologous regions. Homologous regions are placed in one of the bin classes in the x-axis which are based on inter- or intrachromosomal percent identities. Permutation p-values are based on the number of random simulations that obtained more overlaps than observed (*$\leq$ 0.05 and ***$\leq$0.001).

(Smit et al., 2013-2015) by at least 1 bp. From 397,537 masked regions, 2,594 (1.24 Mb; 2.18%) overlap homologous regions by at least 1 bp. However, only 985 kb is covered by both (31.5% and 1.73% of the total length in homologous and masked regions respectively).

Genomic regions which are rich in CpG sites and TSSs show a high recombination rate in birds (Singhal et al., 2015). Thus, we inferred these two features to understand the association of highly recombinant regions with CNVs. We identified 6,861,240 CpG sites in the great tit autosomes, ranging from 12,725 on chromosome LGE22 to 845,266 on chromosome 2. All CNV breakpoints windows contain more

CpG sites than expected by chance and the number of sites increases along with the breakpoint-frequency (correlation coefficient = 0.59, p-value = 0.00017, Figure 2.5A). Similarly, TSSs have positive overlap correlation with CNV breakpoint frequencies (up to 50% of breakpoints with frequency $\geq$15 overlap with TSSs, Figure 2.5B). Results from CpG sites and TSSs are expected to be comparable given the known high prevalence of CpG islands at TSSs (Singhal et al., 2015; Derks et al., 2016).

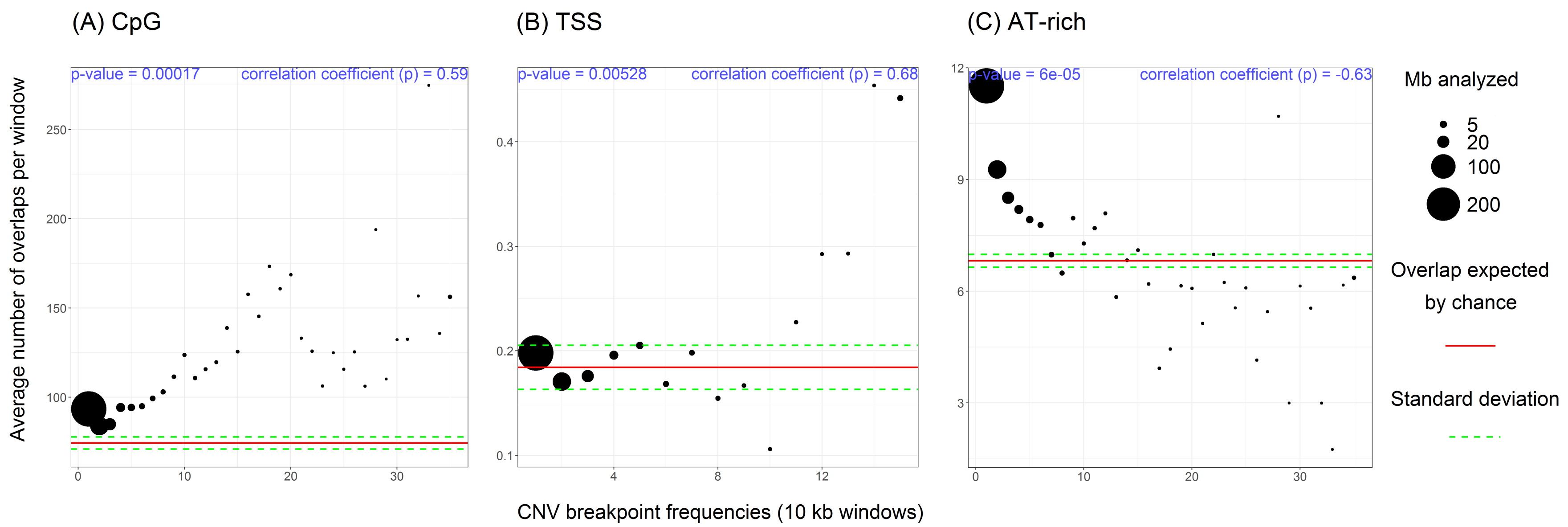


Figure 2.5: Overlap of CNV breakpoints with functional features and regions prone to breakage. A: CpG sites. B: Transcription start sites (TSSs). C: AT-rich intervals. CNV breakpoints observed in 30 birds or more are grouped together for CpG and AT-rich features. Otherwise, in TSSs we grouped those with 10 or in frequency because most of high frequent CNV breakpoints are small groups and can impair confident comparison with more scarce features as TSSs (in comparison with CpG or AT-rich sites).

AT-rich intervals have been reported at genomic regions known to be prone to breakage, thereby allowing complex rearrangements (Carvalho et al., 2013). Thus, we identified 629,840 AT-rich intervals, of which the majority is 8 bp in size but that can be up to 100 bp in size. CNV breakpoint frequencies have a strong negative correlation with AT-rich intervals (Figure 2.5C).

To verify a possible technical bias underlying the observed correlations, we evaluated the correlation between signal variability in SNP probes outside our CNVRs and the GC ratio of the region. The GC ratio could be relevant because it can lead to a so-called GC wave (Diskin et al., 2008), which is a well-known bias in the detection of CNVs from SNP-arrays (causing variation in hybridization intensity). We inferred the Log R Ratio (LRR) values in non-CNV probes and estimated its standard deviation median for each tile of 10 kb in the genome. We correlated these medians with the GC ratio and found a very low positive correlation coefficient (0.02; *p*-value=0.059) with the LRR standard deviation (SD) median. This low correlation is expected because we corrected all LRR values for this GC wave before CNV detection.

Gene enrichment and functional analysis

The genomic coordinates of all 8,008 CNVRs identified overlap with 6,857 of the 16,541 annotated unique genes (41.45%) for great tit (build 1.1 Laine et al. 2016). Using these overlapping genes we performed an enrichment analysis looking for pathways (Kyoto encyclopedia of genes and genomes, KEGG) and gene ontology (GO) gene sets prevailing in genes located within (i) CNVRs and (ii) CNV breakpoints seen in at least four birds.

Proteins of genes overlapping CNVRs were significantly overrepresented for 15 KEGG biological pathways (Table 2.2, which are mostly related to neuronal and cardiac processes. All significant KEGG pathways were compared with 10,000 random enrichments and we found all processes enriched in CNVRs with permutation *p*-value $\leq$ 0.001. In accordance with KEGG results, we found 77 GO gene sets mostly related with neuronal, cardiac and ion transport pathways. The GO gene sets with lowest *p*-values where synaptic membrane, postsynapse and postsynaptic membrane respectively.

In order to determine if similar enrichment is also reflected in more frequent CNVs, we performed the gene enrichment using the CNV breakpoint windows (frequency $\geq$4, subset analyzed in the Figure 2.4). These CNV breakpoints overlap 1,012 genes which are enriched for five KEGG pathways and six GO gene sets, as presynaptic active zone, homophilic cell adhesion and neuron recognition. From these 1,012 genes, a subset of 68 overlap homologous regions in the great tit genome, 18 have

Table 2.2: Biological pathways enriched at CNVRs in the great tit genome.

ID	Description	Number of proteins	Ajusted p-value	Protein ratio
hsa05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	59	5.15×10^{-6}	0.728
hsa04020	Calcium signaling pathway	126	1.16×10^{-4}	0.583
hsa04360	Axon guidance	127	3.99×10^{-4}	0.57
hsa04724	Glutamatergic synapse	78	8.2×10^{-4}	0.609
hsa04514	Cell adhesion molecules (CAMs)	75	8.2×10^{-4}	0.638
hsa04925	Aldosterone synthesis and secretion	60	8.2×10^{-4}	0.61
hsa04713	Circadian entrainment	67	3.1×10^{-3}	0.604
hsa00220	Arginine biosynthesis	19	3.15×10^{-3}	0.826
hsa04970	Salivary secretion	48	1.34×10^{-2}	0.615
hsa04022	cGMP-PKG signaling pathway	105	1.73×10^{-2}	0.591
hsa05410	Hypertrophic cardiomyopathy (HCM)	55	1.73×10^{-2}	0.536
hsa04740	Olfactory transduction	29	1.73×10^{-2}	0.674
hsa05010	Alzheimer's disease	78	3.84×10^{-2}	0.545
hsa04750	Inflammatory mediator regulation of TRP channels	60	4.92×10^{-2}	0.561
hsa05414	Dilated cardiomyopathy	57	4.92×10^{-2}	0.564

ID = pathway identification code; Description = pathway name; Number of proteins = number of protein names with genes overlapping CNVRs; Adjusted *p*-value = enrichment FDR corrected *p*-value; Protein ratio = ratio between protein names with genes in CNVRs and all protein names assigned to a specific pathway.

SNP alleles previously described as under selection (Laine et al., 2016) and five overlap homologous regions and are under selection concomitantly.

Genome Synteny with zebra finch and chicken at great tit CNVRs

We compared the great tit genome with the genomes of chicken and zebra finch to identify synteny blocks. For the great tit-chicken comparison, we found 13,437 blocks in synteny ranging in size from 181 bp to 2.15 Mb. The number of blocks varied from 11 on chromosome LGE22 to 1,921 on chromosome 2. For the great tit-zebra finch comparison, we found 5,141 synteny blocks ranging in size from 182 bp to 6.19 Mb. The number of blocks varies from 18 on chromosome LGE22 to 605 on chromosome 2.

We then inferred to what extent the identified CNVs overlap with evolutionary breakpoints and whether this overlap differs from overlap with regions randomly chosen within the genome. We found 3,090 CNVRs (38.58%) overlapping evolutionary breakpoints (with chicken and zebra finch concomitantly), a number that is consistently higher than expected by chance (*p*-value 9.99e-05). We observed 7,022 genes overlapping the evolutionary breakpoints, which are enriched for biological pathways mostly related to neuronal and cardiac processes. At least eight genes that have previously been reported (Volker et al.) to be located at CNV regions in chicken and four in zebra finch overlap evolutionary breakpoints.

2.4 Discussion

Most studies have focused on single nucleotide changes when studying genetic associations with phenotypes and evolution. However, also variation in genomic structures such as CNVs are shown to be associated with a wide range of phenotypes (Clop et al., 2012; Weischenfeldt et al., 2013) and evolutionary phenomena like speciation (Perry et al., 2006, 2008; Paudel et al., 2015) and adaptation (Kondrashov, 2012; Qian & Zhang, 2014). We here therefore used a high density SNP array to identify CNVs as well as their inheritance and architecture in the great tit genome. We detected CNVs covering a large percentage (28.09%) of the great tit genome. Because CNV identification based on SNP Affymetrix arrays are prone to high false discovery rates, we used the mother-daughter family structure of our data to access relative CNV confidence. The relative number of inherited events is higher for CNVs supported by more SNP probes, especially for CNVs with more than 40 probes. The low inheritance of the shorter CNVs suggests a relative high false negative call rate. On the other hand, most of the CNVs tested by qPCR were successfully validated (15/16) and all of these had less than 25 probes suggesting a low false positive call rate of the Affymetrix array. Regarding the exact number of copies, the disparity between SNP-array and qPCR results can be explained by the inherent resolution of each technology. SNP-array data have limited power to infer the exact number of copies whereas qPCR may be considered a gold standard and consequently is more reliable to infer the number of copies.

We evaluated the overlap pattern of CNVs with five genomic features that have known role in structural variation formation and recombination: (i) Homologous regions, or segmental duplications, which support CNV formation through non-allelic homologous recombination (Sharp et al., 2005; Carvalho & Lupski, 2016). (ii) Repetitive features like transposable elements and retrotransposons which account for a substantial fraction of copy-number differences (Schrider et al., 2013; Dennenmoser et al., 2017) and mutually explain recent and ongoing phenotypic adaptation (Schmidt et al., 2010). (iii) Functional CpG and (iv) TSSs that harbor high recombination rate in birds (Singhal et al., 2015). (v) AT-rich regions are prone to break and subsequently produce complex rearrangements (Carvalho & Lupski, 2016; Zhang & Freudenreich, 2007; Fungtammasan et al., 2012; Deem et al., 2011; Carvalho et al., 2013). All these five genomic features display non-random overlap with CNVs and their breakpoint frequencies.

Homologous regions, at least one kb in size and with at least 90% of sequence identity, reflect recent segmental duplications in the genome (Khurana et al., 2010) and can increase the chance of a triplication event in subsequent generations by more than 100-fold (Liu et al., 2014). Thus, apart from positive selection or drift, the

CNV frequency may have increased due to a higher rate of rearrangement at these genomic intervals. We find a significant positive correlation between, CNV breakpoints seen in at least four birds, and regions containing segmental duplications. How similar these genomic homologies are, is also determinant for CNV formation and can reveal its evolutionary history (Perry et al., 2008). Over time, duplicated regions that are fixed decrease in identity, which consequently decreases the chance of recombination mechanisms, such as non-allelic homologous recombination, to act upon them (Bailey & Eichler, 2006). Therefore, CNVs arising from this mechanism are relatively rarer at duplications with lower homology. This is reflected by the increasingly overlap of CNV breakpoints (frequency $\geq$4) and homologous regions with higher sequence identity.

Most of homologous regions overlap repetitive elements masked in the genome, like transposable elements. However, both features display different genomic length distribution and coverage. Repetitive elements cover around ten times more nucleotides, but are usually smaller in length when compared with overlapping homologous regions. In addition, masked regions overlap CNV breakpoint windows more than expected by chance but do not differ between breakpoint frequencies like homologous regions. The number of transposable elements in the great tit genome is comparable with other bird genomes, but they cover a relatively smaller fraction of the whole genome sequence length. The relative coverage in great tit is 1.24% whereas other bird species vary from 4.1 to 9.8% (Hillier et al. 2004; Warren et al. 2010; Zhang et al. 2014a, for a review see Kapusta & Suh 2016). The coverage of transposable elements found here for the build 1.1 is comparable to previous version of the genome (2.06 Mb in this study and 1.95 Mb previously in Laine et al. 2016). Remarkably, transposable elements in great tit genome display distinct CpG hypermethylation between tissues, albeit their expression is correlated only with non-CpG methylation (Derks et al., 2016).

We also evaluated whether the CNV breakpoints are positively correlated with the presence of functional sequences like CpG sites and TSS. It has been shown that in birds recombination prevails at transcription start or end sites and CpG islands (Singhal et al., 2015). The overlap of CpG sites and TSSs with CNV breakpoints increases with breakpoint frequencies in this great tit population. This result suggests a higher CNV mutation rate at these regions, although it is complex to disentangle mutation rate from selection of the CNVs at these regions.

AT-rich intervals have repeatedly been reported as common fragile sites (Carvalho & Lupski, 2016; Zhang & Freudenreich, 2007; Fungtammasan et al., 2012), which are more prone to break induced replication (Franchitto, 2013). This mechanism has a high risk of undergoing template switching (Carvalho et al., 2013; Deem et al., 2011), resulting in complex structural variants. Therefore, as AT-rich intervals are expected to easily break during meiosis, each meiosis breakage might produce CNVs with distinct breakpoints and gene content in the population (Carvalho & Lupski, 2016). CNV breakpoint frequencies in this great tit population are negatively correlated with AT-rich sites, in agreement with the expectancy that lower number of CNVs will share breakpoint positions among individuals in fragile sites throughout genome.

We also performed a functional enrichment for genes within (i) CNVRs and (ii) CNV breakpoints seen in at least four birds. A large proportion of the great tit genes overlaps with CNVRs (41.76%) and these CNV breakpoints (6.12%). Although CNVRs overlap almost seven times more genes, pathways in CNVRs as well as in these CNV breakpoints were enriched to neuronal processes and structure like axion guidance and glutamatergic synapse; cardiac or muscular processes like arrhythmogenic right ventricular cardiomyopathy and calcium signaling. Interestingly, genes related to neuronal functions were previously shown to be under positive selection in great tit (Laine et al., 2016). Moreover, a comparative CNV analysis among different bird species such as chicken, turkey and common quail found a gain in leucine rich repeat and fibronectin type III domain containing 5 (*LRFN5*), which is involved in presynaptic differentiation, to occur just in quails (Skinner et al., 2014). In this great tit population, *LRFN5* is located within CNVR7101 (frequency $\geq$5.4%) that harbor gains and losses. Calcium signaling, that is also enriched in great tit CNVRs, is a key process in neuronal physiology mainly due to its role on neuron buffering (Blaustein, 1988) and in muscle activity by troponin-tropomyosin complex (Stewart & Levy 1970, for a review on calcium signaling see Clapham 2007). However, the high rate of false negative of the CNVs identified here hampered efforts to find which genes are under selection, or that display high LD with SNP alleles at genes previously found to be under selection (Laine et al., 2016).

We identified a median of 12 CNVs per bird, which is comparable to 11.75 found by Skinner et al. (Skinner et al., 2014) that evaluated different bird species, which in turn is comparable to the situation in mammals (Skinner et al., 2014). The same study also claimed that CNVRs in birds could have a slightly higher association with genes than in mammals, but the limited number of samples prevented a more robust conclusion at that time. Here we found 66% of the CNVRs harboring genes, value that increases to 78.3% when considering only polymorphic CNVRs. These proportions are comparable with the 70% that has been found previously (Skinner et al., 2014). Therefore, the large population analyzed here plus the prevalence of bird CNVs on genes may explain the striking proportion of 41.45% great tit genes with CNVs.

To shed light on the evolutionary implications of CNVs and their associated genomic architecture, we compared the great tit genome with the genomes of two other birds: chicken and zebra finch. As expected, because of the higher evolutionary proximity we found a higher degree of synteny between the two songbirds, great tit and zebra finch. The overrepresentation of CNVs at evolutionary breakpoints suggests a critical role in speciation. Moreover, we found biological pathways that are related to neuronal and cardiac processes enriched in both CNVs and evolutionary breakpoints. Syntenic regions among zebra finch and chicken with known CNVs harbor at least nine genes that are at evolutionary breakpoints. These genes are involved in signalling and neuronal pathways.

2.5 Conclusions

CNVs can be challenging to detect and interpret using SNP arrays due to biological and technical variability. The qPCR validation and the intrinsic genomic architecture of the CNVs identified here point to a substantial number of false negatives. The genomic features enriched in CNVs (homologous regions, masked regions, CpG sites, TSSs and AT-rich intervals) support specific mechanisms of the formation of CNVs. Moreover, CNVs are enriched at evolutionary breakpoints, neuron and cardiac related genes and a subset harbors SNP alleles under selection (Laine et al., 2016). Therefore, we expect the CNVs identified here to be valuable for future studies on the great tit genome, but the non-random distribution and inheritance patterns of CNVs indicate that they should be interpreted in the light of their genomic architecture and false negative rate.

Chapter 3

Genome-wide association of copy number variation with egg-laying date in a wild songbird

Vinicius H. da Silva[1,2,3], Dirk-Jan De Koning[3], Phillip Gienapp[2], Veronika N. Laine[2], Kees van Oers[2], Jon Slate[4], Richard P.M.A. Crooijmans[1], Martien A. M. Groenen[1], Anna M. Johansson[3], Marcel E. Visser[1,2]

[1]Animal Breeding and Genomics, Wageningen University & Research, Wageningen, The Netherlands
[2]Department of Animal Ecology, Netherlands Institute of Ecology (NIOO-KNAW), Wageningen, The Netherlands
[3]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden
[4]Department of Animal and Plant Sciences, The University of Sheffield, Sheffield, England

Submitted

Abstract

Timing of breeding is a life-history trait that has clear effects on reproductive success, especially in species living in seasonal environments. In songbirds, such as great tits (*Parus major*), there is genetic variation in egg-laying date but the underlying genomic variation is poorly understood. Recently, the association between egg-laying date and single nucleotide polymorphisms (SNPs) has been studied, but whether structural variants, such as copy number variation (CNV), explain variation in egg-laying date has not yet been explored. Therefore, we explored the relationship between CNVs and egg-laying date for two long-term study population of great tits (in the Netherlands and the United-Kingdom) but we did not find such an association. However, two association analyses, which independently used (i) CNVs and (ii) their raw intensity signals (i.e. log R ratios), concomitantly highlighted suggestive regions harboring genes related to egg-laying dates. These genes are associated with traits that may play a role in seasonal timing such as circadian clock, reproductive success and mammalian pregnancy.

3.1 Introduction

Organisms living in seasonal environments need to time their breeding so that it coincides with the often short period when conditions, often set by food availability, are favourable (Hau, 2001; Visser et al., 2006; Munguía-Rosas et al., 2011). This seasonal timing (or phenology) is in most species temperature dependent. Increasing temperatures due to global climate change have repeatedly led shifts in phenology but these shifts are often occurring at different rates across species at different trophic levels (Visser & Both, 2005). These results in phenological mismatches between trophic levels (Thackeray et al., 2016) and consequently to selection on consumer phenology (Visser et al., 1998). A textbook example of phenological mismatch is between peak caterpillar abundance and egg-laying date of the great tit (*Parus major*, Visser et al. 2012), a songbird that has been extensively studied in ecology and evolution and has excellent genomic resources (Laine et al., 2016; Kim et al., 2018). Due to the mismatched egg-laying date is under directional selection in great tits but because egg-laying date is a complex trait, affected by both genetic and environmental factors (Noordwijk et al., 1980; Gienapp et al., 2005; Wilkin et al., 2007), the response to this selection is difficult to forecast. The heritability of egg-laying dates ranges from low to moderate (i.e. h^2 from 0.14 to 0.4) depending of the average temperature in the spring preceding a breeding season (Husby et al., 2011; Gienapp et al., 2017). Insight into the genomic variation underlying phenotypic variation in egg-laying date will contribute to a better understanding on how animals in the wild can adapt to their changing world. Therefore, identification of the genetic variants that are associated with timing of egg-laying will help our understanding of the molecular mechanisms underlying breeding timing in great tits, and the way selection may act on this mechanism.

Genome-wide association study (GWAS) is a common method to link phenotypic variation to genomic variation, and evolutionary studies usually focus on genotypes at single nucleotide polymorphisms (SNPs) as the source of genomic variation (Morin et al., 2004). However, structural variants have been increasingly linked with a wide range of phenotypes in humans (Ionita-Laza et al., 2009), livestock (Clop et al., 2012) and wild populations (Prunier et al., 2017). Among these structural variants, copy number variations (CNVs) are commonly studied and can be classified as deletions or duplications of genomic intervals larger than one kilobase (kb, Feuk et al. 2006). In the great tit, the association of genetic variants with egg-laying date has been addressed by a SNP-based GWAS in an environment-dependent manner (Gienapp et al., 2017) whereas CNVs in this species were only used to investigate genomic architecture (da Silva et al., 2018). Moreover, as the SNP-based GWAS (Gienapp et al., 2017) have not convincingly found genes associated with egg-laying date, CNVs might be worth exploring as an alternative source genetic variability.

Pioneering studies of CNV associations with phenotypes have been performed on psychiatric disease risk in humans (Joober & Boksa, 2009; Chao et al., 2009; Morrow, 2010; Levy et al., 2012; Green & MacLeod, 2016; Kendall et al., 2017) and the medical importance of CNVs on cognition is clear (Kirov, 2015). Moreover, several other human traits and diseases, such as HIV susceptibility (Gonzalez, 2005) and Hemophilia (Antonarakis et al., 1995) have been also linked with changes in copy number, which, with further validation, could be used for diagnosis and personalized medical treatments. In livestock animals several diseases, such as osteopetrosis in cattle (Meyers et al., 2010) and intersex syndrome in sheep (Pailhoux et al., 2001), as well as production traits such as meat tenderness (da Silva et al., 2016) and milk production (Xu et al., 2014) in cattle have also been associated with CNVs. Thus, breeding programmes in different livestock species may increasingly make use of CNV information to decrease the incidence of genetic disorders and speed up the genetic gain.

Although well studied in humans and livestock, to the best of our knowledge, studies on phenotype-CNV associations are rare in other wild species (e.g. Prunier et al. 2017) and have never been performed in great tits. Thus, in wild species, understanding the effects of CNVs on ecologically relevant phenotypes (e.g. breeding timing) could improve our understanding of the genetics underlying natural phenotypic variation. CNV-related genetic variation may not be detected in a traditional GWAS, unless strong CNV-SNP linkage-disequilibrium exists, and this problem has been particularly ignored in the molecular ecology/ecological genetics literature. The low number of CNV association studies in wild animals may be partially attributable to a lack of 'gold-standard' CNV-GWAS protocols, because CNVs can show a complex technical/biological variability. Most studies make use of heterogeneous in-house association strategies or generic paid software (i.e. 'black boxes'). Moreover, current efforts are largely focused on rare variants, which may have limited effect on the fitness of wild species. Thus, to address these limitations we implemented further developments to the comprehensive open-source R/Bioconductor package CNVRanger, which may allow digestible, customized and reproducible CNV-GWAS. Thus, we used this package to explore population-specific associations of CNVs with egg-laying date using long-term studies of wild great tit populations from both the Netherlands and the United Kingdom.

3.2 Material and methods

Population description and genotype-CNV calling

We used great tits from long-term study sites in the Veluwe area in the Netherlands (NL) and Wytham Woods in the United Kingdom (UK). A total of 2,648 birds from NL and 1,736 from UK were genotyped at Edinburgh Genomics (Edinburgh, United Kingdom) on a custom made Affymetrix® great tit 650K SNP chip (Kim et al., 2018). We identified CNVs in these populations based on SNP probe intensities (log R ratios - LRRs) and allele frequencies (B allele frequencies - BAFs) with the PennCNV software (Wang et al., 2007). Detailed procedures for genotyping and CNV detection/filtering are described in our previous CNV study on the NL population (da Silva et al., 2018). After quality control, we identified a total of 2,175 NL and 1,349 UK birds with at least one CNV. From all birds with CNV information, 2,133 NL and 268 UK birds were also phenotyped for egg-laying date and therefore used for the genome-wide association analysis with CNVs.

We compared the great tit CNV data-sets identified in the NL (da Silva et al., 2018) and two independent studies in the UK population (i.e. (i) reported here and (ii) CNVs previously reported in an independent study by Kim et al. 2018). To display the genomic intersection (i.e. common genomic intervals overlapped by CNVs) we used `GRange` objects (Lawrence et al., 2013), harboring CNV ranges belonging to each respective data-set, into the `UpSet` function that is implemented in the ComplexHeatmap Bioconductor/R package (version 1.20, Gu et al. 2016).

Gene annotation and enrichment analysis

We used gene annotation version 101 from the general feature format (GFF) file from National Center for Biotechnology Information (NCBI) great tit genome 1.1 (`https://www.ncbi.nlm.nih.gov/assembly/GCF_001522545.2`). Of the 17,545 unique gene names, 16,541 could be assigned to autosomal chromosomes which were then used in the enrichment analysis. To identify KEGG pathways for all CNVRs identified in the UK population (the NL population was previously analyzed in da Silva et al. 2018), great tit gene names were converted to human Entrez Ids with `bitr/bitr` kegg and subsequently analysed for enrichment with `enrichKEGG` functions. These functions were implemented in the ClusterProfiler Bioconductor/R package version 3.4.1 (Yu et al., 2012). We used *Homo sapiens* as the organism in the enrichment analysis (i.e. *org.Hs.eg.db* Bioconductor/R package version 3.7, Carlson 2017) due to a high accuracy in gene and pathway annotation. The p-values were adjusted by false discovery rate (FDR), also known as the Benjamini and Hochberg method (Benjamini & Hochberg, 1995).

Linkage-disequilibrium between SNPs and CNVs

We used previously inferred SNP genotypes, which were filtered in the NL population (for details on SNP genotype calling see da Silva 2019 et al. IN REVIEW) to scan for linkage-disequilibrium (LD) between SNPs and CNV segments. The method to infer CNV segments is explained in detail in the section about genome-wide association. First, we identified all SNPs in a genomic window of 1 Mb up- and downstream from each CNV segment breakpoint (i.e. start and end, respectively). Then, the r^2 and adjusted p-values (i.e. q-values) for each pairwise comparison between CNV segments and neighbouring SNPs were obtained with the `calculateLDSNPandCNV` function in the CNVrd2 Bioconductor/R package (Nguyen et al., 2014). As a default, the q-value for each comparison was determined based on the number of tests per CNV segment.

Genome-wide association with egg-laying date

Because mean egg-laying dates differ between years (as it is strongly affected by spring temperature, Gienapp et al. 2005) and among habitats, we fitted the following model to all recorded egg-laying dates (i.e. birds with and without genotypes) and used the year and area estimates from this model to 'pre-correct' the recorded phenotypes of the genotyped individuals:

$$y_{i,j} = \mu + \beta_j + \beta_a + pe_i + \varepsilon$$

with $y_{i,j}$ being the phenotype of individual i in year j, μ the overall intercept, β_j and β_a the fixed effects for year (as factor) and area (Buunderkamp-NL, Westerheide-NL, Roekelse Bos-NL, Hoge Veluwe-NL, Oosterhout-NL or Wytham Woods-UK), respectively and pe_i the random permanent environmental effect of individual i. We performed this two-step approach, instead of fitting year and area directly in the GWAS models that are described below, because not all individuals in all years were genotyped, which could have led to inaccuracy and/or bias in the estimates for year-area combinations with few genotyped individuals.

To identify phenotypic variation associated specifically with CNVs, we constructed association models that used PCA results based on CNV genotypes, in addition to pedigree information. By doing so, we ensured that results are not caused by any population structure/relatedness, which, if ignored, could cause spurious associations between specific CNVs and the phenotype. The use of a genomic relationship metric as PCA, in addition to non-genomic family information, may be relevant as the pedigree information is not available for all of the analyzed birds, limiting the accuracy of estimates of variance caused by family structure. For the PCA analysis, we used the snpgdsPCA function in SNPRelate R/Bioconductor package version 1.10.2 (Patterson et al., 2006; Zheng et al., 2012) for all autosomes. As we assumed that gain and losses will rarely share a common origin, both were considered different loci in the PCA analysis even when overlapping. Resulting eigen vectors one and two were used in the models mentioned below.

Given our previous evidence that CNV calling in the NL great tit population has a high rate of false negatives (da Silva et al., 2018), most of the CNV regions may have underestimated frequencies. This high rate of 'false negative CNVs' make association analyses difficult and may generate unreliable *p*-values. To address this problem, we performed a two-step association analysis. First, we used only the intensity strength from each SNP probe underlying CNVs (i.e. the same LRR values used for the CNV identification) in a linear mixed model that considered any known pedigree information. Then, the LRR results were compared with CNV states using the same model. We used LRR and CNV based models collectively, i.e. considering *p*-values from both models simultaneously for all CNV segments, to decrease the number of spurious associations.

As individual CNVs can display distinct breakpoints due to both biological and technical reasons (Abyzov et al., 2011; Alkan et al., 2011), and nearby probes may reflect the same CNV, we established CNV segments to be used as the loci in the association analysis. To construct these CNV segments, we first assigned the corresponding CNV state for each of the SNP probes overlapping a CNV call. Thus, we estimated the CNV frequency in each probe and selected only those with frequency above 5%. Then, these selected probes were used to construct CNV segments based on CNV-genotype similarity. In other words, the percentage of the birds with a given CNV state between subsequent probes defined the boundaries of each CNV segment in this population (minimum similarity of 90%). A simplification of the concept using 75% as the threshold is exemplified in the Figure 3.1.

A raw *p*-value was generated independently for each probe for each model. Raw *p*-values were corrected using genomic inflation. The probe with the lowest raw *p*-value was chosen to represent each corresponding CNV segment. We applied the Benjamini and Hochberg method (FDR, Benjamini & Hochberg 1995) on the assigned raw *p*-values to obtain *q*-values for each CNV segment.

We first identified suggestive associations in the LRR based model (q-value <0.1) which also have a significant association in the CNV based model (assigned raw p-value <0.05). Raw instead of *q*-values were used from the CNV based association for the following reasons: (i) a high rate of false negative CNVs may generate biased *q*-values in the CNV based association and (ii) independent association tests pointing to the same trend (i.e. LRR and CNV based models) inadvertently confer

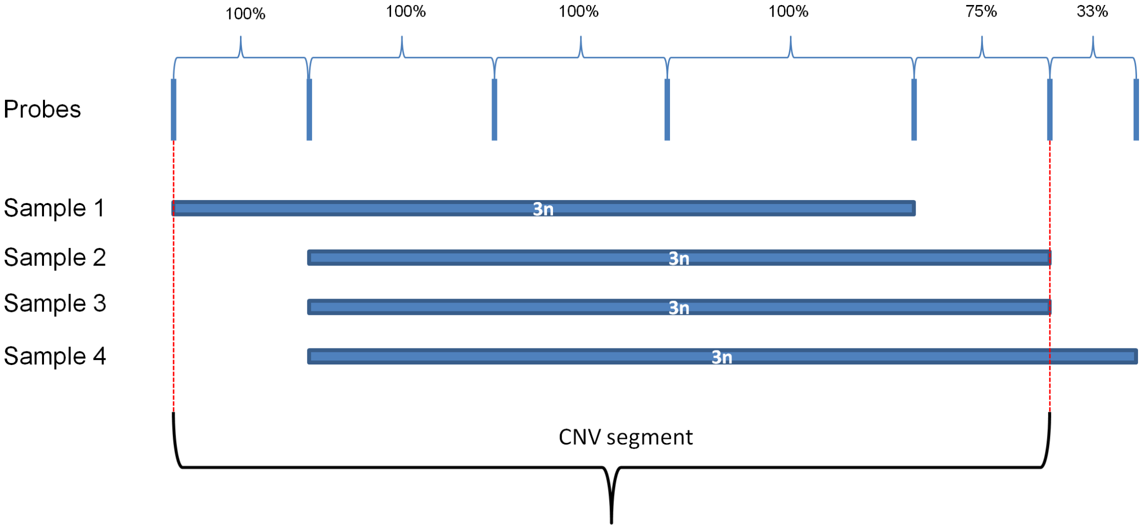


Figure 3.1: Concatenation of CNV segments. Example with a threshold of 75% for similarity between subsequent SNP probes.

robustness to the conclusions. The GWAS was performed separately for the great tit populations from (i) Netherlands (NL) and (ii) United Kingdom (UK).

LRR based model:

$$y'_{i,j} = \mu + age_i + pe_i + eigen1_i + eigen2_i + LRR_i + a_i + e_{i,j}$$

CNV based model:

$$y'_{i,j} = \mu + age_i + pe_i + eigen1_i + eigen2_i + CNV_i + a_i + e_{i,j}$$

$y'_{i,j}$ being the pre-corrected phenotype of individual i in year j, μ the overall intercept, age_i the age of individual i (as factor, 1^{st} year breeder versus older), pe_i the random permanent environmental effect of individual i, $eigen1_i$ and $eigen2_i$ the eigen vectors one and two from a CNV-based PCA analysis, CNV_i as the number of copies at the CNV segment or the log R ratio (LRR) at the representative probe (i.e. LRR and CNV based models) and a_i the sparse relatedness matrix calculated from pedigree (i.e. non-genomic family information). The model was fitted with the `relmatLmer` function from lme4qtl R package (Ziyatdinov et al., 2018). The GWAS procedure described in this study is implemented in the development branch of the CNVRanger R/Bioconductor package (version 0.99.18, Geistlinger & da Silva 2019). Therefore, the results presented in this study can be easily with the `cnvGWAS` function, from the CNVRanger package, by using the parameter `method.to.run=lmm`.

3.3 Results

CNVs across great tit populations in the Netherlands and the United-Kingdom

Using a high density custom SNP array, we previously identified CNVs in 2,175 birds from a Dutch great tit population (da Silva et al., 2018). In this study, we used the same methods to detected additional CNVs in 1,349 birds from Wytham Woods, Oxford, UK. After quality control, we found 20,828 CNVs which were subsequently merged into 6,450 CNV regions (CNVRs) in the UK population.

The CNVRs in the UK population cover 25.55% (235.72 Mb) of the autosomes. Coverage for different chromosomes ranged from 20.45% of chromosome 3 to 76% of chromosome 25LG1. The sizes of the CNVRs were variable ranging from 1 kb to 2.88 Mb with a mean size of 36.54 kb. The number of birds with CNVs mapped onto a given CNVR ranged from 1 (0.07%) to 357 (26.46%) of the 1,349 birds with at least one CNV identified. We found 148 CNVRs that occur in more than 1% of the population (> 13 birds) which we denote as 'polymorphic CNVRs', as previously suggested (Itsara et al., 2009).

The CNVRs from the UK population overlapped 7,338 of the 16,541 genes in the great tit genome (build 1.1, Laine et al. 2016). CNVRs showed enrichment for cell signaling, neuronal development and cardiac functions (Table 3.1), in accordance with the CNVRs identified in the NL population (da Silva et al., 2018).

Table 3.1: KEGG pathways significantly enriched for genes overlapping CNVRs in the UK great tit population.

ID	Description	q-value
hsa04514	Cell adhesion molecules (CAMs)	0.0129
hsa04740	Olfactory transduction	0.0129
hsa05410	Hypertrophic cardiomyopathy (HCM)	0.0246
hsa04360	Axon guidance	0.0246
hsa04392	Hippo signaling pathway - multiple species	0.0246
hsa04921	Oxytocin signaling pathway	0.0452
hsa05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	0.0452
hsa04720	Long-term potentiation	0.0452
hsa04925	Aldosterone synthesis and secretion	0.0467

From 6,450 CNVRs found in the UK population, 2,227 (28.96%) did not overlap with CNVRs identified in NL birds. From 8,008 CNVRs previously found in the NL population, a total of 3,754 (21.33%) did not overlap with CNVRs identified in the UK population.

Although a considerable number of CNVRs are population-specific, most of them are unique or low-frequency CNVRs. There was only one polymorphic CNVR which was uniquely found in NL population. This unique population-specific polymorphic CNVR is located on Chromosome 4 (at ≈54.63-54.73 Mb, CNVR 6317). This CNVR is present in 30 NL birds (1.38%) and there are no genes mapped to this genomic region.

A subgroup of the UK birds included here was also previously analyzed in an independent study (Kim et al., 2018), where the same custom array as well as the same software (i.e. PennCNV, Wang et al. 2007) was used to generate the CNV calls. We compared four different CNV data-sets reported in great tits, i.e. (i) CNVs identified in the NL population (da Silva et al., 2018); (ii) the UK population analyzed in this study; (iii) the UK population with and (iv) without the filtering criteria defined in Kim et al. 2018 (Figure 3.2). Approximately 26.17 Mb of the great tit genome harbors at least one CNV in all four data-sets.

Linkage-disequilibrium between CNVs and SNPs

Before performing the CNV-GWAS, we checked the linkage-disequilibrium (LD) between CNVs and SNPs. The justification for this step is that high LD among the two types of polymorphism would imply that a SNP-based GWAS should be sufficient to detect associations caused by CNVs. In contrast, if LD is low, then using CNV genotypes adds new information to a GWAS. We used CNV-SNP genotypes identified in the NL population to understand the LD between these two polymorphism types in the great tit genome (i.e. CNV and SNP genotypes that we previously published in da Silva et al. 2018 and da Silva 2019 et al. IN REVIEW, respectively). We performed a total of 292,583 CNV-SNP comparisons to infer LD (Figure 3.3). In general, the LD between these variants is low as SNP genotypes rarely tag CNV states. Only one comparison had an r^2 value above 0.5, indicating strong LD between SNPs and CNVs on Chromosome 1A, but not elsewhere. The CNV segment with the highest number of significant comparisons, as well as the highest r^2 value for a single comparison, is located within a tentative breakpoint of a large inversion on chromosome 1A (explored in detail elsewhere da Silva 2019 et al. IN REVIEW). In total 5,806 comparisons were significant after multiple correction (q-value ≤ 0.05), albeit displaying relatively low r^2 values. These comparisons represent 57 CNV segments (13.25% of all segments with frequency $\geq 1\%$), which have a median of five significant SNP-CNV comparisons each. The number of significant comparisons among CNV segments ranged from 1 to 1,249. The majority of SNPs

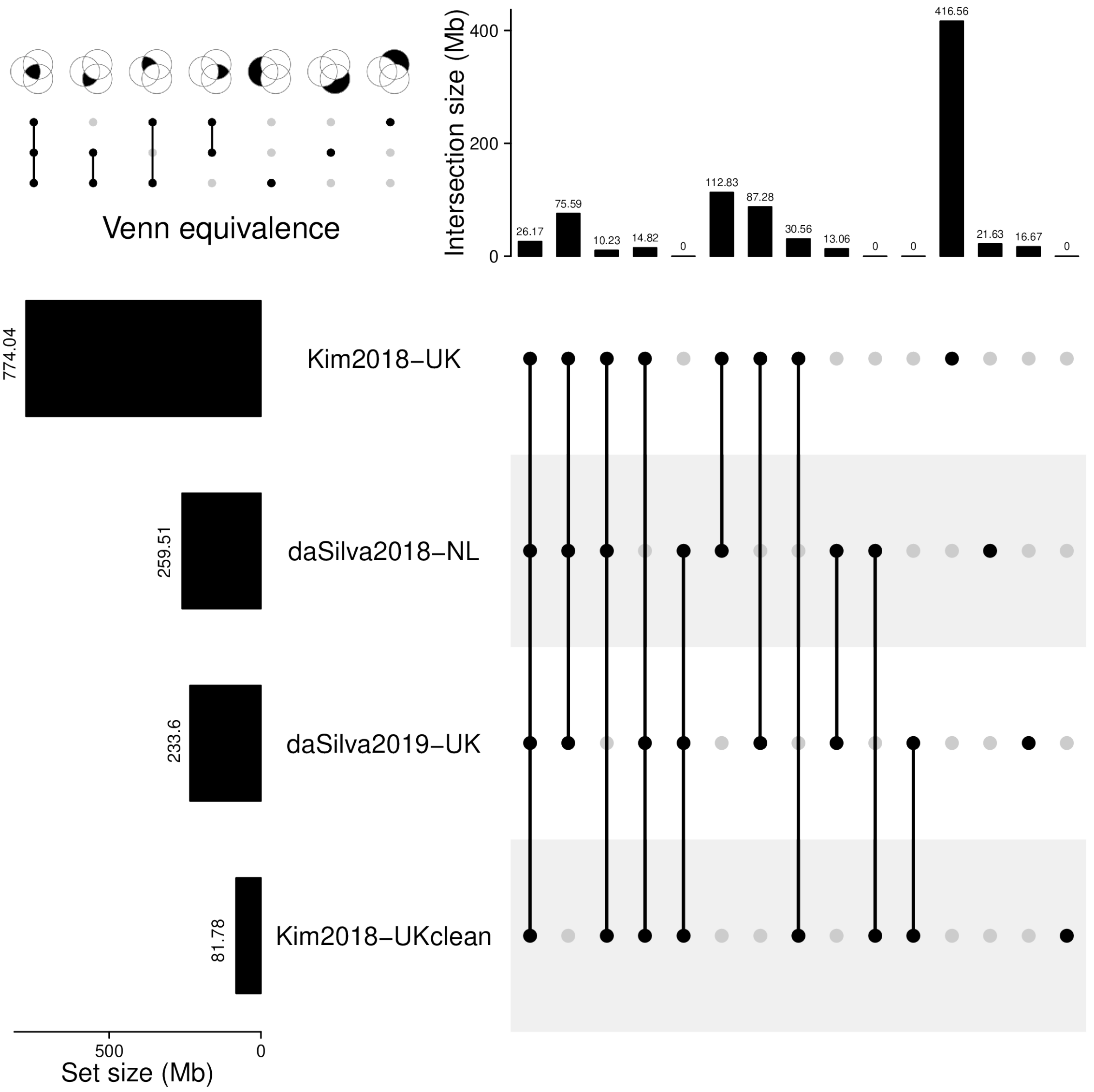


Figure 3.2: Comparison among CNV data-sets, reported in great tits, from different studies and populations. 'Kim2018-UK' represents all CNVs reported in the UK population by Kim et al. 2018. 'Kim2018-UKclean' represents a subset of the same CNV data-set after a strict filtering performed by Kim and colleagues (the filtering was based on the standard deviation of cluster distances). 'daSilva2018-NL' includes all CNVs previously reported by us in a NL population (da Silva et al., 2018). 'daSilva2019-UK' represents the CNV data-set identified here, but in the same UK population used in Kim et al. 2018. CNVs from 'daSilva2018-NL' and 'daSilva2019-UK' data-sets were filtered as described in da Silva et al. (2018).

near CNVs are not in LD with these CNVs. Nevertheless, SNPs closer to CNVs are more likely to have significant r^2 values (correlation coefficient = -0.57 and p-value < 0.0001).

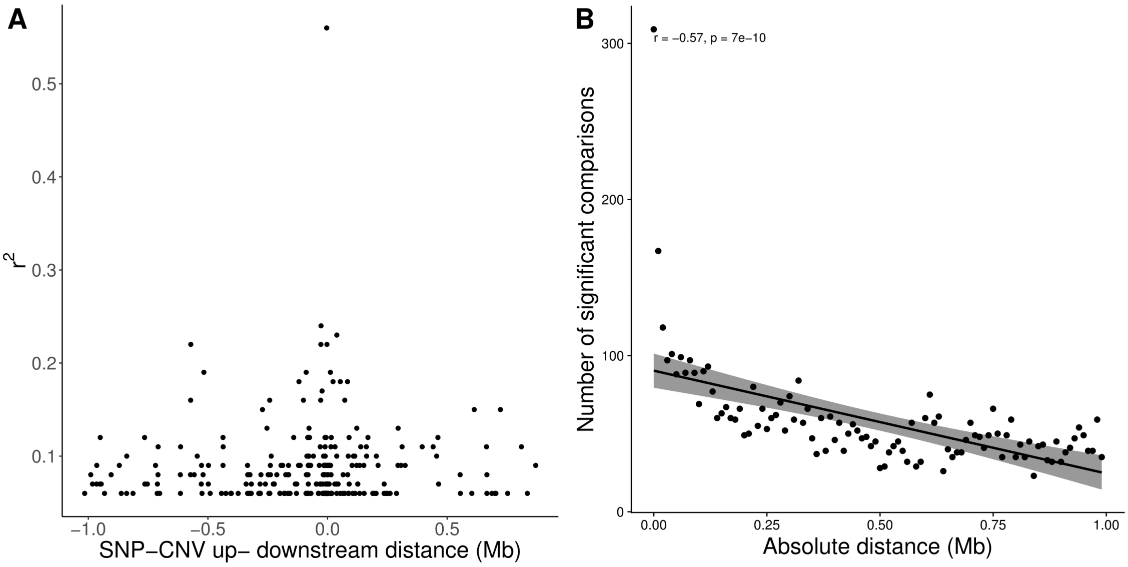


Figure 3.3: CNV-SNP linkage-disequilibrium analysis. A-) Distribution of r^2 values for pairwise comparisons of CNV segments and SNPs. The x-axis depicts the distance of neighbor SNPs from the center of each CNV segment. B-) Correlation coefficient of the number of significant SNP-CNV pairwise comparisons (p-value ≤ 0.05).

Egg-laying date association with CNVs

CNVRs are a coarse-grained map of CNV loci and are a valid approach for summarizing CNVs in a population. However, using CNVRs as loci in a GWAS usually leads to an oversimplification of the actual individual CNV genotypes. Thus, we performed a genome-wide association using CNV segments (**?**) concatenated based on genotype similarity of subsequent probes among all birds (i.e. from both NL and UK populations). First, we verified the CNV genotype per SNP probe to identify a total of 369 probes (Figure 3.4) for which at least 5% of the birds show overlap with a CNV (177 of the total of 3,524 birds, of which 2,175 are from NL and 1,349 from UK). These probes generated 42 distinct CNV segments, which have 12.76 kb in average ranging from 1 bp (only one probe) to 83.47 kb and are supported by 8.78 probes in average ranging from 1 to 63. Although CNV segments (i.e. CNV loci) were jointly inferred, the association analysis was carried out separately for the NL and UK populations.

The NL population has 2,133 birds which were phenotyped for egg-laying date and that had at least one CNV detected. In total 13 CNV segments display a suggestive q-value < 0.1 in the LRR based association. From these 13 CNV segments, five showed a significant raw p-value < 0.05 in the CNV based association. These five CNV segments are located on chromosomes 1, 2, 10 and 27, respectively (Figure 3.5 and Table 3.2).

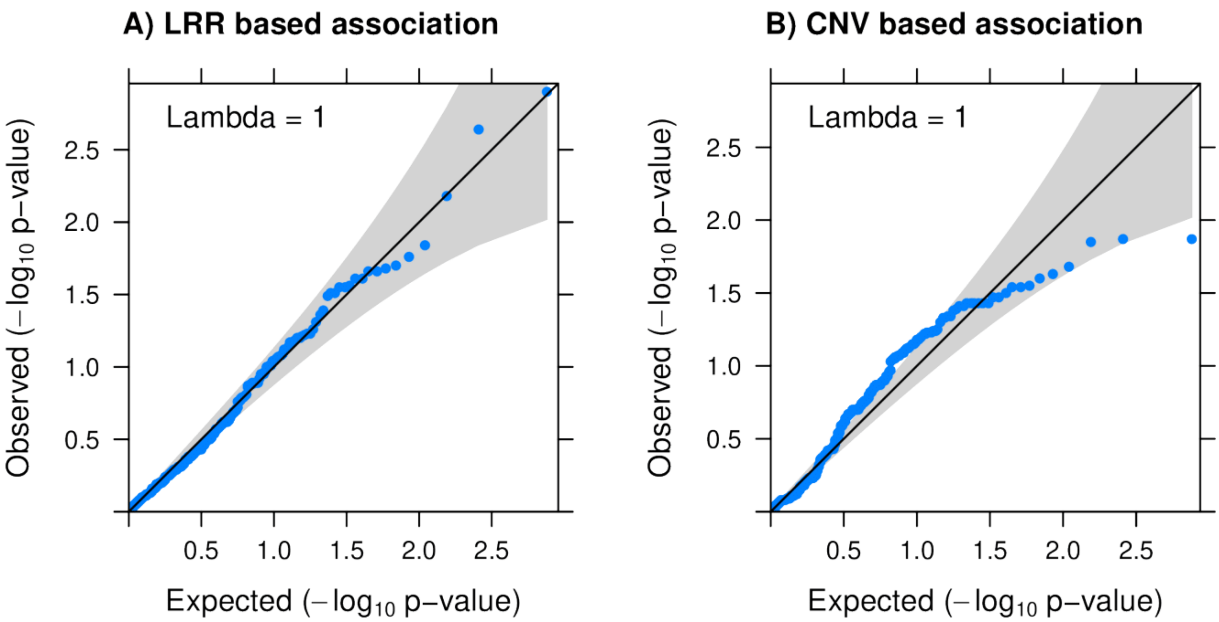


Figure 3.4: QQ-plots with the raw *p*-values for each of the 369 SNP probes supporting the CNV segments used as the CNV loci in the genome-wide associations. A-) LRR based association. B-) CNV based association.

Table 3.2: Five CNV segments above the suggestive thresholds in the CNV association with egg-laying date in a Dutch great tit population (q-value<0.1 in the LRR based association and raw *p*-value<0.05 in the CNV based association).

chr	start	end	CNV birds	*q*-value (lrr based)	raw *p*-value (cnv based)	genes
2	885391	942971	284	0.0976	0.0135	MYL3,TMIE
27	637857	668662	412	0.0987	0.0253	KPNB1,NPEPPS
1	98147555	98184039	827	0.0476	0.0341	none
10	19090667	19096193	194	0.0976	0.0392	ITGA11
10	1658760	1724249	284	0.0976	0.0458	STRA6,CCDC33

The CNV segment located on chromosome 1 (98.15-98.18 Mb) does not overlap any annotated gene. The CNV segment on chromosome 2 (0.88-0.94 Mb) overlaps two genes, coding for Myosin light chain 3 (*MYL3*) and Transmembrane Inner Ear (*TMIE*) respectively. We found two CNV segments on chromosome 10 (1.66-1.72 Mb and 19.09-19.10 Mb) which overlap with the Integrin Subunit Alpha 11 (*ITGA11*), Stimulated By Retinoic Acid 6 (*STRA6*) and Coiled-Coil Domain Containing (*CCDC33*) genes. The CNV segment on chromosome 27 (0.63-0.67 Mb) overlaps the Karyopherin Subunit Beta 1 (*KPNB1*) and Aminopeptidase Puromycin Sensitive (*NPEPPS*) genes. The UK population has 268 birds which were phenotyped for egg-laying date and that had at least one CNV detected. Among all five suggestive CNV segments found in the NL population, using the same threshold only the segment on chromosome 1 (98.15-9818 Mb, which does not overlap any annotate genes) was also suggestive in the UK population.

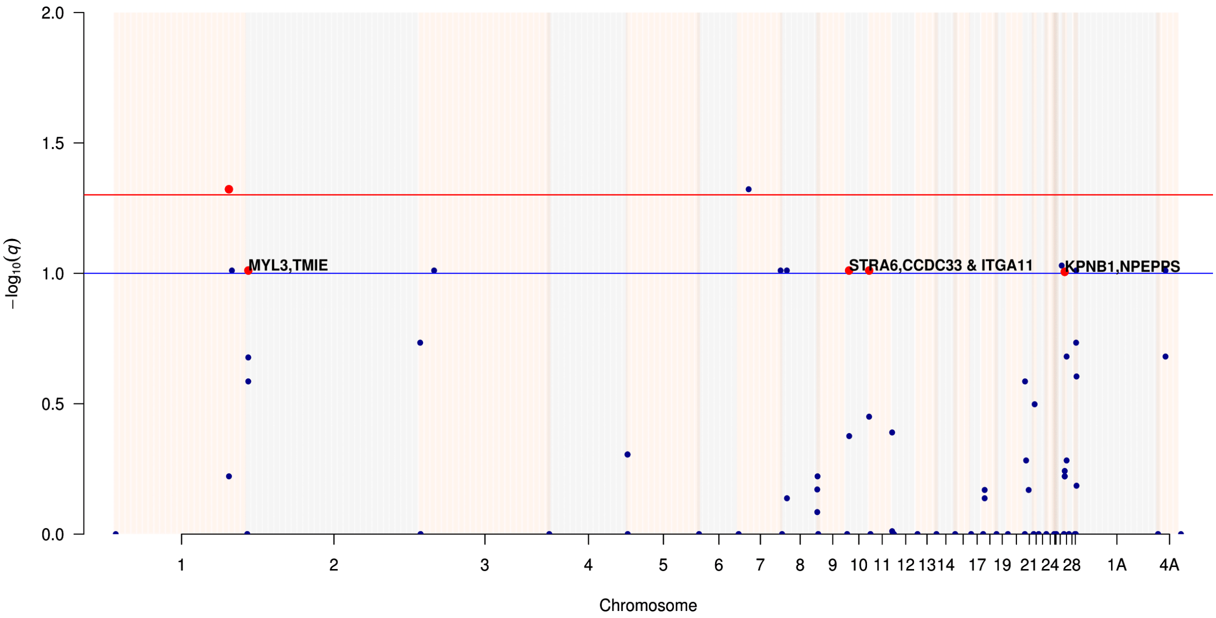


Figure 3.5: Genome-wide association of egg-laying date with CNVs from a population of great tits in the Netherlands. CNV segments with more than 5% in frequency display the q-values for the LRR based association. The red and blue lines represent 0.05 and 0.1 thresholds, respectively. Red dots represent five CNV segments which obtained concomitantly suggestive q-value (<0.1, blue line represented in the figure) in the LRR based association and significant raw p-value (<0.05, not represented in the figure) in the CNV based association.

3.4 Discussion

Breeding time is associated with the fitness in bird species (Perrins, 1970; Grüebler & Naef-Daenzer, 2010; Gienapp & Bregnballe, 2012). However, the plastic nature of seasonal timing has challenged efforts to find genetic variants underlying the ample phenotypic variation. Gienapp et al. (2017) recently addressed the genome-wide SNP associations of breeding time in great tits in an environment-dependent manner. However, the effect of structural variants, such as CNVs, has not yet been explored and these might reveal independent regions that are important to disentangle the genomic architecture of breeding time. Thus, to justify an association analysis using CNVs only, it is important to explore the linkage-disequilibrium (LD) between SNPs and CNVs.

CNVs are challenging to detect and interpret using SNP arrays due to technical and biological variability (Abyzov et al., 2011; Alkan et al., 2011). Technical variability can be linked to the platform, array design, and the software used for the CNV detection, among others (Carter, 2007; Zhao et al., 2013; Winchester et al., 2009). SNP arrays that are optimized for CNV identification (i.e. a higher number of probes at known CNV regions) might display an improved reliability. However, they can be costly to design and produce. Specifically for the custom-made SNP array used in this study (Kim et al., 2018), we detected the main source of bias to be the high number of false negative CNVs (da Silva et al., 2018). Therefore, using this information, we applied a hybrid strategy of association (i.e. CNVs and LRRs) which may reduce the bias caused by this high incidence of false negative CNVs in our data-set. Our strategy can partially tackle this bias because LRR values should, to some extent, reflect the CNVs (Yau & Holmes, 2008) that failed to be identified with the PennCNV software (Wang et al., 2007). On the other hand, LRR values can be noisy and generate false positives. Thus, as we identified a high number of false negatives in our NL CNV data-set, the combination of both approaches may help to find real associations here (and to remove false ones).

The biological variability of CNVs is likely to be mainly due to their usually complex breakpoints and also because they may possess multiple allelic states (i.e. it is usually difficult to determine the exact number of copies in regions that are especially repetitive). Furthermore, complex regions of the genome can harbor several CNVs with distinct origins, caused by different breakpoints. Thus, to partially tackle this biological variability we defined CNV loci based on the genotype similarity of nearby probes to generate CNV segments. Using this strategy we assume that CNVs with different origins affecting the same gene would have a comparable biological effect, which might be not true if overlapping CNVs affect different genomic features (e.g. different number of introns and exons).

Previous studies in humans have found that CNVs are less likely than SNPs to be in high LD with flanking SNPs (Schrider & Hahn, 2010). In accordance, we found low LD between CNVs and SNPs, which justifies performing a GWAS using CNV genotypes. Therefore, CNVs may provide different genetic information and could potentially point to independent genomic regions that are associated with egg-laying date. In fact, none of the top genomic regions found in the previously SNP based GWAS (Gienapp et al., 2017) were reflected on the results of the present study.

The CNV-GWAS performed here identified five regions that may be relevant for egg-laying date in the NL population of great tits, one of which also seen in the UK population with the same directional effect (i.e. CNV-birds usually have later egg-laying dates in comparison with 2n-birds). The UK population has less than 300 phenotyped birds, which limits the power to detect associations and explains the low number of relevant CNV segments in comparison with the NL population. Four out of the five highlighted CNV segments are located within genes, making it likely that they affect gene expression.

The CNV segment located on chromosome 1 does not overlap with any annotate genes, but it is the only segment that was independently found in both populations.

The indirect effect on genes that are located nearby cannot be discarded (e.g. a CNV which is overlapping an enhancer), but requires further investigation. The segment on chromosome 2 overlaps the *TMIE* gene, which is required for maturation of sensory hair cells in the cochlea and associated with recessive non-syndromic deafness (DFNB) (Naz et al., 2002; Mitchem et al., 2002). Two segments that are located on chromosome 10 overlap two genes that are associated with progesterone levels in the pregnant cervix in mammals (*ITGA11*, Ji et al. 2011) which is crucial in sexual mammalian reproduction and could therefore play a role in avian breeding timing. The last CNV segment on chromosome 27 overlaps *KPNB1*, which mediates the circadian clock function (Lee et al., 2015). Circadian clocks are linked to seasonal timing by providing reference for photoperiodic time measurement and most likely also by associations with circannual rhythms (Helm & Visser, 2010). Although we describe possible CNV associations with egg-laying dates in great tits, the results presented here should be treated carefully. Robust association of CNVs with quantitative phenotypes is not a trivial task, lacking a clear well defined 'gold standard' (i.e. given the above-mentioned technical and biological limitations). In addition, the known high false negative rate reported for the CNV calling in this study might not fully be tackled by our hybrid GWAS strategy.

3.5 Conclusions

Seasonal timing is a complex polygenic trait that can be affected by environmental factors like spring temperature, altitude and food availability (Noordwijk et al., 1980; Gienapp et al., 2005; Wilkin et al., 2007) making it a challenge to unravel the underlying genetic variation. Nevertheless, this study provides a first glance of the role of more complex variants such as CNVs by exploring their effect on egg-laying date of great tit species, known to be shifting due to global warming (Visser et al., 2006).

Chapter 4

CNVRanger: association analysis of CNVs with gene expression and quantitative phenotypes

Vinicius H. da Silva[1,2,3], Marcel Ramos[4], Martien A. M. Groenen[1], Richard P.M.A. Crooijmans[1], Anna M. Johansson[3], Luciana C. de A. Regitano[5], Luiz L. Coutinho[6], Ralf Zimmer[7], Levi Waldron[4], Ludwig Geistlinger[4]

[1]Animal Breeding and Genomics, Wageningen University & Research, Wageningen, The Netherlands
[2]Department of Animal Ecology, Netherlands Institute of Ecology (NIOO-KNAW), Wageningen, The Netherlands
[3]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden
[4]Graduate School of Public Health and Health Policy, City University of New York, New York, USA
[5]Embrapa Pecuária Sudeste, São Carlos, Brazil
[6]Animal Science Department, University of São Paulo, Piracicaba, Brazil
[7]Institute of Bioinformatics, Ludwig-Maximilians-Universität München, München, Germany

Oxford Bioinformatics, doi:10.1093/bioinformatics/btz632

Abstract

Copy number variation (CNV) is a major type of structural genomic variation that is increasingly studied across different species for association with diseases, production traits, and evolutionary footprints. Established protocols for experimental detection and computational inference of CNVs from SNP array and next-generation sequencing data are available. However, only limited options exist for further interpretation of CNV data and integration with gene expression and quantitative phenotypes. We present the `CNVRanger` R/Bioconductor package which implements a comprehensive toolbox for structured downstream analysis of CNVs. This includes functionality for summarizing individual CNV calls across a population, assessing overlap with functional genomic regions, and genome-wide association analysis with gene expression and quantitative phenotypes.

4.1 Introduction

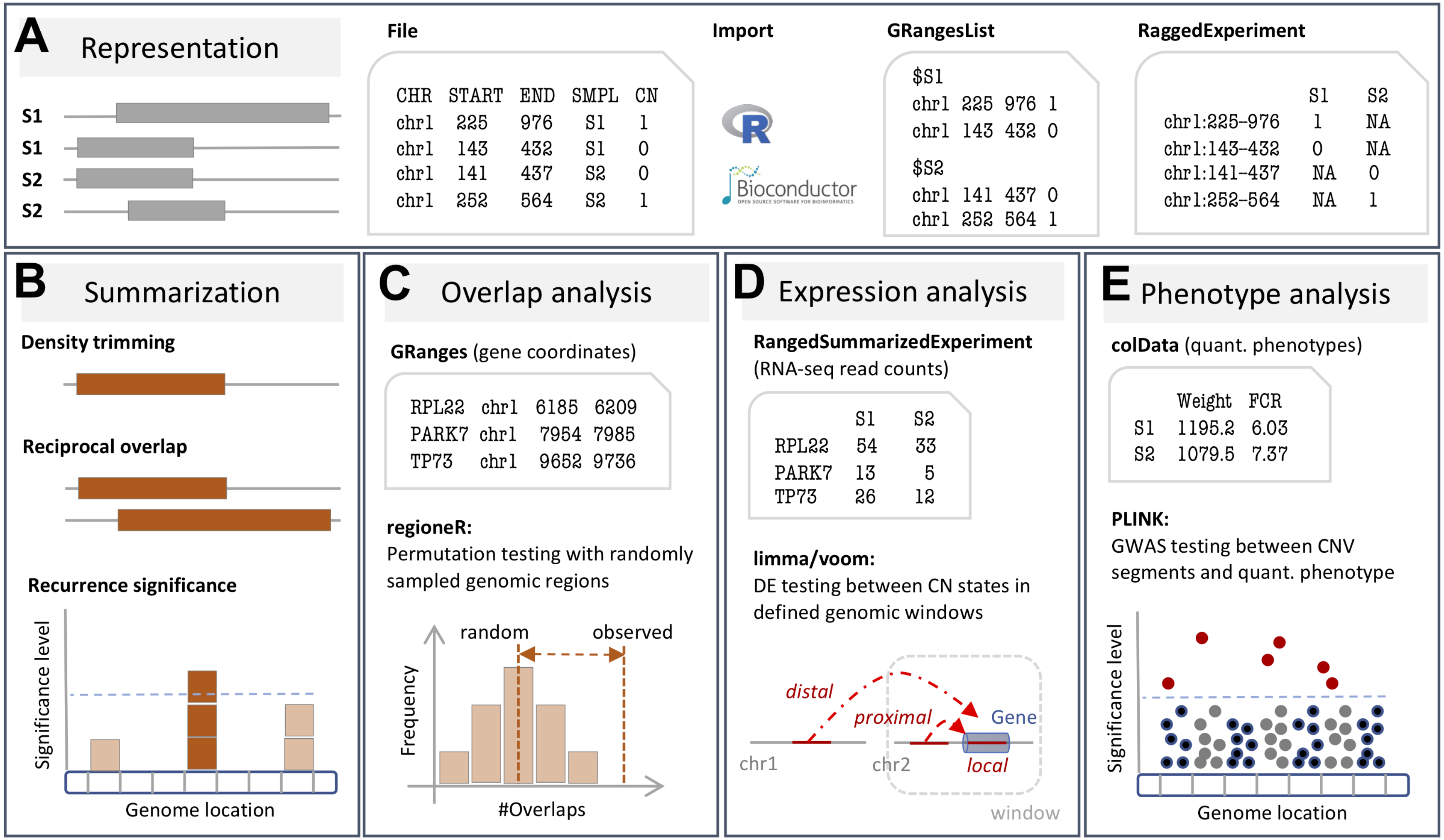
Copy number variation (CNV) is a frequently observed deviation from the diploid state due to duplication or deletion of genomic regions (Conrad et al., 2010). CNVs can be experimentally detected based on comparative genomic hybridization, and computationally inferred from SNP-arrays or next-generation sequencing data. These technologies for CNV detection have in common that they report, for each sample under study, genomic regions that are duplicated or deleted with respect to a reference genome. Such regions are denoted as *CNV calls* in the following and are typically the starting point for subsequent downstream analysis.

In previous work, we developed, described, and applied functionality for analyzing CNVs across a population, including association analysis with gene expression and quantitative phenotypes (da Silva et al., 2016; Geistlinger et al., 2018; da Silva et al., 2018). To allow straightforward application to similar datasets, we generalize these concepts and provide refined implementations in the `CNVRanger` R/Bioconductor package.

4.2 Features

4.2.1 Reading and accessing CNV data

The `CNVRanger` package reads CNV calls given in a general file format, providing at least chromosome, start position, end position, sample ID, and integer copy number for each call (Fig. 4.1A). Once imported into `R`, the CNV data is stored for efficient representation and manipulation in `Bioconductor` (Huber et al., 2015) data structures as implemented in the `GenomicRanges` (Lawrence et al., 2013) and `RaggedExperiment` (Morgan & Ramos, 2017) packages.



```
CHR  START  END  SMPL  CN
chr1  225   976   S1   1
chr1  143   432   S1   0
chr1  141   437   S2   0
chr1  252   564   S2   1
```

```
$S1
chr1 225 976 1
chr1 143 432 0
$S2
chr1 141 437 0
chr1 252 564 1
```

```
              S1   S2
chr1:225-976  1    NA
chr1:143-432  0    NA
chr1:141-437  NA   0
chr1:252-564  NA   1
```

```
RPL22  chr1  6185  6209
PARK7  chr1  7954  7985
TP73   chr1  9652  9736
```

```
       S1   S2
RPL22  54   33
PARK7  13    5
TP73   26   12
```

```
    Weight  FCR
S1  1195.2  6.03
S2  1079.5  7.37
```

Figure 4.1: The `CNVRanger` package **(A)** imports CNV calls given in a general format into `R`, and stores them in dedicated `Bioconductor` data structures, and **(B)** implements three frequently used approaches for summarizing CNV calls across a population: (i) the `CNVRuler` procedure that trims region margins based on regional density (Kim et al., 2012), (ii) the reciprocal overlap (RO) procedure that requires calls to sufficiently overlap with one another (Conrad et al., 2010), and (iii) the `GISTIC` procedure that identifies recurrent CNV regions (Beroukhim et al., 2007). `CNVRanger` also **(C)** builds on `regioneR` (Gel et al., 2015) for overlap analysis of CNVs with functional genomic regions such as genes, promoters, and enhancers, **(D)** implements RNA-seq expression Quantitative Trait Loci (eQTL) analysis for CNVs by interfacing with `edgeR` (Robinson et al., 2010), allowing to restrict analysis to specified genomic windows to e.g. detect cis-eQTLs, and **(E)** interfaces with `PLINK` (Purcell et al., 2007) for traditional genome-wide association studies (GWAS) between CNVs and quantitative phenotypes.

4.2.2 Summarizing individual CNV calls across a population

For the analysis of CNVs in a population study, `CNVRanger` implements three frequently used approaches for defining recurrent regions (Fig. 4.1B). The `CNVRuler` (Kim et al., 2012) method trims low-density areas that would otherwise inflate the size of the resulting CNV region, by default trimming region margins that are covered by <10% of the total number of calls within a region. The reciprocal overlap (RO) procedure merges calls with sufficient mutual overlap (Conrad et al., 2010). For example, an RO of 0.51 between calls A and B requires A to overlap at least 51% of B, and B to also overlap at least 51% of A. Particularly in cancer, it is important to distinguish driver from passenger mutations, i.e. to distinguish meaningful events from random background aberrations. The `GISTIC` (Beroukhim et al., 2007) method identifies those regions of the genome that are aberrant more often than would be expected by chance, with greater weight given to high amplitude events (high-level copy-number gains or homozygous deletions) that are less likely to represent random aberrations.

4.2.3 Overlap analysis with functional genomic regions

Once recurrent CNV regions have been defined, `CNVRanger` allows to assess whether and to which extent these regions overlap with functional genomic regions such as genes, promoters, and enhancers (Fig. 4.1C). As a certain amount of overlap can be expected just by chance, an assessment of statistical significance is needed to decide whether the observed overlap is greater (enrichment) or less (depletion) than expected by chance. `CNVRanger` therefore builds on the `regioneR` package (Gel et al., 2015), which implements a general framework for testing overlaps of genomic regions based on permutation sampling. We use the package to repeatedly sample random regions from the genome, matching size and chromosomal distribution of the CNV regions. By recomputing the overlap with the functional features in each permutation, statistical significance of the observed overlap can be assessed.

4.2.4 CNV-expression association analysis

The `CNVRanger` package implements association testing between CNV regions and RNA-seq read counts based on `edgeR` (Robinson et al., 2010), which applies generalized linear models based on the negative-binomial distribution while incorporating normalization factors for different library sizes. For CNV regions with only one CN state deviating from the $2n$ reference group, this reduces to the classical 2-group comparison as previously described (Geistlinger et al., 2018). For multi-allelic CNVs (e.g. $0n$, $1n$, $2n$), `edgeR`'s ANOVA-like test is applied to test for significant expression differences in any non-diploid group with respect to the $2n$ group. Assuming distinct modes of action, we distinguish between (i) local effects (*cis*), where expression changes coincide with CNVs in the respective genes, and (ii) distal effects (*trans*), where CNVs supposedly affect trans-acting regulators such as transcription factors (Fig. 4.1D). Due to power considerations and to avoid detection of spurious effects, stringent filtering of (i) not sufficiently expressed genes, and (ii) CNV regions with insufficient sample size in groups deviating from $2n$, is carried out when testing for distal effects. Local effects have a clear spatial indication and the number of genes locating in or close to a CNV region of interest is typically small; testing for differential expression between CN states is thus generally better powered for local effects and less stringent filter criteria can be applied.

4.2.5 CNV-phenotype association analysis

Specifically developed for CNV calls inferred from SNP-chip data, `CNVRanger` allows to carry out a probe-level genome-wide association study (GWAS) with quantitative phenotypes (Fig. 4.1E). CNV calls from other sources such as sequencing data are also supported by using the start and end position of each call as the corresponding probes. As previously described (da Silva et al., 2016), we then construct CNV segments from probes representing common CN polymorphisms (CNPs, allele frequency >1% as default), and carry out a GWAS as implemented in `PLINK` (Purcell et al., 2007) using a standard linear regression of phenotype on allele dosage. For CNV segments composed of multiple probes, the segment p-value is chosen from the probe p-values, and multiple testing correction is carried out using the FDR method (Benjamini & Hochberg, 1995) per default. This is similar to a common approach used in differential expression analysis of microarray gene expression data, where typically the most significant probe is chosen in case of multiple probes mapping to the same gene. Results can then be displayed as for regular GWAS via a Manhattan plot.

Chapter 5

The genomic complexity of a large inversion in great tits

Vinicius H. da Silva[1,2,3], Veronika N. Laine[1,4], Mirte Bosse[1], Lewis G. Spurgin[5], Martijn F.L. Derks[1], Kees van Oers[2], Bert Dibbits[1], Jon Slate[6], Richard P.M.A. Crooijmans[1], Marcel E. Visser[1,2], Martien A.M. Groenen[1].

[1]Animal Breeding and Genomics, Wageningen University & Research, Wageningen, The Netherlands
[2]Department of Animal Ecology, Netherlands Institute of Ecology (NIOO-KNAW), Wageningen, The Netherlands
Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden
[4]Department of Molecular and Cellular Biology, Harvard University, Cambridge, USA
[5]School of Biological Sciences, University of East Anglia, Norwich, England
[6]Department of Animal and Plant Sciences, The University of Sheffield, Sheffield, England

Genome Biology and Evolution, 2019 11:1870–1881

Abstract

Chromosome inversions have clear effects on genome evolution and have been associated with speciation, adaptation and the evolution of the sex chromosomes. In birds, these inversions may play an important role in hybridization of species and disassortative mating. We identified a large (≈64 Mb) inversion polymorphism in the great tit (*Parus major*) that encompasses almost 1,000 genes and more than 90% of Chromosome 1A. The inversion occurs at a low frequency in a set of over 2,300 genotyped Dutch great tits with only 5% of the birds being heterozygous for the inversion. In an additional analysis of 29 resequenced birds from across Europe we found two heterozygotes. The likely inversion breakpoints show considerable genomic complexity, including multiple copy number variable segments. We identified different haplotypes for the inversion, which differ in the degree of recombination in the center of the chromosome. Overall, this remarkable genetic variant is widespread among distinct great tit populations and future studies of the inversion haplotype, including how it affects the fitness of carriers, may help to understand the mechanisms that maintain it.

5.1 Introduction

Inversions are structural intra-chromosomal mutations resulting in the reversal of gene/sequence order. Chromosomal inversions represent an important class of polymorphism that are of particular interest in evolutionary studies (Hoffmann & Rieseberg, 2008; Kirkpatrick, 2010). Numerous studies have shown inversions to be important factors in speciation and adaptation (reviewed in Hoffmann & Rieseberg 2008). Studies of hominin evolution indicate a crucial role of inversions in the process, with more than one thousand inversions arising in both the human and chimpanzee lineages since they shared a common ancestor (Hellen, 2015). Red fire ants (*Solenopsis invicta*) provide an interesting example of how inversions can promote adaptation; whether or not ant colonies contain a single queen or multiple queens depends on which inversion genotype is present the colony. The two social forms are genetically isolated (Keller & Ross, 1998; Wang et al., 2013). In passerines, inversions are significantly more common in clades with more sympatric species, which suggests that inversions may often evolve or be maintained because they suppress recombination between the genomes of hybridizing species (Hooper & Price, 2017). In both the white-throated sparrow (*Zonotrichia albicollis*) and the ruff (*Calidris pugnax*), morphs with different sexual behaviours are determined by inversions (Küpper et al., 2015; Lamichhaney et al., 2016; Tuttle et al., 2016). The inversion in the white-throated sparrow is very large, harboring ≈1,000 genes, and lethal in homozygous state (Tuttle et al., 2016).

To explain how inversions are maintained in a population it is important to understand the different mechanisms underlying selection on inversions. There can be meiotic drive if the inversion harbors alleles that alter segregation distortion (Kirkpatrick, 2006). Selective advantages can also occur when an inversion affects the expression of advantageous genes located within or closely linked to the inversion (Puig et al., 2004). The effect of the inversion on gene expression is well documented in red fire ants (Wang et al., 2008, 2013; Nipitwattanaphon et al., 2013; Lucas et al., 2015; Huang et al., 2018). In this species, gene expression differences between the monogyne and polygyne social forms are greatest in the inversion, suggesting that the inversion plays a key role in morphological and behavioural differences between the two forms. In addition, selective advantages of an inversion can be the result of recombination disruption in heterozygotes, which can preserve advantageous alleles. Moreover, reduced crossing-over within the inversion is associated with higher recombination rate elsewhere in the genome (Stevison et al., 2011), which in turn can modulate selection (McGaugh et al., 2012).

In many cases, recombination is suppressed between an inverted haplotype and the wild haplotype. As a result of this lack of recombination in heterozygous inversion carriers, strong linkage disequilibrium between loci within the inverted region can rapidly build up. Although the lack of recombination can maintain advantageous variants without disruption throughout generations (i.e. supergenes, reviewed in Thompson & Jiggins 2014), there are also possible costs associated with the suppression of recombination. Each of the inversion haplotypes will behave as a single heritable entity that can help to retain certain alleles in the population even when they are subject to purifying selection (i.e. deleterious recessive alleles can be maintained if they are found within inversion polymorphisms by a "hitchhiking" effect, Kirkpatrick 2006). As a consequence, deleterious recessive alleles can accumulate in regions of low recombination, such as an inversion, as they are no longer effectively removed by purifying selection. Moreover, throughout evolution an inversion becomes structurally more complex than the non-inverted counterpart and often experiences a degenerative process (Tuttle et al., 2016). This degenerative process has been reported to be associated with a size increase in young supergenes (Stolle et al., 2018). In general, an increase in the number of gene copies can alter trans- and cis-gene expression, which might generate novel phenotypic variation (Geistlinger et al., 2018).

Inversions may harbor complex genomic rearrangements at their breakpoints (Calvete et al., 2012), given that inversion breakpoints are more likely to happen at complex parts of a chromosome (Carvalho & Lupski, 2016). Apart from changing the gene order, inversions also often involve gene duplications that can lead to genetic novelty and subsequent adaptation (Furuta et al., 2011). In mosquitoes from the species complex *Anopheles gambiae*, haplotypes involving structural rearrangements at the breakpoint of a paracentric inversion have shed light on the origin and evolution of their malaria vectorial capacity (Sharakhov et al., 2006). The presence of repetitive regions at inversion breakpoints is recurrent and in both inversions and repetitive regions can share the same mechanism of formation, such as non-allelic homologous recombination (NAHR) (Carvalho & Lupski, 2016; Kehrer-Sawatzki & Cooper, 2008). Understanding structural variations linked to inversion breakpoints may help to clarify the possible functionality and evolutionary history of inversions.

Genetic markers like SNPs and sequence data can be used to identify inversions polymorphism given the distinct population genetic structure caused by LD patterns within inversions. Thus, methods that are based on principal components analysis (PCA) can detect the unusual genetic structure of inversions (Ma & Amos, 2012). In this study, we describe a 64.2 Mb putative inversion on Chromosome 1A in great tits (*Parus major*), a widely studied songbird in ecology and evolution (Kvist et al., 2003; Visser et al., 1998; Husby et al., 2011) with a broad range of genomic resources such as a high density SNP array (Kim et al., 2018), reference genome and methylome analysis (Laine et al., 2016) as well as copy number variation (CNV) maps (da Silva et al., 2018; Kim et al., 2018).

5.2 Material and methods

5.2.1 Population description, genotyping and sequencing.

A total of 2,322 great tits were genotyped using a custom made Affymetrix® great tit 650K SNP chip (Kim et al., 2018) at Edinburgh Genomics (Edinburgh, United Kingdom). SNP calling was done following the Affymetrix® best practices workflow by using the Axiom® Analysis Suite 1.1. After sample filtering, 26 birds with dish quality control (DQC, Nicolazzi et al. (2014)) <0.82 and SNP call rate <95% were discarded. SNPs with minor allele frequency (MAF) <1% and call rate <95% were removed. Only autosomes were used in this study. After filtering, 2,296 birds and 514,799 SNPs were kept for subsequent analysis. The genotyped birds were from our long-term study populations on the 'Veluwe' area near Arnhem, the Netherlands (52°02' N, 5°50' E). More information regarding the origin of the birds and the *in vitro* DNA procedures are described by da Silva et al. (da Silva et al., 2018). The raw genotype data used in this study was submitted to GEO (GSE105131). Filtered genotypes and the source code to perform all analyses described below are available at Open Science Framework (OSF, `https://osf.io/t6gnd/?view_only=821507ec135b44778d8b80254c24633b`).

In addition to the birds genotyped on the SNP chip, we also used sequence data from 29 birds (10 from the Wytham Woods population in Oxford (UK), 19 birds sampled from 15 other European populations). Each bird was sequenced at an average depth of around 10x using paired-end sequencing libraries. Details of sequencing analysis, as well as information regarding the origin and sample quality of each bird are provided elsewhere (Laine et al., 2016).

5.2.2 Identification and characterization of a large inversion on Chromosome 1A.

Population structure between SNP-typed individuals was explored using a principal components analysis (PCA) approach, previously applied for the study of inversions (Ma & Amos, 2012), using the `snpgdsPCA` function in SNPRelate R/Bioconductor package (v. 1.10.2) (Patterson et al., 2006; Zheng et al., 2012). Each autosome was analysed separately.

Following PCA, we estimated the fixation index (F_{ST}) in a SNP-wise fashion, using the `Fst` function available in snpStats R/Bioconductor package (v. 1.26.0) (Clayton, 2015) to compare birds in different clusters identified by visual inspection (i.e. subpopulations) of PCA plots. As SNP heterozygosity is expected to be higher within the inversion in carriers (i.e. birds with two different inversion haplotypes), the ratio of heterozygous birds (i.e "AB") for each SNP was assigned within each subpopulation. The SNP-wise F_{ST} and heterozygosity values were used to define the likely breakpoints of the inversion.

Pairwise D' values, (Lewontin & Kojima, 1960) using all birds, were calculated to assess the linkage disequilibrium. To aid visualization of the patterns revealed by the SNP data, SNPs were pruned to retain loci with MAF >0.4 and an LD threshold of 0.05 (using genomic windows with a maximum size of 500 kb). Pruning was performed with the `snpgdsLDpruning` and `snpgdsLDMat` functions within the SNPRelate R/Bioconductor package (v. 1.10.2) (Zheng et al., 2012). A total of 214 SNPs was retained and used in the LD analysis plot. We produced a graphical representation of the LD map using the `LDheatmap` function from the LDheatmap R package (v. 0.99-2) (Shin et al., 2006). The function used to infer LD in this study makes use of the expectation-maximization (EM) algorithm (Excoffier & Slatkin, 1995), which is able to infer LD from unphased data. In addition, the R^2 (Zaykin et al., 2008) estimator was used for comparison with results from D' because each estimator may respond differently to low frequency alleles (Wray, 2005).

5.2.3 Inference of structural complexity at Chromosome 1A.

We used copy number variation (CNV) data obtained from SNP intensity information from the same Dutch great tit population, as described previously (da Silva et al., 2018), to evaluate if certain CNVs are associated with normal/inverted phases. Moreover, we identified CNVs in the 29 resequenced birds from different European populations (Laine et al., 2016)). First, we used the *.bam* file of each sample, containing reads mapped onto the reference genome build 1.1 using BWA (Li & Durbin, 2009), to extract map locations with samtools (Li et al., 2009) as described in CNV-seq manual (Xie & Tammi, 2009a). CNVs were called with the default parameters of CNV-seq (Xie & Tammi, 2009b). CNV-seq uses coverage information to calculate a $\log_2$ transformed ratio between the subject samples (inv-norm only, because inv-inv birds were absent from the dataset) and wild-type samples (norm-norm). A positive ratio is associated with copy-number gain (duplication), while a negative ratio is associated with copy-number loss (deletion).

In addition, we used Lumpy (Layer et al., 2014) with default parameters, incorporated in the speedseq pipeline (Chiang et al., 2015) to predict the exact breakpoints of the CNV events and to predict inversion events from sequence data. Information from split and discordant mapped reads was used to describe the structure of a CNV complex in one of the inversion breakpoints (details in the supplementary section 3.4- Patterns in split reads supporting the CNV complex).

5.2.4 Inversion detection by PCR-RFLP.

As genotyping with SNP array can be time consuming and expensive, we designed an alternative method to type the Chromosome 1A inversion, based on a PCR followed by a restriction enzyme digestion (PCR-RFLP). For this, we used the SNP with the second highest F_{ST} value (i.e. AX-100689781) because it almost perfectly captures the inversion (99.32% of the inv-norm birds have AB genotype and 98.95% of the norm-norm birds have the AA genotype). The SNP with the highest F_{ST} value did not allow distinguishable fingerprints *in silico* because there are no restriction enzymes which differentially cut the two alleles. Instead, we choose SNP AX-100689781 which is located close to the downstream breakpoint of the inversion, at position 65,878,384 in the great tit genome build 1.1 (Laine et al., 2016) (details in the supplementary section primer design and enzyme search). This SNP is located within the first intron of the gene *PIK3C2G*. We genotyped 42 birds by PCR-RFLP which had also been genotyped with the SNP-chip.

For each PCR-RFLP reaction we used 6μl of DNA (10ng/μl). The PCR was performed with OneTaq 2X mastermix (New England Biolabs) and 1μl of primermix (primer sequences are given in the supplementary section primer design and enzyme search). The PCR program had steps of: 95°C for 5 min, 34 cycles of 95°C for 30 seconds, 55°C for 45 seconds, 72°C for 90 seconds and a final elongation step of 72°C for 10 min. The digestion reaction was done for 5 hrs at 37°C using 3μl of the PCR product, 0.4μl of the enzyme *SspI* (10U/μl, New England Biolabs), 1μl of the *SspI* buffer 10X and 5.6μl of sterile deionized water (MQ). The PCR-RFLP was analyzed on a 3% agarose gel. The restriction fragments were checked on the Geldoc XR+(Biorad) gel documentation system with the software Image Lab (v. 5.2.1).

5.3 Results

5.3.1 Population structure for Chromosome 1A reveals a large inversion.

We found a large putative inversion on Chromosome 1A. Based on visual inspection of the principal component analysis (PCA) (Patterson et al., 2006), we classified the clustering patterns separately for each autosome in the great tit genome (Sup

Figure 5.6). Plots for whole chromosomes may reveal obvious substructure if the inversion is relatively large. Although additional chromosomes display some population structure (e.g. chromosomes 5 and 7, Sup Figures 5.6 and 5.7), the variation within PCA clusters is greater, and the F_{ST} values across these chromosomes less conclusive, relative to the patterns seen on Chromosome 1A. Moreover, this unusual PCA pattern, which was most likely reflecting an inversion, was briefly reported elsewhere (Bosse et al., 2017). Therefore, the remainder of this paper considers the likely inversion polymorphism on Chromosome 1A. Chromosome 1A displayed clear population structure for the first eigenvector (Figure 5.1a, First and Second eigenvectors explain 2.28 and 0.50% of the variance, respectively), with two subpopulations that are genetically distinct. The larger subpopulation comprises 2,179 birds and the smaller one contains only 117. Among these 117 birds, ten display intermediate values in Eigenvector One. Analysis of the ten birds' genotypes indicates that they are carrying a distinct haplotype, derived from the inversion, rather than representing a distinct inversion genotype from the rest of these birds (e.g. the ten being heterozygotes and the remainder being homozygous for the inversion haplotype). The genotypes and LD patterns in the center of the inversion are discussed in detail in a subsequent section (i.e. Linkage-disequilibrium and haplotypes across the inversion).

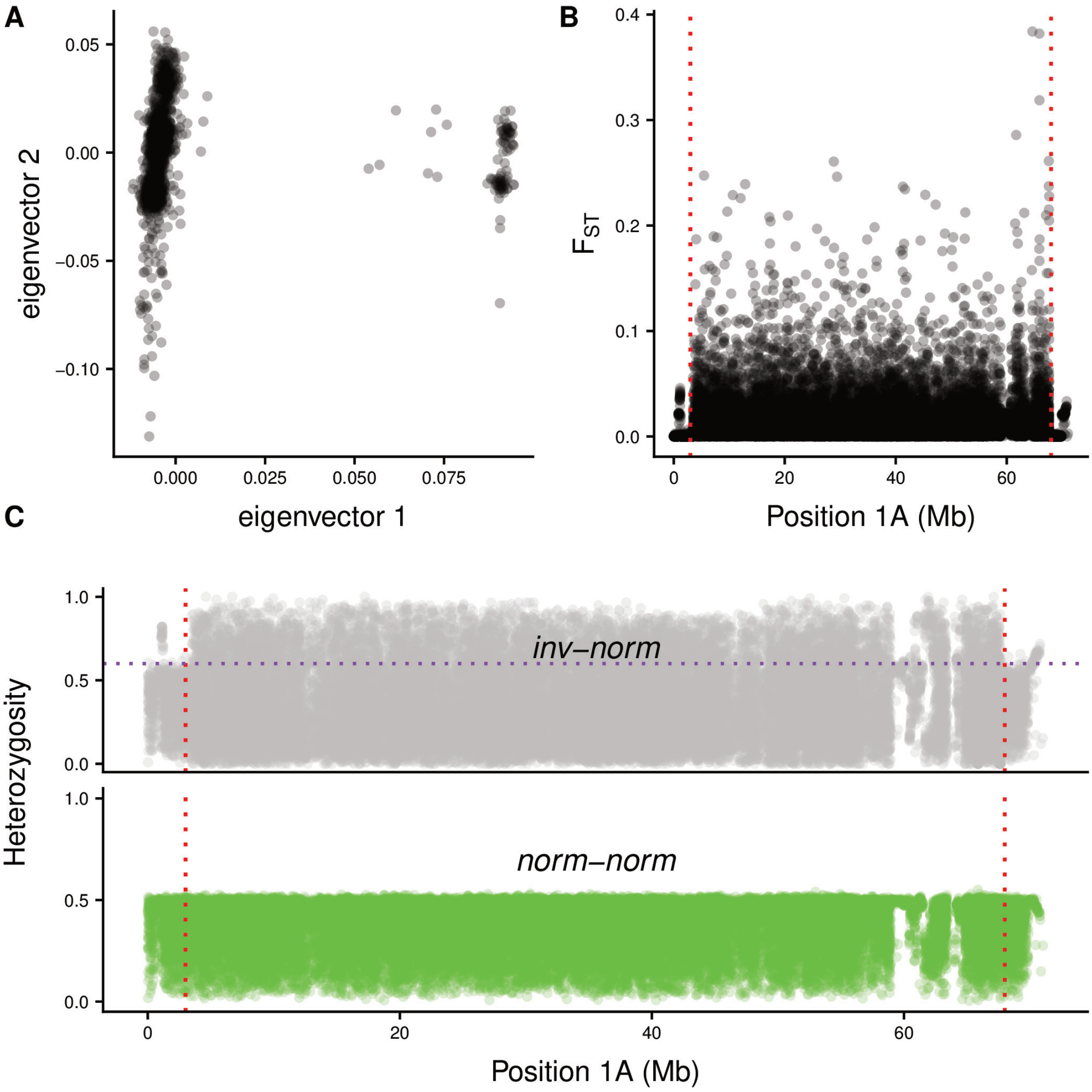


Figure 5.1: A) PCA: based on the SNPs located on Chromosome 1A, a principal component analysis revealed two distinct subpopulations. The distinction is given by Eigenvector One, which gave the initial evidence of inversion carriers. **B)** F_{ST}**:** these two subpopulations display highly differentiated SNPs across the whole of Chromosome 1A, except at regions near to telomeres. **C) Heterozygosity:** each subpopulation exhibits a particular heterozygosity level across the Chromosome 1A. The inv-norm subpopulation has many SNPs with high heterozygosity within the region bounded by the tentative breakpoints given by F_{ST} analysis ($\approx$3 to 68 Mb, delimited by the red dashed lines). The purple dashed line represents the maximum expected in norm-norm birds. SNPs above this threshold are considered informative.

We obtained high F_{ST} values between the two PCA plot subpopulations across almost the whole of Chromosome 1A except for the most distal SNPs on the chromosome (Figure 5.1b). The heterozygosity level in each of these subpopulations across

Chromosome 1A is also strikingly different (Figure 5.1c). The heterozygosity level for the smaller subpopulation is greater than for the larger subpopulation, except for markers close to the telomeres. This suggests that the smaller subpopulation contains birds heterozygous for the inversion polymorphism. The heterozygosity patterns are consistent with the pattern shown by the F_{ST} analysis, in terms of where the inversion is located on the chromosome. In addition, the F_{ST} values of the SNPs located on Chromosome 1A have a significantly different distribution than SNPs in the rest of the genome (Wilcoxon rank sum test with continuity correction p-value ≈ 0.0002).

The PCA, F_{ST} and heterozygosity results support the existence of a pericentric inversion in the smaller PCA subpopulation (117 birds). This putative inversion comprises ≈90% of the length of the chromosome (≈64.2 Mb) and is present only in heterozygous state in this great tit population (given the PCA clustering in addition to the high levels of heterozygosity of the SNPs at Chromosome 1A in inv-norm birds, Figure 5.1a-c).

5.3.2 Linkage-disequilibrium and haplotypes across the inversion.

We used the unphased SNP genotypes from all birds to characterize linkage-disequilibrium (LD) across Chromosome 1A by calculating D' (Lewontin, 1964). As expected for regions with low recombination, a large LD block which overlaps the whole inversion was identified (Figure 5.2a). This LD block is not present in norm-norm birds (Figure 5.2b), suggesting that recombination is only restricted in birds heterozygous for the inversion. On the other hand, when R^2 is used as a measure of LD inference, an LD block is only observed in the middle of the chromosome (from position ≈24.6 to 48.8 Mb, Figure 5.2c). This R^2 LD block overlaps the region that causes the two distinct genotype distributions among the 117 inv-norm birds (Figure 5.2d).

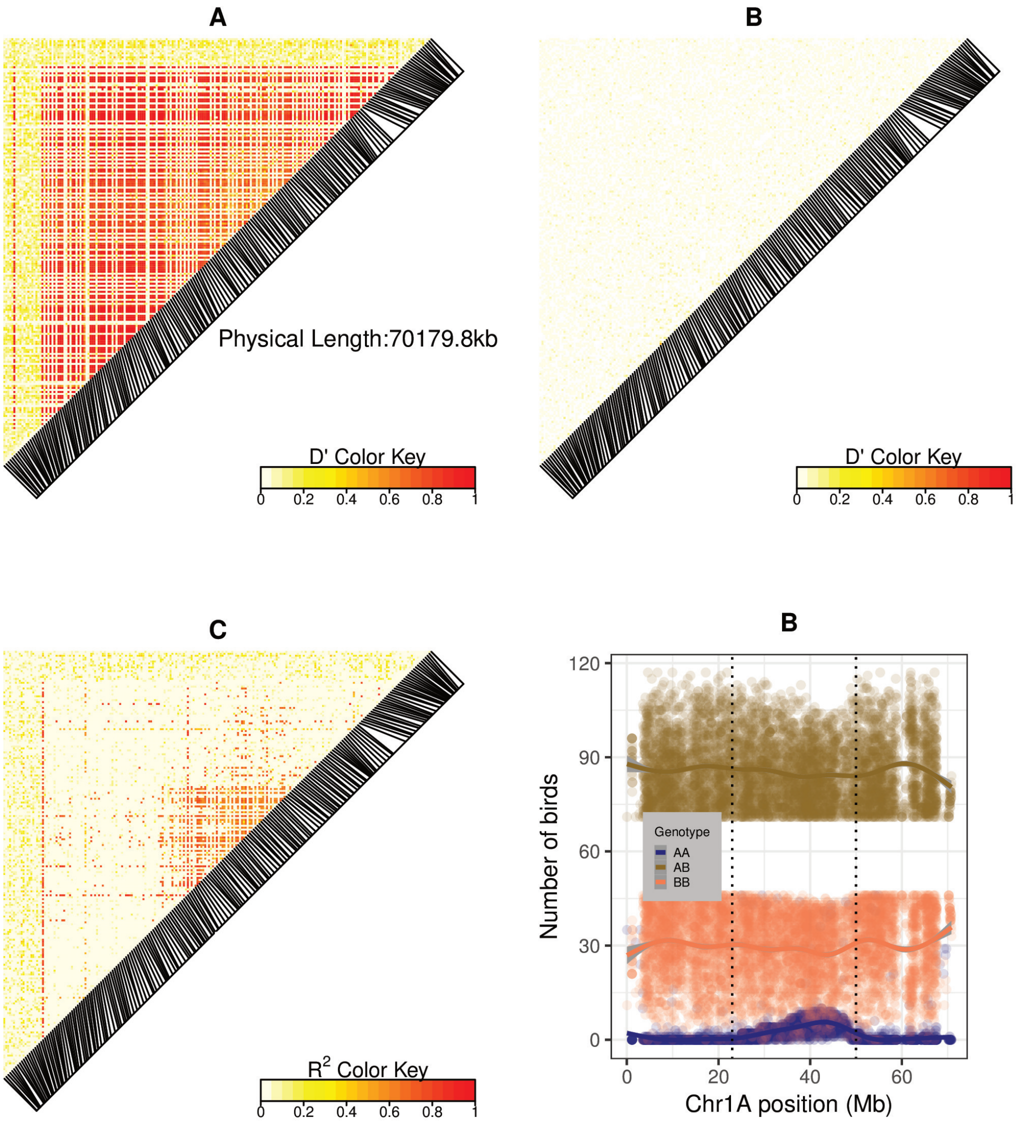


Figure 5.2: The pairwise linkage-disequilibrium on the Chromosome 1A. A) D' measured in 2,296 great tits. **B)** D' measured in 2,179 norm-norm birds. Figures in the lower panels (C and D) support possible recombination events in the center of the inversion. In other words, possible recombination in the center of the inversion is supported by the distinct genotype distribution in comparison with the rest of the inversion and confirmed by R^2. As R^2 metric has reduced power to detect LD among SNPs with low allele frequency, the LD is reflected only in the center of the inversion. **C)** R^2 measured in 2,296 great tits reveals an LD block only in the middle of the chromosome. The full inversion does not show elevated LD, due to the limitation of R^2 at dealing with low frequency SNP alleles outside the center of the inversion. **D)** Genotype frequency of informative SNPs (heterozygosity > 0.6) across Chromosome 1A in the inv-norm subpopulation. The vertical dotted line roughly indicates the genomic region of middle block which harbors a higher number of birds with "AA" genotypes when compared to the rest of the inversion. Along with the LD pattern from R^2 method, the genotype frequencies suggest a different genetic structure at the center of the inversion.

Allele phasing was not possible in the inv-norm birds as the phasing was clearly random in inv-norm birds (data not shown). Therefore, a detailed analysis of genetic diversity within the different inversion haplotypes was not possible. Instead, we used genotype information to explore putative inversion haplotypes. In the center of the inversion (a 20-55 Mb window was used, which is a 5 Mb up- and downstream extension of the LD block in the center due to uncertainty over the precise breakpoint locations), the genotype frequencies (i.e. the ratio of genotypes "AA", "AB" and "BB", where "A" is the major and "B" the minor allele in the general population) is substantially different between the ≈10% of the inv-norm birds (ten birds, Figure 5.10) and the remainder of the inv-norm birds. The number of "AA" SNP genotypes (i.e. homozygous for the major allele, which is rare in the inversion) in these ten birds is greater than in the other inv-norm birds. A total of 107 birds (91.4%) have between 4 and 30 (mean = 11.61, standard deviation = 4.95) SNPs with genotype "AA" while the remaining 10 birds have substantially more "AA" genotypes (range = 146-1,382; mean = 892.4; standard deviation = 394.2; Figure 5.3). To a certain extent the ten birds with distinct haplotypes can also be distinguished from the other inv-norm birds, by the PCA analysis due to their intermediate values in eigenvector one (0.053 to 0.076). These ten birds are from four different areas in Netherlands (2 birds from Buunderkamp; 3 birds from Westerheide; 2 birds from Roekelse Bos; 2 birds from Hoge Veluwe and one birds from an unknown location).

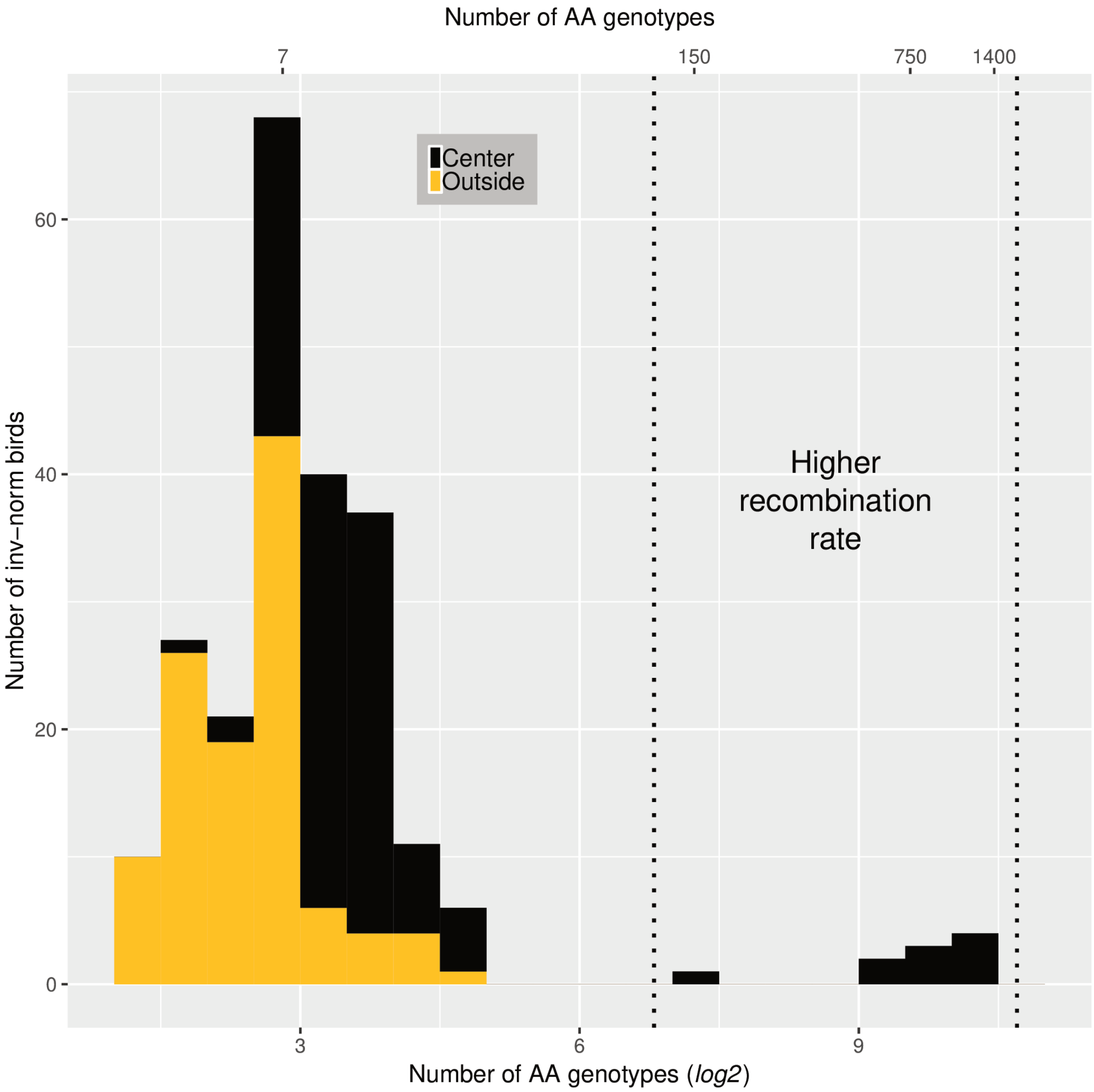


Figure 5.3: Genotype distribution within/outside the center of the inversion (20-55 Mb) in inversion carriers. The number of genotypes is represented on a log_2 scale to improve the visualization but untransformed values are shown on the upper x-axis. Based on the number of "AA" genotypes it is possible to identify inv-birds birds which harbour a different genotype distribution at the center of the inversion and therefore possibly have different inversion haplotypes (black bars among the dashed lines).

5.3.3 Complex genomic structure at the inversion breakpoint.

Inversion breakpoints can provide insight in the evolutionary history of the inversion (Sharakhov et al., 2006). The downstream breakpoint of the Chromosome 1A inversion harbors a previously identified CNV region, '2802', located at position 64.83-67.67 Mb (Figure 5.4a, da Silva et al. 2018). Of all 2,296 birds analyzed for the inversion, 2,021 were also previously analyzed for copy number variations. This includes 1,921 birds classified as norm-norm and 100 as inv-norm. Among the norm-norm birds, 217 harbor CNVs at the inversion breakpoint (11.29%) whereas 1,704 have two copies as expected in the diploid state. By contrast, 96% of the inv-norm birds have an individual CNV call mapped at the CNVR 2802. At this CNVR, 94.8% of all individual CNV calls are gains.

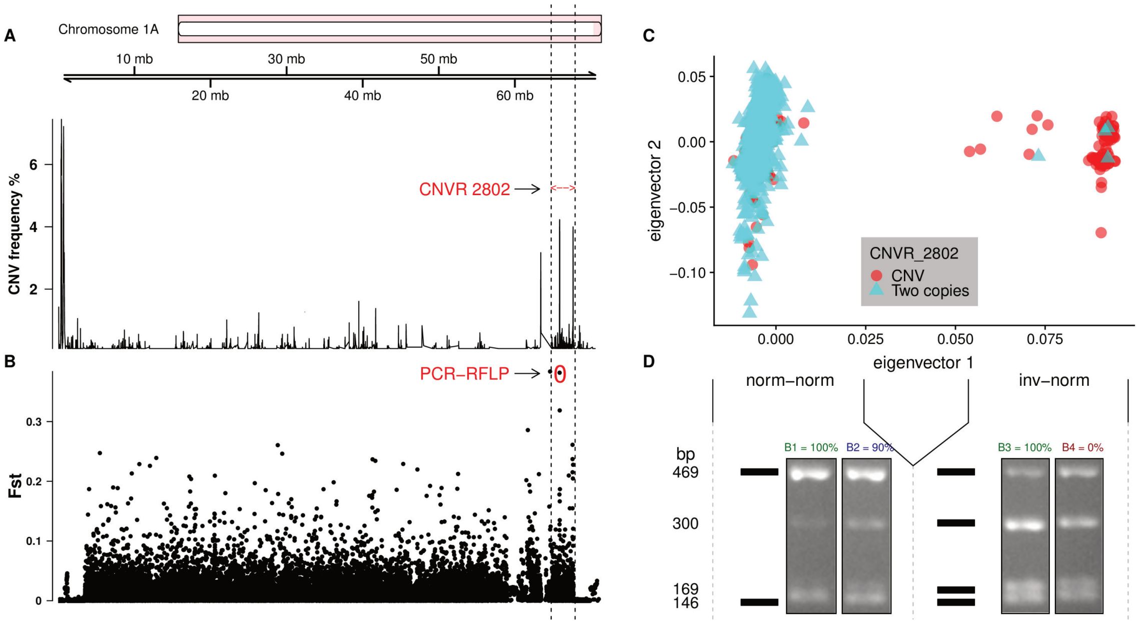


Figure 5.4: CNVs in the inversion breakpoint. A) CNV frequency across the Chromosome 1A and the genomic interval of the previously identified CNV region '2802' (≈64.83-67.67 Mb, da Silva et al. (2018)), which is located at the inversion breakpoint. **B)** F_{ST} values across the chromosome. A red circle is highlighting the SNP used to the PCR-RFLP analysis. **C)** A CNV in the inversion breakpoint is present in the vast majority of inv-norm birds whereas is rarely found in norm-norm birds. **D)** Digestion pattern of the PCR-RFLP at the SNP AX-100689781. The black bars represent the expected gel patterns alongside each of the two observed patterns in each subpopulation (i.e. norm-norm and inv-norm). Distinct copy number genotypes are evidenced by the allele intensities in the gel after electrophoresis. The values above each gel picture depicts the fingerprint name and the degree of confidence to tag a specific karyotype state (i.e. percent of the birds with concordant inversion genotype between SNP array and PCR-RFLP). Green was used in highly confident profiles, blue in the medium confidence one and red for B4, which has high heterozygosity (expected in inv-norm) but was only identified in two norm-norm birds. To differentiate between fingerprints note the distinct intensities of subsets of bands; between B1 and B2 the greatest difference is mainly at the 300/169 bp bands and between B3 and B4 the greatest difference is between the 469/300 bp bands.

5.3.4 Inversion detection with PCR-RFLP.

We looked for SNPs with the highest F_{ST} possible, which concomitantly allowed different DNA fingerprints of their SNP genotypes to be obtained by restriction digest. For the SNP with the second highest F_{ST} value (Figure 5.4b), "AA" and "AB" genotypes (i.e. associated with norm-norm and inv-norm karyotypes, respectively), our genotype assay produced two distinct *in silico* profiles when the PCR fragments were digested by the enzyme *SspI* (Figure 5.4d, represented by the black bars). In a diploid region, we would expect a profile with four bands (i.e. "AB") in an inv-norm bird whereas a profile with two bands (i.e. "AA") would be norm-norm. However, as the SNP is placed in a repetitive region (i.e. containing a CNVR and segmental duplications), the obtained profiles are more complex. We obtained instead four different profiles, which differ in the intensity in each of the four possible fragments (Figure 5.4d). Profile B3 was only identified in inv-norm samples whereas the profiles B1, B2 and B4 were mostly, but not exclusively observed in norm-norm samples. However, birds with the profile B2, in 90% of the cases, are norm-norm and in 10% inv-norm. Unexpectedly, the profile B4, which shows high heterozygosity as in the inversion, was only identified in two norm-norm birds (0% of confidence, i.e. expected to be found in inv-norm but only found in norm-norm birds). The SNP is located in the first intron of the *PIK3C2G* gene.

5.3.5 Assessing breakpoint complexity from sequencing data.

We classified 29 birds for the inversion from distinct European populations by whole genome resequencing (Laine et al., 2016) based on the presence of the CNV complex at the breakpoint. A total of 27 birds were classified as norm-norm and two as inv-norm. We used sequencing data from the two inv-norm birds, one from France and another from Belgium, to characterize CNVs across the inversion. At the downstream breakpoint, we detected a CNV (gain state) in both birds in agreement with the results from the Dutch great tit population, which suggests a high correlation of the inversion with a gain state at the downstream breakpoint (Figure 5.4c). None of the other 27 resequenced birds without the inversion showed CNVs at this region. The CNVs that we identified in the two inv-norm resequenced birds point to a substantial increase in the number of copies instead of only a single copy gain. The log_2 values from CNV-seq at that region suggest around ten copies in the inverted phase involving three CNVs that are part of the same structural complex (the regions between 65.87-65.90, 67.56-67.58 and 67.64-67.65 Mb, which together comprise $\approx$50.43 kb). In addition, we identified an increase of around 100 copies in a region upstream to the CNV complex (63.44-63.46 Mb, $\approx$20 kb), which in turn is followed by an increase of around ten copies (63.46-63.56 Mb, $\approx$100 kb). It is unclear if these events are part of the same complex (Sup Fig 4 shows the estimated number of copies in each of the above mentioned CNV regions). Considering only the three CNVs which are part of the complex, the inverted Chromosome 1A is at least 500 kb larger than the reference (i.e. the normal non-inverted) haplotype. However, summing the CNV complex with other upstream CNV regions that are also only present in sequenced inv-norm birds (i.e. a region with ≈100 copies followed by other regions with ≈10 copies), suggests that the inverted chromosome may be up to 3.5 Mb larger than the normal chromosome.

As split reads from sequencing data are useful to reveal complex rearrangements in the genome, we evaluated their pattern in the CNVR. We identified split reads in this region that support a complex genomic rearrangement involving different CNVs. Split reads and discordantly mapped paired reads show that this region contains a complex rearrangement of three intervals which are arranged in a different order and orientation when compared to the reference genome (supplementary section patterns in split reads supporting the CNV complex, Figure 5.5).

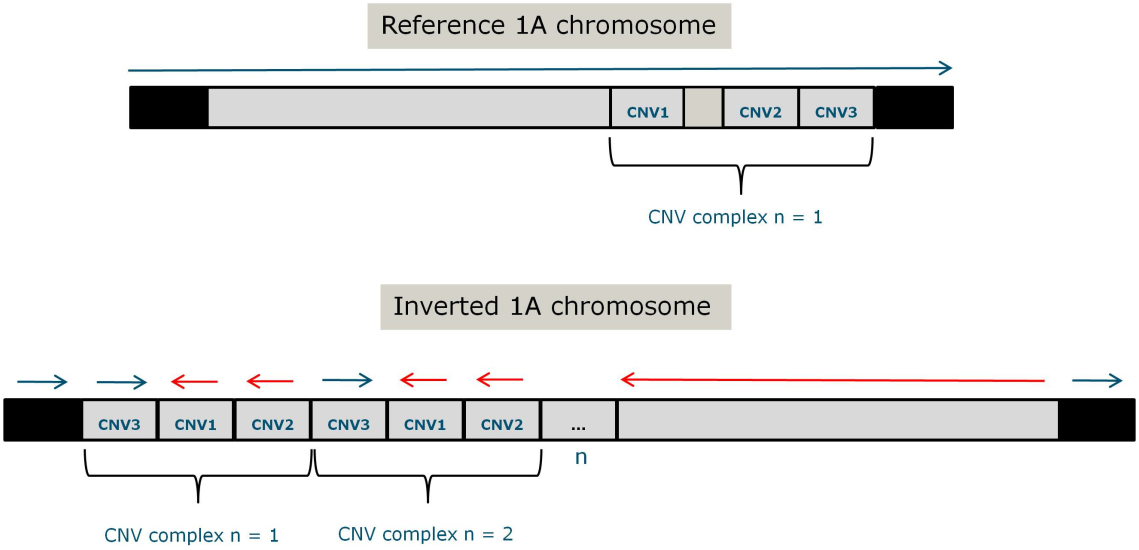


Figure 5.5: Representation of the whole Chromosome 1A with the complex structural rearrangement in the downstream breakpoint of the inversion. Blocks in grey represent the inversion region whereas those in black are genomic regions outside the inversion. CNVs identified by sequencing in the two inv-norm birds which were sequenced are labeled as CNV1-3 for simplicity. Horizontal curly brackets define the structural complex which encompasses CNVs 1-3. The above chromosomal representation displays the chromosome as shown in the reference genome (Laine et al., 2016). The below representation displays the expected genomic structure in the inversion. CNVs are relatively larger than their real length for schematic purposes.

In addition, Lumpy (Layer et al., 2014) was used to predict the exact breakpoints of the inversion. We were unable to infer the whole inversion event from sequencing data, but interestingly one large inversion was unique to the two inv-norm samples that were sequenced. The inversion boundaries are from 62.15 to 63.55 Mb, with a length of 1.4 Mb on the reference genome. For the two inv-norm samples, 9 (sample name = 233) and 8 (sample name = 973) reads supported this 1.4 Mb inversion event. The coordinates of the inversion start lies within a single copy region, while the coordinates of the inversion end are located in the CNV complex (65.87-67.65 Mb). Therefore, we hypothesize that at least one of the inversion breakpoints is within the large complex; however, the precise coordinates are difficult to predict.

5.3.6 Gene content and functionality at the inversion breakpoint.

Genomic regions around the inversion breakpoints can have a different structure and nucleotide diversity compared to the rest of the inversion (Andolfatto et al., 2001; Branca et al., 2011; Hoffmann & Rieseberg, 2008). The CNV complex overlaps 32 genes associated with a broad range of phenotypes in other species (for details on the phenotypes associated with each gene see supplementary section 3.3 Genes overlapping the CNVR at the CNV complex). It is perhaps noteworthy that three genes (*BPGM*, *CALD1* and *PIK3C2G*) could potentially be broken in the inverted haplotype, given that sequencing data shows CNVs only partially overlapping them.

5.4 Discussion

Here we have described a large putative inversion on Chromosome 1A of the great tit (Bosse et al., 2017) that covers more than 90% of the chromosome and contains almost 1,000 genes. The inversion is present in 5% of the analyzed Dutch population as well as in two out of 29 resequenced individuals from other European populations; one carrier was from Belgium and the other from France, indicating that the inversion is present in other great tit populations as well. In this study, the inversion was analyzed with a SNP array and by shotgun sequencing. Although the most likely explanation for suppressed recombination is an inversion (Kirkpatrick, 2010), we acknowledge that methods such as FISH (Bishop, 2010) and long read sequencing (Shao et al., 2018) need to be used to confirm the inversion hypothesis. It is feasible, though unlikely given the size of the region, that suppressed recombination leading to chromosomal divergence could arise without a chromosomal inversion (Bergero et al., 2007, 2008, 2013; Natri et al., 2013). For clarity in this discussion we refer to the putative inversion found here simply as inversion.

In the Dutch population, among the 2,296 birds analyzed after filtering, no homozygous bird was found. Given that very large inversions can cause homozygous lethality in songbirds (Tuttle et al., 2016), we investigated if this great tit population has significantly fewer homozygous inverted birds than expected. However, given the low frequency of the inversion, and assuming Hardy-Weinberg Equilibrium, we would expect less than two homozygous inverted birds and it is thus unclear whether the complete absence of homozygotes is due to a deleterious recessive effect of the inversion or whether homozygotes are present in the population but not sampled in this study. A possible lethal effect of this inversion could be tested by exploring the frequency of genotypes among offspring of mated carriers. Given the structural complexity and large size of this inversion, a relevant biological effect could be expected. A CNV complex located at the downstream breakpoint encloses 32 genes involved in a wide range of biological processes, which could significantly change the amounts of the transcripts/proteins due to copy number changes in the genes located at the CNV complex. Future studies of this inversion polymorphism will be directed to test the lethality hypothesis and to measure the relative fitness of wildtype homozygotes, inversion carriers and inversion homozygotes. Indeed, this future goal was one motivation for developing a cheap and quick method (based on PCR-RFLP) to more easily type inversion karyotypes.

To identify the inversion without SNP array data, we selected the SNP with highest F_{ST} value that concomitantly would produce a PCR-RFLP profile capable of distinguishing between inversion carriers and non-carries. The selected SNP is located at the first intron of the *PIK3C2G* gene, which is within the CNV complex at one of the putative inversion breakpoints. Along with *PIK3C2G*, several other genes are also located in the CNV complex and these genes have crucial roles in a broad range of processes from cell cycle to gene silencing (supplementary section 3.3 Genes overlapping the CNVR at the CNV complex). Resequenced birds showed a high number of copies within that genomic region (≈10 copies in two inv-norm birds). Moreover, the PCR-RFLP gel intensities support at least four genotypes (three for norm-norm and one for inv-norm birds). Thus, this substantial copy number change in inv-norm birds could underlie distinct patterns in gene expression and consequently phenotypic variation. Interestingly, such complex rearrangements at inversion breakpoints have a key evolutionary roles in other species e.g. an effect on malaria vectorial capacity in mosquitoes (Sharakhov et al., 2006).

A CNV complex located at the breakpoint seems to be older than the inversion. Assuming a single origin for this complex, the CNV sequences may be older than the inversion given that it is present in virtually all inv-norm birds whereas it occurs at low frequency in norm-norm birds. More than 10% of the norm-norm birds have at least one CNV overlapping the CNV complex. In addition, a repetitive structure is usually found at inversion breakpoints underlying their mechanisms of formation (such as non-allelic homologous recombination - NAHR, Hoffmann & Rieseberg (2008); Carvalho & Lupski (2016)). Thus, it is possible that the inversion is a result of the CNV sequences, which underpinned the mechanism of the inversion formation. However it remains possible that CNVs are present in the inversion only due to a 'hitchhiking' effect and thus did not necessarily contribute to the inversion's formation. The hypothesis that CNVs might have underpinned the formation of the inversion remains speculative and needs further investigation. Considering the size of all CNVs associated with the inversion (i.e. complex with $\approx$10 copies and another complex of $\approx$10 copies with an additional region with $\approx$100 copies, identified by sequencing) the inverted chromosome is estimated to be approximately 3.5 Mb larger than the reference sequence reported in genome build 1.1. The greater length of chromosomes harboring the inversion is in line with the hypothesis of degenerative expansion in young supergenes (Stolle et al., 2018). However, genetic variation is not only present in the CNV complex but also at the center of the inversion.

Allele phasing in inv-norm birds is challenging because phasing strategies like BEAGLE assume Hardy-Weinberg equilibrium Browning & Browning (2007); this assumption is often violated at inversion genotype-informative SNPs (i.e the vast majority of the genotype-informative SNPs significantly deviate from HWE). Thus, we used the genotype distribution (i.e. the proportions of "AA", "AB" and "BB" genotypes) to partially explore the haplotypes in the inversion. There are at least two (and perhaps three or more) putative inversion haplotypes, which are reflected by the number of AA genotypes at the center of the inversion (located at $\approx$20-55 Mb of the Chromosome 1A, Figure 5.3, note the log scale and three distinct groups). In the LD analysis, only the R^2 metric reflected the variation within inv-norm birds due to SNPs in a block in the center of the inversion. The R^2 method has a constraint to deal with low frequency alleles (Wray, 2005) whereas D' is not highly dependent upon allelic frequencies (Hedrick, 1987). Interestingly, in the inv-norm population, the frequency of the less common genotype in the informative SNPs at the R^2 LD block (Figure 5.2a) is not as low as in the rest of the inversion (Figure 5.2b). Thus, the distribution of allele frequencies in the inv-norm birds may explain why the R^2 metric does not describe elevated LD, outside the center of the inversion, and is consistent with the hypothesis of a higher recombination rate in the center. In other words, because the two different LD measures are not equally sensitive to rare alleles, and because the allele frequencies seem to be different in the center of the inversion than elsewhere, one metric finds a pattern that the other misses. Presumably this is because occasional recombination has caused allele frequencies and LD patterns to be slightly different in the center than in the rest of the inversion. Due to the expected very low rates of recombination within the inversion in heterozygotes (Kirkpatrick, 2010), we did not expect multiple haplotypes for the inversion. However, on timescales of 10^5 generations or longer, even this limited recombination works as an important source of variation within inversions (Kirkpatrick, 2010). Indeed, gene conversion and multiple crossing overs, at least far from the breakpoints, are possible within inversions (Andolfatto et al., 2001; Hoffmann

& Rieseberg, 2008; Korunes & Noor, 2018). Thus, rare recombination events may explain distinct haplotypes found in the center of the inversion. Moreover, as CNVs can underlie mechanisms of formation and be prone to errors, independent inversion events and errors during meiosis cannot be discarded.

It is unclear whether the inversion has any phenotypic effects. Nevertheless, the CNVs identified by sequencing at the CNV complex directly overlap at least three genes, including *CALD1* involved in smooth muscle contraction (Walsh, 1994), *BPGM* underlying oxygen sensing in blood cells (Petousi et al., 2014) and the above mentioned *PIK3C2G* gene (the other 29 genes overlap a CNVR in the same region but do not overlap partially CNVs identified by sequencing). In other songbird species, such as zebra finches (*Taeniopygia guttata*), sperm morphology and motility is associated with an inversion in the Z Chromosome (Kim et al., 2017). Moreover, inversions in zebra finches can have strong additive effects on several morphological traits and increase mortality rates (Knief et al., 2016). In white-throated sparrows, which display different plumage morphs and sexual behavior, a large inversion involving up to 1,000 genes and lethal in its homozygous state, has a profound role in disassortative mating (Tuttle et al., 2016). However, there is no evidence of distinct morphs in great tit. Thus, if the inversion is underlying any kind of mate choice it may be reflected by a more subtle trait or behaviour.

5.5 Conclusions

Apart from songbirds, large inversions can underlie a number of phenotypes in nature, ranging from mimicry and crypsis in butterflies and moths (Nadeau et al., 2016) to meiotic drive in mice (Lyon, 2003). Our detailed characterization of the variability and complexity of this large inversion provides the foundation for further studies aiming to discover the phenotypic effects and the evolutionary role of this inversion.

5.6 Supplemental material

5.6.1 Supplemental figures

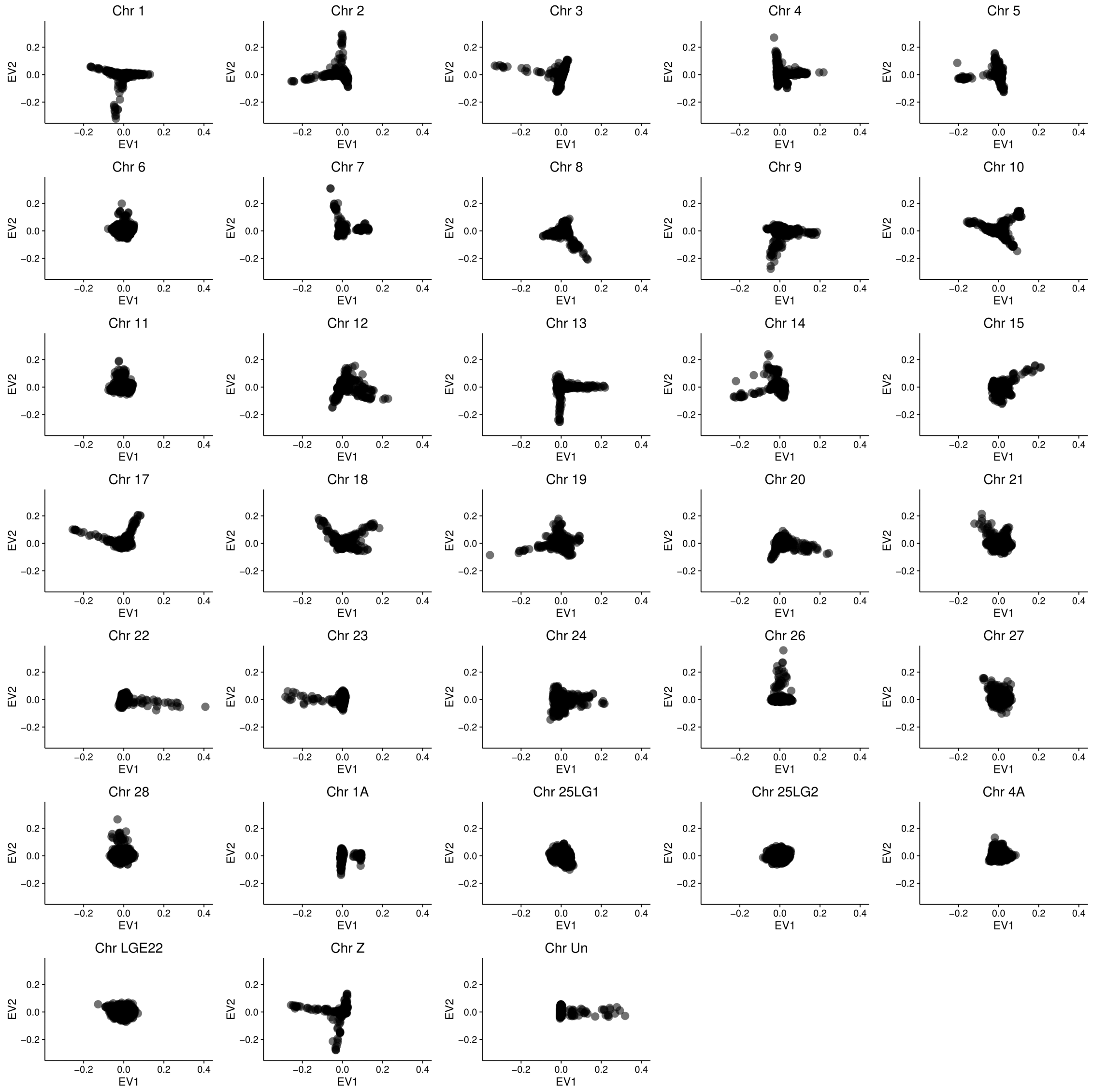


Figure 5.6: PCA for all autosomes in the great tit genome build 1.1.

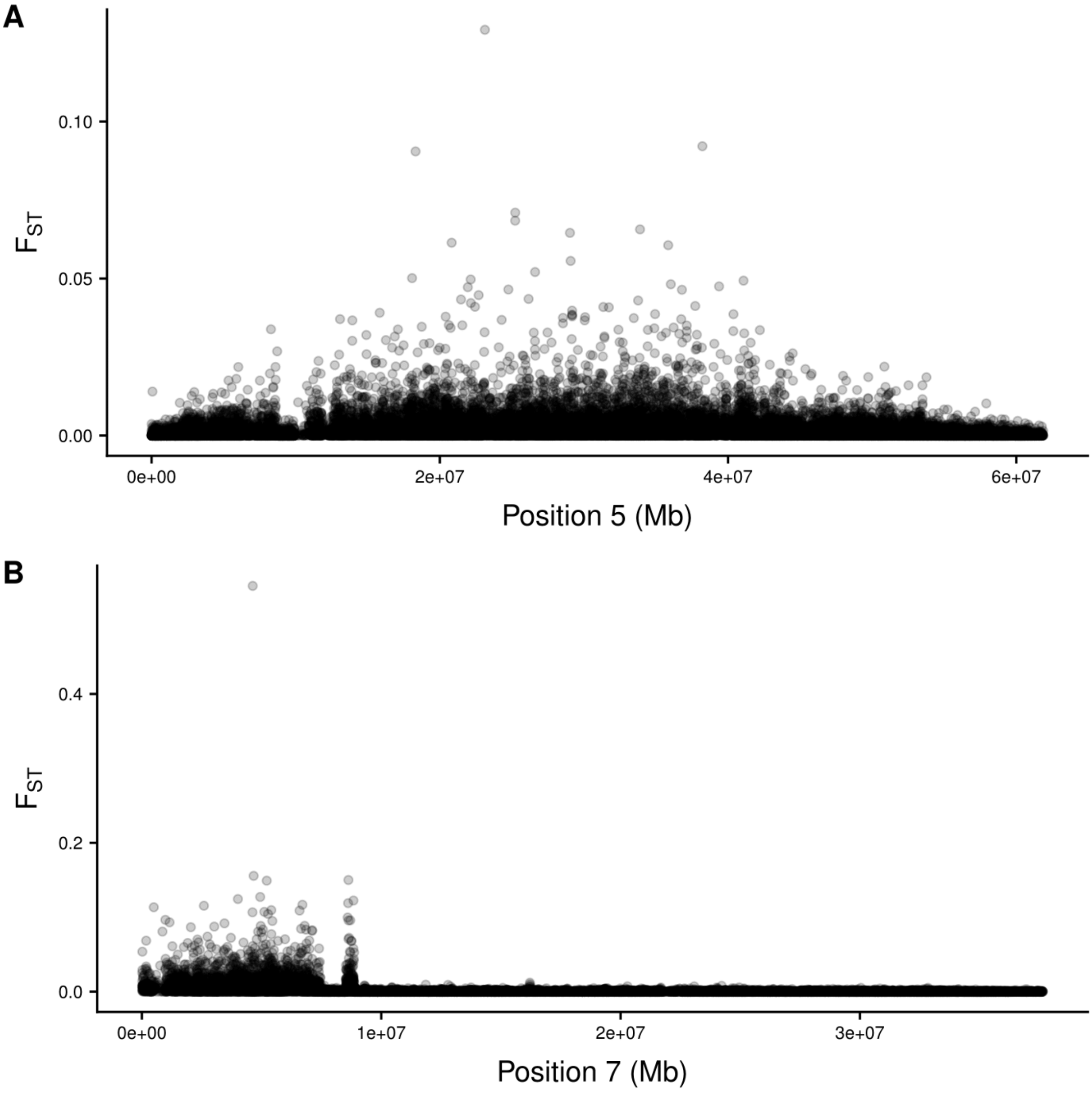


Figure 5.7: A-) F_{ST} across the Chromosome 5. B-) F_{ST} across the chromosome 7.

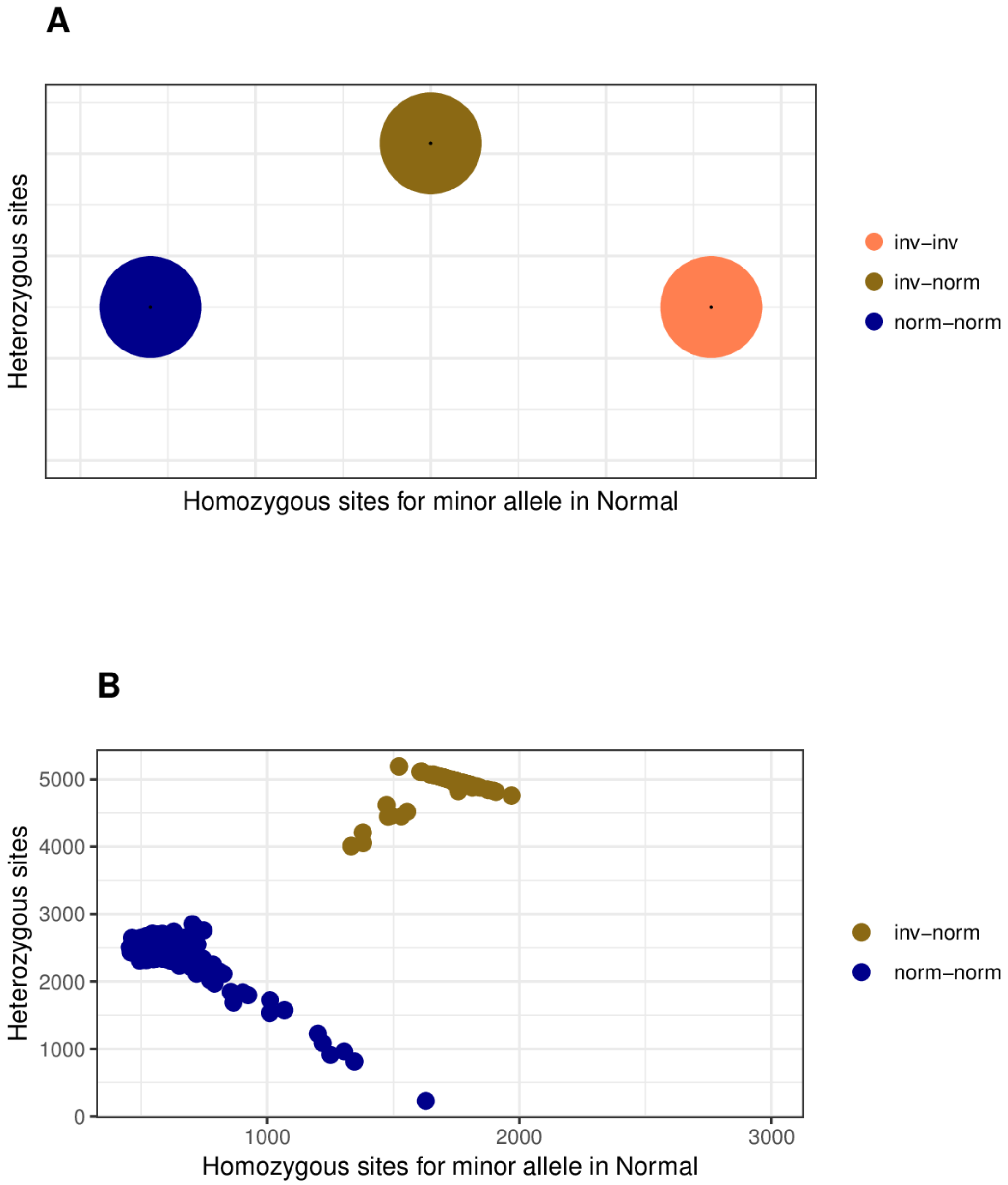


Figure 5.8: Cluster patterns, using all informative SNPs on Chromosome 1A, in each of the possible diploid karyotypes of a chromosome-wide inversion (i.e. norm-norm in dark blue, inv-norm in brown and inv-inv in orange, from left to right). The x-axis is the count trend of each karyotype for homozygous SNPs for the alternative allele in the normal phase. The y-axis is the count trend of each karyotype for heterozygous SNPs. Therefore, the expectations presented in the upper panel are based on the following assumptions: (i) inv-norm birds should have higher number of heterozygous SNPs across the chromosome 1A in comparison with inv-inv or norm-norm and (ii) inv-norm birds should have an intermediate number of homozygous SNPs for the minor allele in norm (i.e. “BB”) in comparison with inv-inv or norm-norm. **A)** Expected clustering patterns. **B)** Cluster results from 2,296 great tits which were colored based on the classification from PCA analysis.

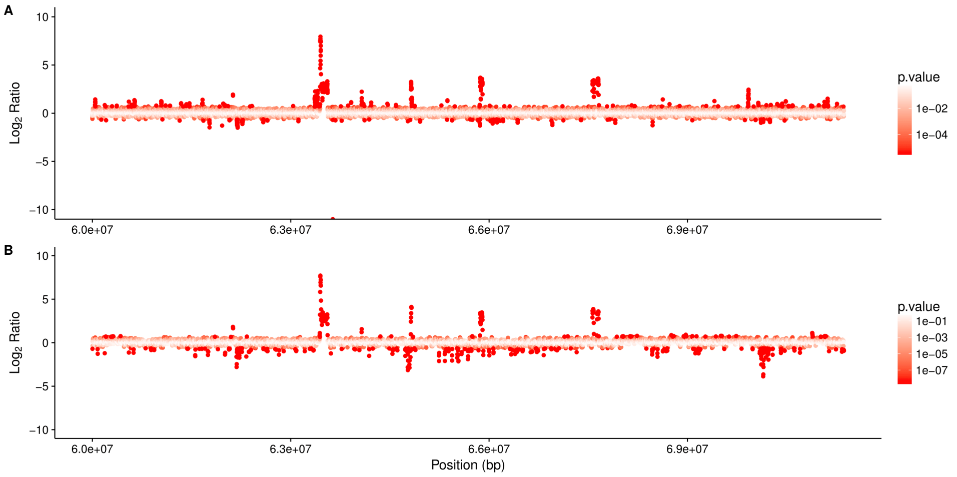


Figure 5.9: The x-axis represents the genomic coordinates of the CNV complex (i.e. downstream the inversion breakpoint) whereas the y-axis display the log_2 ratio that reflects the relative copy number across the complex (relative to a norm-norm bird). Thus, the anti-log of the log_2 ratio can be roughly interpreted as the absolute number of copies (i.e. if log_2 ratio = 3.333, then the anti-log is $2^{3.333}$ = ≈10 copies). A and B show respectively a female from France and a male from Belgium, which were classified as inv-norm based on sequencing data.

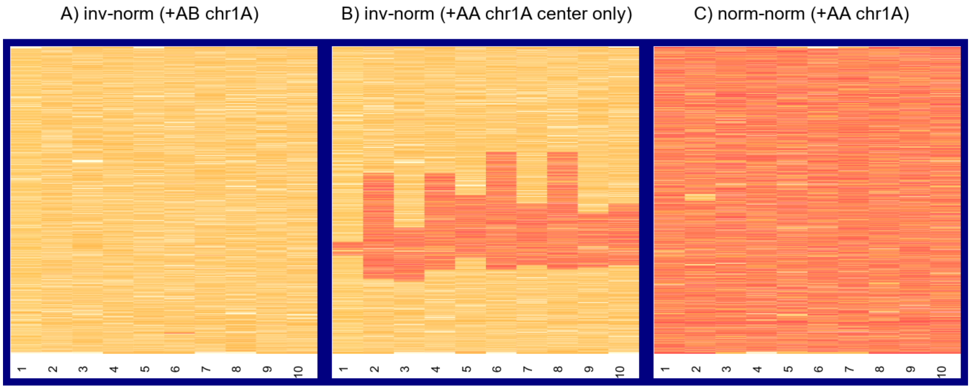


Figure 5.10: We used 4,124 informative SNPs (i.e. heterozygosity >0.6 in the inv-norm subpopulation), which are located in the center of the Chromosome 1A (20-60 Mb), to display the different inversion genotypes distributions in a heatmap. The SNP genotypes are represented by white ("BB"), light orange ("AB") and dark orange ("AA"), respectively. The distinct number of "AA" genotypes in the center of the inversion suggests different haplogroups in approximately 10% of the inv-norm birds (i.e. ten birds). **A)** Ten inv-norm birds selected randomly. **B)** Ten inv-norm birds displaying a distinct genotype distribution at the center of the inversion. **C)** Ten norm-norm birds selected randomly.

5.6.2 Supplementary methods

5.6.3 Classification confirmation for inversion carriers

Although PCA analysis is expected to produce clusters that distinguish inversion karyotypes due to genetic differentiation (i.e. both phases with the inversion, only one or absence of the inversion in both), we confirmed the inversion karyotypes using two sources of information. (i) Number of heterozygous SNPs and (ii) number of homologous SNPs for the minor allele in the normal phase, which are expected to form independent clusters for each inversion genotype in a scatter (XY) plot. For this confirmation strategy, we only used SNPs with heterozygosity value >0.6 in the subpopulation with higher values at eigenvector one (i.e. classified as inv-norm by PCA analysis). Therefore, we reclassified the birds as (i) norm-norm, (ii) inv-norm and (iii) inv-inv based on the XY plot for comparison with PCA classification.

Selection of the SNP used in the RFLP-PCR

All the SNPs supporting the inversion in the chromosome 1A were ranked by F_{ST} value. Thus, possible RFLP-PCR essays were simulated with the R/Bioconductor package DECIPHER (Wright, 2016). The SNP AX-100689781 had the second highest F_{ST} value overall, but had the higher F_{ST} value among possible assays and was then carried forward for the subsequent primer design and enzyme search.

Primer design and enzyme search

In order to design a primer pair and pick a restriction enzyme which is able to differentiate genotypes at SNP AX-100689781, we first imported the reference sequence genome build 1.1 (Laine et al., 2016) with `readDNAStringSet` function from Biostrings R/Bioconductor package (v. 2.44.2) (Pagès et al., 2017). The sequence around the SNP was extracted and then written with `writeXStringSet` function, which is also available in Biostrings package. The candidate restriction enzyme was selected using the group-specific signatures pipeline available in the R/Bioconductor package DECIPHER manual (Wright, 2016). The primers were designed using Primer3plus (Untergasser et al., 2007) and their quality was tested by NetPrimer (`http://www.premierbiosoft.com/netprimer`. The full nucleotide sequence of the amplicon (615 bp) can be copied directly from <NCBI>. The genotype-specific cutting patterns on the PCR amplicon (i.e. generated with the primers in Sup Table 5.1) after digestion by the SspI enzyme is exemplified in the Sup Figure 5.11. The DNA of the selected animals was checked for quality and quantity with Qubit® Fluorometer.

Table 5.1: Primers used in the PCR-RFLP analysis.

	Sequence
Forward	GCCAGGCTCCTTAACATTTTG
Reverse	TCAGAGGGAACTGGATCTGC

5.6.4 Supplementary results

Identification of the inversion carriers

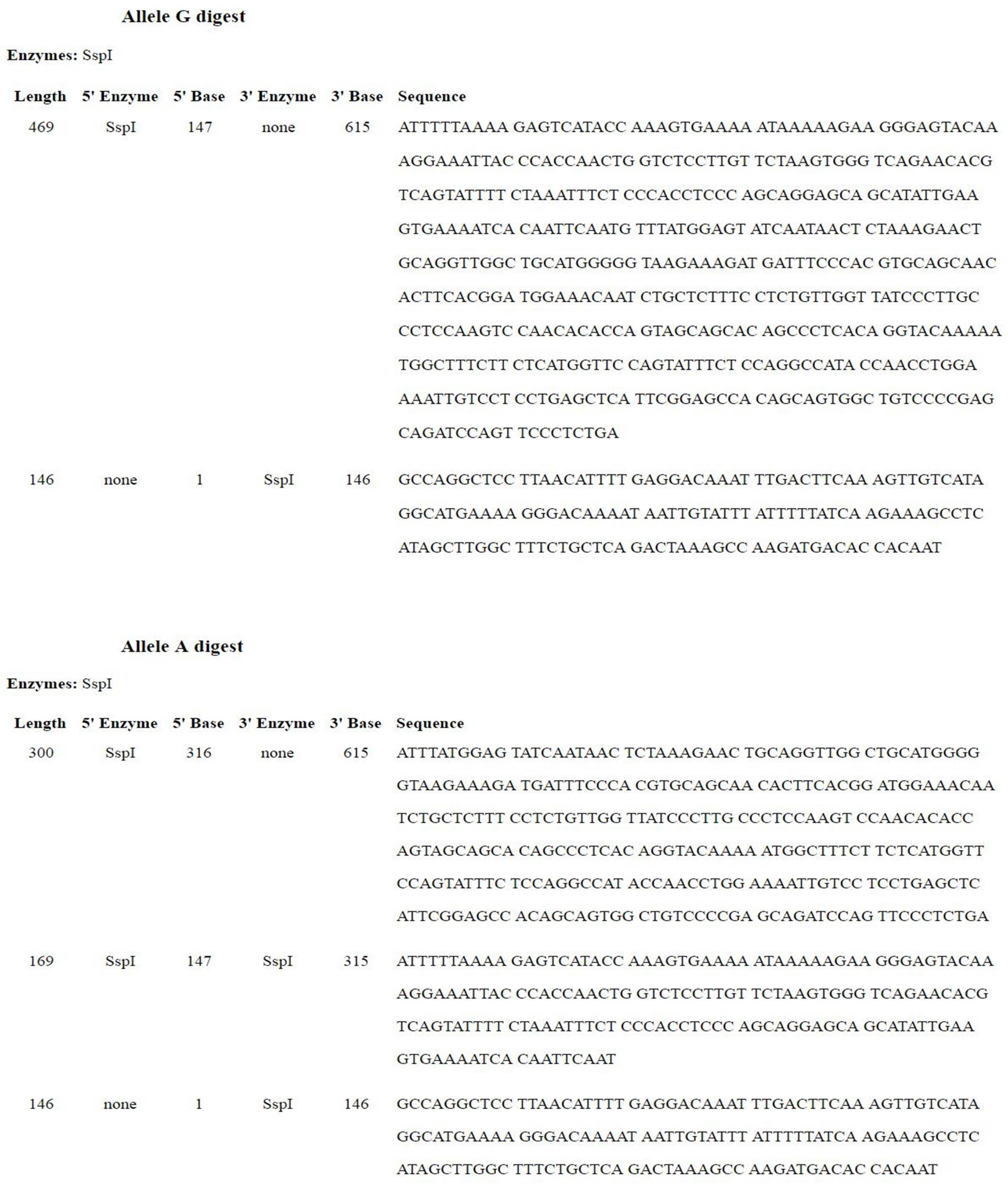
We performed an additional test which relies on the assumption that informative SNPs should cluster birds with the same karyotype, based on the relative number of heterozygous SNPs and SNP genotypes homozygous for the minor allele in the normal phase (Sup Figure 5.8a). Thus, we classified the samples into (i) no inversion as norm-norm (ii) one inverted phase as inv-norm and (iii) two inverted phases as inv-inv (not found in this population) as in the PCA test. The test reflected the PCA clustering results and we therefore classified 117 birds as inv-norm and 2,179 as norm-norm (Sup Figure 5.8b).

Quality of the SNPs used in the LD analysis

To make sure that the high incidence of “AA” genotypes in the center of the inversion for some inv-norm birds is not due to low quality markers, we compared the consistency of genotypes in the reference genome animal which was genotyped twice. We split chromosome 1A into 500 tiles (≈140kb each) and estimated the percentage of concordant genotypes in both assays for each tile. We could not find any indication of low quality SNPs within the R^2 LD block (i.e no lower genotyping quality in the center of the chromosome, t-test p-value = 0.84).

Genes overlapping the CNVR at the CNV complex

The SNP within the CNV complex, used for inversion detection by PCR-RFLP (high F_{ST} value within the inversion), is placed at the first intron of the *PIK3C2G* gene which has crucial role on signaling pathways (Rozycka et al., 1998). Nevertheless, the CNV complex in the inversion breakpoint is a gene-rich genomic interval that encompasses 32 genes (16 with known gene names) that are related to a wide range of processes (Sup Table 5.2). These genes or its paralogs translate proteins involved in the cell cycle (*PDE3A*, *RERG* and *PIK3C2G*) (Begum et al., 2011; Zhao et al., 2017; Rozycka et al., 1998), protein trafficking (*PIK3C2G*) (Rozycka et al.,



Allele G digest

Enzymes: SspI

Length	5' Enzyme	5' Base	3' Enzyme	3' Base	Sequence
469	SspI	147	none	615	ATTTTTAAAA GAGTCATACC AAAGTGAAAA ATAAAAAGAA GGGAGTACAA AGGAAATTAC CCACCAACTG GTCTCCTTGT TCTAAGTGGG TCAGAACACG TCAGTATTTT CTAAATTTCT CCCACCTCCC AGCAGGAGCA GCATATTGAA GTGAAAATCA CAATTCAATG TTTATGGAGT ATCAATAACT CTAAAGAACT GCAGGTTGGC TGCATGGGGG TAAGAAAGAT GATTTCCCAC GTGCAGCAAC ACTTCACGGA TGGAAACAAT CTGCTCTTTC CTCTGTTGGT TATCCCTTGC CCTCCAAGTC CAACACACCA GTAGCAGCAC AGCCCTCACA GGTACAAAAA TGGCTTTCTT CTCATGGTTC CAGTATTTCT CCAGGCCATA CCAACCTGGA AAATTGTCCT CCTGAGCTCA TTCGGAGCCA CAGCAGTGGC TGTCCCCGAG CAGATCCAGT TCCCTCTGA
146	none	1	SspI	146	GCCAGGCTCC TTAACATTTT GAGGACAAAT TTGACTTCAA AGTTGTCATA GGCATGAAAA GGGACAAAAT AATTGTATTT ATTTTTATCA AGAAAGCCTC ATAGCTTGGC TTTCTGCTCA GACTAAAGCC AAGATGACAC CACAAT

Allele A digest

Enzymes: SspI

Length	5' Enzyme	5' Base	3' Enzyme	3' Base	Sequence
300	SspI	316	none	615	ATTTATGGAG TATCAATAAC TCTAAAGAAC TGCAGGTTGG CTGCATGGGG GTAAGAAAGA TGATTTCCCA CGTGCAGCAA CACTTCACGG ATGGAAACAA TCTGCTCTTT CCTCTGTTGG TTATCCCTTG CCCTCCAAGT CCAACACACC AGTAGCAGCA CAGCCCTCAC AGGTACAAAA ATGGCTTTCT TCTCATGGTT CCAGTATTTC TCCAGGCCAT ACCAACCTGG AAAATTGTCC TCCTGAGCTC ATTCGGAGCC ACAGCAGTGG CTGTCCCCGA GCAGATCCAG TTCCCTCTGA
169	SspI	147	SspI	315	ATTTTTAAAA GAGTCATACC AAAGTGAAAA ATAAAAAGAA GGGAGTACAA AGGAAATTAC CCACCAACTG GTCTCCTTGT TCTAAGTGGG TCAGAACACG TCAGTATTTT CTAAATTTCT CCCACCTCCC AGCAGGAGCA GCATATTGAA GTGAAAATCA CAATTCAAT
146	none	1	SspI	146	GCCAGGCTCC TTAACATTTT GAGGACAAAT TTGACTTCAA AGTTGTCATA GGCATGAAAA GGGACAAAAT AATTGTATTT ATTTTTATCA AGAAAGCCTC ATAGCTTGGC TTTCTGCTCA GACTAAAGCC AAGATGACAC CACAAT

Figure 5.11: Restriction enzyme digestion of the PCR amplicon considering a 2n state on the target region (diploid). As the region being analyzed mostly deviates from 2n, the real patterns may diverge in signal intensity as well. As the GG and AG genotypes represent mostly norm-norm and inv-norm respectively, norm-norm and inv-norm birds are expected to show two and four fragments respectively.

1998), muscle contraction (*CALD1*) (Walsh, 1994), recurrent translocation in cancer (*LMO3*) (Chambers & Rabbitts, 2015), spliceosome activity (*STRAP*) (Seong et al., 2005; Chari et al., 2008), brain development (*PLEKHA5*) (Yamada et al., 2012), glucose metabolism (*IAPP*) (Mulder et al., 1996), oxygen sensing in blood cells (*BPGM*) (Petousi et al., 2014), fat production (*MGST1*) (Littlejohn et al., 2016), signalling (*EPS8* and *RERGL*) (Lanzetti et al., 2000; Colicelli, 2004), solute transport (*SLC15A5*) (Hoglund et al., 2011), synapse formation and apoptosis (*PTPRO*) (Jiang et al., 2017; Liang et al., 2017), energy metabolism (*DERA*), (Salleron et al., 2014) and even pigmentation by affecting Polycomb activity (*AEBP2*) (Grijzenhout et al., 2016; Kim et al., 2011), which is a key process in gene silencing (Golbabapour et al., 2013).

To make sure the higher rate of informative SNPs at the CNV complex is not driven by low quality genotypes at this region, we compared the percentage of consistent genotypes at the complex with the genotypes in other regions of the chromosome 1A. We found no significant difference (t-test, p-value = 0.75), what suggests that the number of false positives in this region is not higher than other regions in the chromosome 1A.

Table 5.2: Genes overlapping the CNV complex at the downstream breakpoint of the inversion.

Chromosome	Start	End	Width	Name
chr1A	64843171	64844337	1167	LOC107204104
chr1A	64861670	64908113	46444	LOC107205143
chr1A	64874841	64878856	4016	IAPP
chr1A	64919923	64938780	18858	LOC107205182
chr1A	64947738	64989258	41521	LOC107204204
chr1A	64999708	65223576	223869	PDE3A
chr1A	65224970	65233165	8196	LOC107205022
chr1A	65236702	65339065	102364	LOC107205021
chr1A	65274559	65279283	4725	LOC107205023
chr1A	65355652	65396498	40847	LOC107204113
chr1A	65516912	65560642	43731	AEBP2
chr1A	65577008	65743662	166655	PLEKHA5
chr1A	65862206	66091155	228950	PIK3C2G
chr1A	66109620	66118841	9222	RERGL
chr1A	66427883	66437729	9847	LOC107204286
chr1A	66557323	66617748	60426	LMO3
chr1A	66647333	66649964	2632	LOC107204290
chr1A	66674727	66682085	7359	MGST1
chr1A	66709327	66739543	30217	SLC15A5
chr1A	66789556	66833259	43704	DERA
chr1A	66836525	66844259	7735	STRAP
chr1A	66845766	66857357	11592	LOC107204111
chr1A	66873268	67003015	129748	EPS8
chr1A	67004993	67150264	145272	PTPRO
chr1A	67023437	67032017	8581	LOC107204503
chr1A	67191246	67291500	100255	RERG
chr1A	67330974	67366580	35607	LOC107204153
chr1A	67377799	67401512	23714	LOC107204567
chr1A	67400647	67409947	9301	LOC107204566
chr1A	67410594	67581825	171232	CALD1
chr1A	67622020	67640854	18835	LOC107204149
chr1A	67646418	67680793	34376	BPGM

Patterns in split reads supporting the CNV complex

We manually checked the reads overlapping CNVs which are located nearby to the downstream breakpoint of the inversion (Sup Table 5.3). Interestingly, we found read pairs at the breakpoints of the CNVs 1, 2 and 3 to support their structural rearrangement into a CNV complex (Sup Figure 5.12). However, although the inversion breakpoint is relatively clear in the SNP-array based results (Figure 5.1), CNVs identified with sequencing data indicate that the inversion breakpoint may be placed at the CNV complex. These CNVs belonging to the CNV complex are nearby to gaps in the reference genome, which adds another layer of complexity to the interpretation of these variants. Moreover, it is not completely clear how the $\approx$ 10 copies of the complex are distributed across the genome (e.g. *in tandem* or not). Thus, the actual boundaries of the inversion might differ from the breakpoints found in SNP array results.

Table 5.3: Sequencing coverage in two inv-norm birds

CNV id	CNV location	PHRED quality	French coverage	Belgium coverage
CNV1	65.87-65.90	8677.93	112.832	86.658
CNV2	67.56-67.58	8352.07	110.254	102.649
CNV3	67.64-67.65	8677.93	113.469	103.582
CNVup1	63.44-63.46	9274.26	2105.23	2074.36
CNVup2	63.46-63.56	6293.79	83.6796	68.7332

French coverage = read depth of the sequenced sample from a French population (id = 233, 1A average coverage = 13.15); Belgium coverage = read depth of the sequenced sample from a Belgium population (id = 973, 1A average coverage = 9.55)

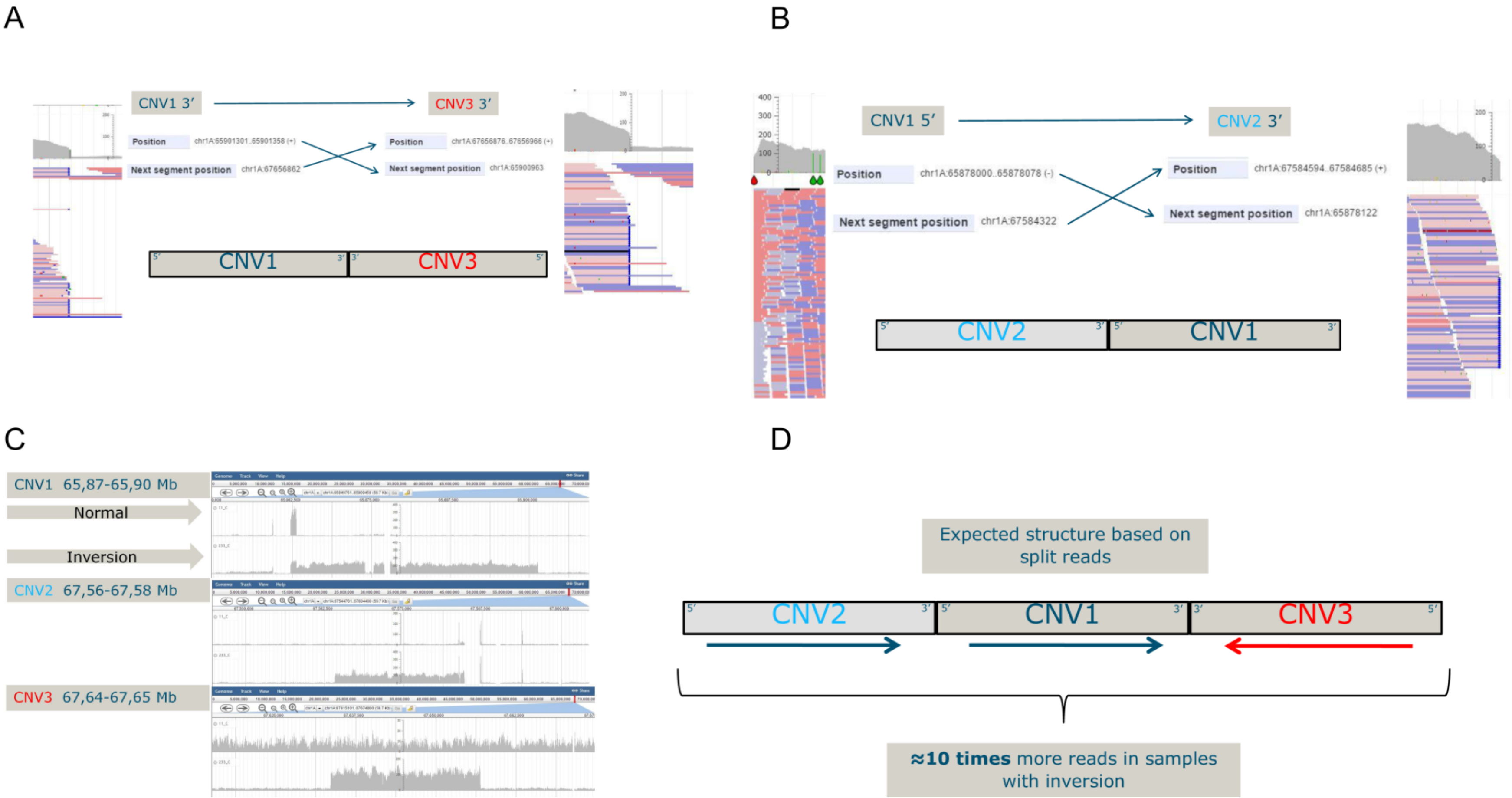

A
CNV1 3′
CNV3 3′
Position chr1A:65901301..65901358 (+)
Next segment position chr1A:67656862
Position chr1A:67656876..67656966 (+)
Next segment position chr1A:65900963
5′ CNV1 3′
3′ CNV3 5′
B
CNV1 5′
CNV2 3′
Position chr1A:65878000..65878078 (-)
Next segment position chr1A:67584322
Position chr1A:67584594..67584685 (+)
Next segment position chr1A:65878122
5′ CNV2 3′
5′ CNV1 3′
C
CNV1 65,87-65,90 Mb
Normal
Inversion
CNV2 67,56-67,58 Mb
CNV3 67,64-67,65 Mb
D
Expected structure based on split reads
5′ CNV2 3′
5′ CNV1 3′
3′ CNV3 5′
≈10 times more reads in samples with inversion

Figure 5.12: A) Split reads supporting the structural rearrangement between CNV1 (65.87-65.90) and CNV3 (67.64-67.65) **B)** Split reads supporting the structural rearrangement between CNV2 (67.56-67.58) and CNV1 (65.87-65.90).

Chapter 6

Selfishness can be deadly: a recessive lethal inversion is maintained by meiotic drive in great tits

Vinicius H. da Silva[1,2,3], Judith E. Risse[2,4], Kees van Oers[2,5], Martijn F.L. Derks[1], Veronika N. Laine[2,6], Mirte Bosse[1], Richard P.M.A Crooijmans[1], Martien A.M. Groenen[1] & Marcel E. Visser[1,2]

[1]Animal Breeding and Genomics, Wageningen University & Research, Wageningen, The Netherlands
[2]Department of Animal Ecology, Netherlands Institute of Ecology (NIOO-KNAW), Wageningen, The Netherlands
[3]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden
[4]Bioinformatics, Wageningen University & Research, Wageningen, The Netherlands
[5]Behavioural Ecology, Wageningen University & Research, Wageningen, The Netherlands
[6]Department of Molecular and Cellular Biology, Harvard University, Cambridge, USA

In preparation

Abstract

Recessive lethal variants can be maintained in large populations by genetic drift, balancing selection through a heterozygote advantage or segregation distortion. We recently reported a large (≈64 Mb) and widespread (≈5% in frequency) inversion on Chromosome 1A of the great tit (*Parus major*). Here, we show that this inversion is recessive lethal as the offspring of 13 wild carrier-by-carrier mating pairs is composed by 62.5% of heterokaryotypes and 37.5% non-carriers while no homorokaryotypes were found. Moreover, carrier-by-carrier pairs had 20% less eggs hatched in comparison with carrier-by-normal and normal-by-normal pairs. In pairs where the father is the carrier, we found twice more carrier offspring than expected by Mendelian law (≈67%, 69 from 103), suggesting that the inversion is a selfish arrangement when transmitted by a male. To maintain the inversion around its observed frequency of ≈2.5%, and taking the segregation distortion strength into account, the carriers should have a fitness disadvantage of ≈12.7%. In the current data set of 612 birds the fitness disadvantage for carriers (i.e. lower number of fledged offspring) is not significant and a larger data-set may be needed to demonstrate such an association. Therefore, the large recessive lethal inversion in the great tit has been maintained by segregation distortion but the molecular mechanism and the fitness disadvantage that is preventing it to have a higher frequency need further research.

6.1 Introduction

Inversions are intra-chromosomal genetic variants that result in a reversal gene order and play a crucial role in the evolutionary history of several species (reviewed in (Hoffmann & Rieseberg, 2008)). Inversions can provide a fitness advantage to the carriers, which may promote, but does not always lead to fixation (Jones et al., 2012; Kapun et al., 2016). The mechanisms underlying the maintenance of polymorphic inversions include overdominance, epistasis/coadaptation and associative overdominance (reviewed in (Faria et al., 2019)). In some cases, an inversion can become a recessive lethal variant by disrupting an essential gene (i.e. deleterious effect at the breakpoints) or by harboring a recessive allele (e.g. single nucleotide polymorphisms or copy number variations at essential genes). Inversions which are homozygous lethal tend to be purged out but they can be maintained when heterozygotes have a fitness advantage over the non-carriers. Lethal alleles under balancing selection have been identified in natural (Ekblom, 2016) and livestock populations (Derks et al., 2018), which usually reaches a frequency plateau after a sufficient number of generations.

Although fitness advantage can drive the maintenance of lethal variants under Mendel's law of equal segregation, some alleles can deviate from Mendel inheritance and show a different rate of transmission than alternative alleles (Sandler & Golic, 1985). Thus, these variants can exert advantage in the intragenomic conflict instead, leading to a fitness advantage to the carrier itself (Sandler & Novitski, 1957). Unequal allele segregation can have a profound evolutionary impact because 'selfish' variants can be maintained across generations even if they have a selective disadvantage to their carriers (review on the evolutionary impacts of meiotic drive in (Lindholm et al., 2016)). Segregation distortion, or meiotic drive, can involve biological processes that are strictly linked to females or males. Elements associated with chromosome structure as centromeres and telomeres can exploit female meiosis asymmetry of some species to promote its preferential inclusion in ova (Fishman & Kelly, 2015; Chmátal et al., 2014; Didion et al., 2015). Otherwise, the male-related meiotic drive is usually linked to their sperm dynamics. In males, drive elements can obtain a higher transmission rate by killing the sperm that lacks the meiotic variant (Wu et al., 1988; Merrill, 1999; Larracuente & Presgraves, 2012) or by improving the motility of carrier sperms (Sutter & Lindholm, 2016; Kim et al., 2017).

In birds, inversions are associated with traits or behaviours related to reproduction such as male morphology (Lamichhaney et al., 2016), improved sperm motility (Kim et al., 2017) and disassortative mating (Tuttle et al., 2016). In ruffs (*Philomachus pugnax*), a nested inversion is associated with different male morphs, i.e. independents (dominant), satellites (submissive) and feathers (mimicry female plumage), which may lead to three different reproductive strategies (Lamichhaney et al., 2016). The white-throated sparrow (*Zonotrichia albicollis*) displays disassortative mating among morphs, which in turn are defined by a large inversion encompassing more than 1,000 genes that is recessive lethal (Tuttle et al., 2016). In the zebra finch (*Taeniopygia guttata*) the majority of the genetic variation in sperm morphology is caused by an inversion polymorphism, which is located in the Z Chromosome (Kim et al., 2017). These results in zebra finch support that meiotic drive is maintaining this sex-linked inversion because heterozygous males have the fastest and most successful sperm.

Here we used great tits (*Parus major*) to investigate meiotic drive and fitness associated with a large and widespread putative inversion, which encompasses almost 1,000 genes and is located on the Chromosome 1A of these species (a detailed analysis on this inversion can be found in (da Silva et al., 2019)).

6.2 Material and methods

6.2.1 Sample description and inversion profiling

A total of 2,296 birds were previously genotyped (da Silva et al., 2019) at Edinburgh Genomics (Edinburgh, United Kingdom) on a custom made Affymetrix® great tit 650K SNP chip (Kim et al., 2018) and then classified for the inversion. In addition, 134 birds (55 chicks, 6 mothers and 73 fathers) were profiled by a PCR-RFLP diagnostic assay (PCR-RFLP profiles are described in detail elsewhere (da Silva et al., 2019)). In the end, a total of 229 females and 182 males belonging to 306 different mating pairs with fitness-related seasonal measurements recorded were used to investigate the inversion effects on fitness. These pairs are classified as 11 carrier-by-carrier, 146 carrier-by-normal (17 carrier (male)-by-normal and 129 carrier (female)-by-normal) and 149 normal-by-normal.

To investigate whether the inversion follows what is expected in a recessive lethal variant, we used 56 chicks from the 11 carrier-by-carrier mating pairs, which had 13 broods profiled for the inversion (two pairs had both a first and a replacement clutch in the same year). Moreover, to detect any sign of segregation distortion (i.e. meiotic drive) in the inversion, we analyzed 27 carrier-by-normal mating pairs (total of 30 broods), 12 carrier (male)-by-normal and 15 carrier (female)-by-normal, which had a total of 105 chicks profiled for the inversion by a PCR-RFLP diagnostic assay (da Silva et al., 2019).

6.2.2 Testing for extra-pair offspring

A substantial proportion of the offspring are not sired by the social male in great tits. Thus, extra-pair paternity rate may be important to clarify real father-offspring relationships to properly investigate the inversion inheritance. After DNA extraction PCR was performed using five microsatellite DNA loci: PmaTAGAn71, PmaGAn27, PmaTGAn33, PmaC25, and PmaD105, as described elsewhere (Saladin et al., 2003). Separation of the PCR fragments took place using an ABI 3130 Genetic Analyzer (Thermo Fisher Scientific, Waltham, MA). The capillary electrophorese results of the ABI were analyzed with the software GeneMapper 5.0 (Thermo Fisher Scientific, Waltham, MA) to determine the sizes of the amplification products. A chick was categorized as extra pair if three or more loci mismatched with the social father.

6.2.3 Inversion association with fitness components

Seasonal measurements such as egg-laying dates, clutch size, number of hatched eggs and number of fledged chicks have been recorded in our long-term study great tit populations on the 'Veluwe' area close to Arnhem (52°02' N, 5°50' E, the Netherlands) since the 1955. In this study area nest boxes are widely available so almost the entire population breed in boxes and can be monitored.

The mean of the seasonal measurements (i.e. fitness components) differ between years (as they are strongly affected by spring temperature and other environmental variables, (Gienapp et al., 2005)) and among areas. We therefore fitted the following model to all fitness components for the entire population (i.e. birds with and without genotypes):

$$y_{i,j} = \mu + \beta_j + \beta_a + pe_i + \varepsilon$$

with $y_{i,j}$ being a fitness component i in year j, μ the overall intercept, β_j and β_a the fixed effects for year (as factor) and area (Buunderkamp-NL, Westerheide-NL, Roekelse Bos-NL, Hoge Veluwe-NL or Oosterhout-NL), respectively and pe_i the random permanent environmental effect of mother i. We then used the year and area estimates from this model to correct the fitness components of the genotyped individuals for year and area effects. We performed this two-step approach, instead of fitting year and area directly in the linear mixed models that are described below, because not all individuals in all years were genotyped (and for some year/area combinations only very few individuals), which could have led to inaccuracy and/or bias in the estimates for year-area combinations by using only the few mating pairs that have full family and inversion genotype information available.

We used a linear mixed model to detect the association strength between our four fitness-related measures and the inversion haplotype of the respective mating pairs. As there were no homozygous inversions detected (da Silva et al., 2019), we only have birds which are heterozygous for the inversion (carriers) or those that are non-carriers in this analysis. Therefore, we compared (i) egg-laying date, (ii) clutch size, (iii) number of hatched eggs and (iv) number of fledged chicks among all three possible mating pairs combinations for the inversion: (i) carrier-by-carrier, (ii) carrier-by-normal and (iii) normal-by-normal.

$$y'_{i,j} = \mu + inv_i + age_i + mother_i + father_i + e_{i,j}$$

$y'_{i,j}$ being the fitness component corrected for year and area effects in the mating pair i in year j, μ the overall intercept, inv_i the mating pair combination (i.e. carrier-by-carrier, carrier-by-normal or normal-by-normal), age_i the age of the mother (the age of the father is mostly unknown), $mother_i$ is the random effect of the mother and $father_i$ is the random effect of the father from each respective brood.

The association between fitness components and mating pairs can expose how the combination of the inversion genotype in the parents affect these seasonal measurements. However, the individual association with fitness might be also useful to better understand what is maintaining this inversion and can be more easily plugged into subsequent simulations. Thus, we associated the inversion genotype of each individual with the fitness component that would best reflect fitness among our measurements (i.e. number of fledged chicks).

$$y'_{i,j} = \mu + inv_i + age_i + sex_i + pe_i + e_{i,j}$$

$y'_{i,j}$ being the number of fledged birds corrected for year and area effects in the bird i in year j, μ the overall intercept, inv_i the bird genotype (i.e. carrier or non-carrier), age_i the age bird if available, sex_i the sex of the bird and pe_i random effect of the individual.

The models were fitted with the `lmer` function from lme4 R package (version 1.1-21, Bates et al. (2015)). The models were fitted using REML and the p-values derived using the Wald chisquare test with `Anova` function implemented in the car R package (Fox & Weisberg, 2011). A post-hoc test to explore differences between means in different mating pairs while controlling the family error rate was carried with the Tukey method (Tukey, 1949), which is implemented in the `emmeans` function from the R package emmeans (version 1.3.3, (Lenth, 2019)).

6.2.4 Simulations on drift-selection and statistical power of the fitness association

To investigate fitness advantage/disadvantage of heterozygotes that would be needed to explain the maintenance of a variant with the same singularities of the inversion (i.e. recessive lethal and selfish and with frequency of 2.5%), we empirically simulated drift-selection scenarios. In all tested scenarios (i) the fitness of the homozygote was set to zero and (ii) the gamete proportions transmitted to each subsequent generation were intentionally weighted to account for the observed segregation distortion in males (the inversion is inherited in 70% of the offspring instead 50% in carrier males whereas maintained under Mendelian law for carrier females). We used the effective population size of 5.7x10^5 individuals, as previously estimated from pairwise sequential Markovian coalescent analysis (Laine et al., 2016). We modified the source code of Shiny/R package driftR (`https://cjbattey.shinyapps.io/driftR/`) to perform the drift-selection simulation as described above.

Next, we estimated the number of birds that would be required to find the fitness difference as predicted by the drift-selection simulation that is described above. The observed data-set (i.e. observed number of fledged birds) and the above described model was used to simulate an association analysis with `simr` R package (version 1.0.5, Green & MacLeod (2016)), which can predict the sample size required to significantly expose the expected effect of this inversion on individual fitness. In the model used for the simulation, the fixed effect of the inversion genotype (i.e. carrier) was modified, as suggested in `powerSim` function from `simr` R package (version 1.0.5, Green & MacLeod (2016)), to reflect the expected fitness advantage/disadvantage of heterozygotes that were obtained by drift-selection simulation.

6.3 Results

6.3.1 Inheritance patterns of a recessive lethal variant

In inversion carrier-by-carrier pairs we expect 66.65% of the chicks to be carriers and 33.35% to be non-carriers (i.e. here defined as ‘normal’) assuming an inheritance model where the inversion is a fatal recessive allele. The 11 carrier-by-carrier pairs, which produced 13 broods, had 62.5% (35 chicks) of the offspring as carrier and 37.5% (21 chicks) as normal, in agreement with homozygous lethality.

6.3.2 Inheritance patterns displaying segregation distortion

The carrier-by-normal pairs are expected to have half of the offspring as carriers and the other half as normal if we assume that the inversion follows the Mendelian inheritance. In carrier-by-normal pairs, the inversion inheritance clearly deviates from what is expected for a genetic variant following the Mendelian law (i.e. 50% carriers and 50% normal chicks). We found that the offspring in these pairs follows

Mendelian inheritance only when the mother is the carrier. In this case, from 102 chicks, 50 (49%) are normal and 52 (51%) are carriers. By contrast, when the father is the carrier, from 103 chicks, 34 (33%) are normal and 69 (67%) are carriers. Note however that not all offspring will be sired by the social male. We therefore determined which of the offspring of the carrier males were extra-pair offspring and found that 14 out of 34 were extra-pair, and these were all normal offspring. Thus, the percentage of carrier chicks of carrier males is 77.5%.

6.3.3 Association of the inversion with fitness

We evaluated the effect of the combination of the inversion genotype in mating pairs on two traits; (i) egg-laying dates and (ii) clutch size; and two fitness measurements; (i) number of hatched eggs and (ii) number of fledged chicks. Only the number of hatched eggs was significantly lower in carrier-by-carrier (5.9 eggs in average) in comparison with carrier-by-normal (7.38 eggs) and normal-by-normal (7.37 eggs) pairs (Tukey multiple comparison p-values 0.0026 and 0.001, respectively, Figure 6.1). In fact, the ratio between clutch size and the number of hatched eggs is clearly different among carrier-by-carriers and the other two pair classes, which further supports the homozygous lethality of the inversion.

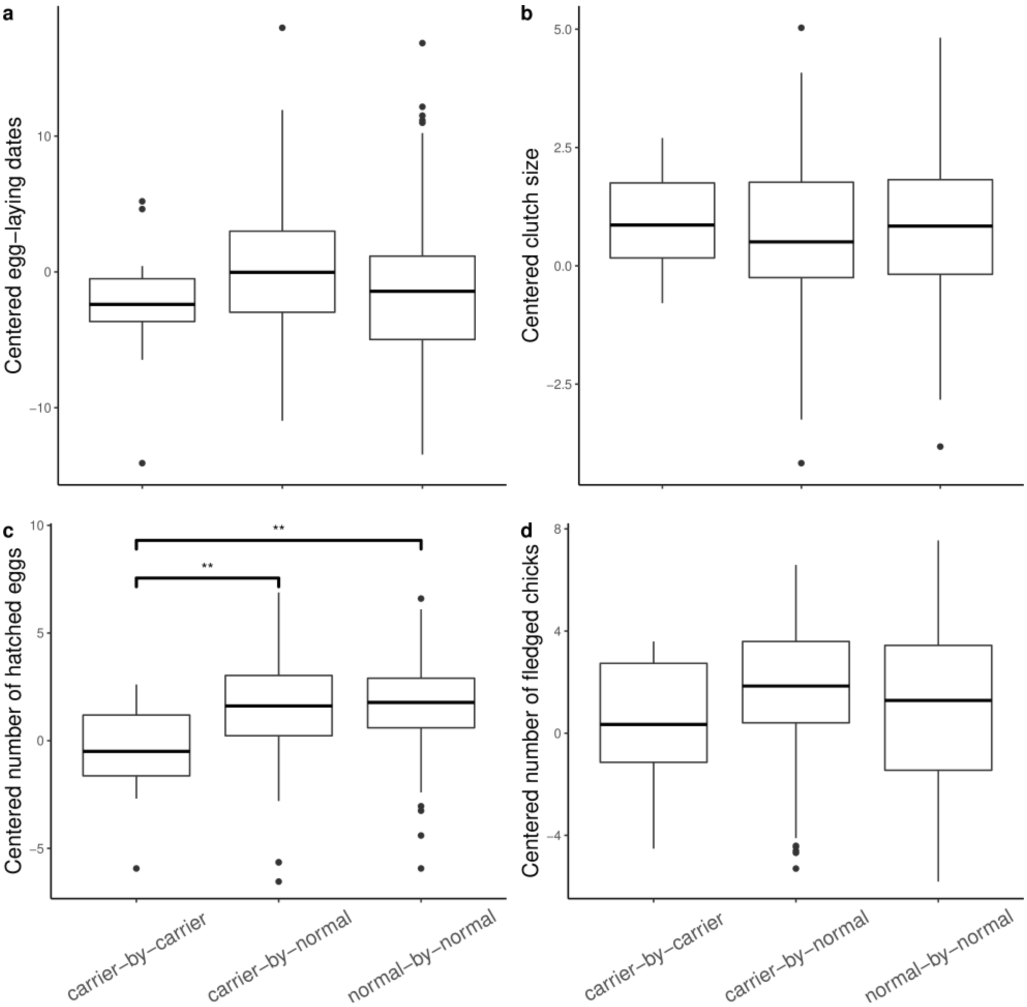


Figure 6.1: Association of fitness-related measurements with mating pair classes. A lower number of hatched eggs was observed in carrier-by-carrier in comparison with carrier-by-normal and normal-by-normal mating pairs (p-values 0.0026 and 0.001, respectively). All 'carrier' birds harbor a large inversion on the Chromosome 1A, for which they are heterozygous. All 'normal' birds are non-carriers. The inversion is fatal in homozygous condition. The number of hatched eggs is significantly higher in mating pairs with at least one non-carrier (i.e. normal-by-normal and carrier-by-normal) compared to carrier-by-carrier. None of the other fitness-related measurements significantly differ between mating pair classes.

Although we found no direct association between mating pair classes and the majority of the fitness-related measurements, fitness advantage/disadvantage may be expressed at individual level. Thus, we additionally associated individual genotypes (i.e. carrier or normal) with the measurement that would be best reflect fitness, i.e. their number of fledged chicks. In accordance with the results using mating pair classes, being carrier is also not significantly associated with the number of fledged chicks (p-value $= 0.55$).

To quantify the expected fitness advantage/disadvantage caused by the inversion on the carriers, we simulated a drift-selection scenario in which an allele follows all the inversion singularities (i.e. 2.5% in frequency plateau, recessive lethality and selfishness). The inversion frequency in the population should reach a plateau around the observed frequency of 2.5% in approximately 400 generations after its formation, when the relative fitness disadvantage of the carriers is assumed to be approximately 12.7% (Figure 6.2a). In a drift-selection scenario that the carriers have no fitness disadvantage but the inversion is equally selfish, the allele frequency should be much higher than what is observed in our population (≈14.5% in frequency, Figure 6.2b).

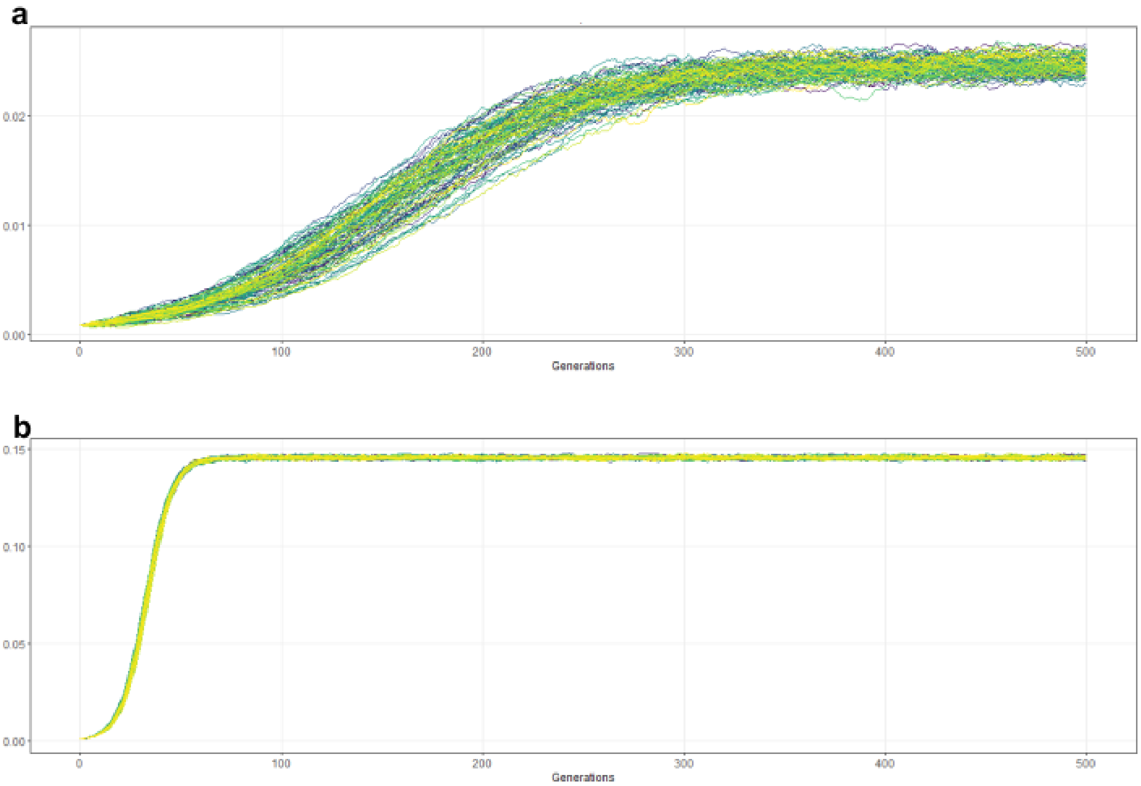


Figure 6.2: Drift and selection of the inversion in great tit species. The y-axis is the frequency of the inversion allele under a drift-selection scenario that considers its selfish nature (i.e. 70/30 inheritance ratio) and recessive lethality (homozygotes have relative fitness, i.e. w, set to 0). A-) Assuming the w disadvantage of the carriers to be ≈-12.7%. B-) Assuming w carriers and non-carriers to be equal, which shows the expected increase in the inversion frequency due to drift alone.

As we have a limited number of birds concomitantly profiled for (i) the inversion and (ii) number of fledged offspring, the expected relative fitness disadvantage of 12.7% might be undetectable with our current statistical power. Therefore, we estimated the sample size that would be required to reach a significant association. Using results and settings from the drift-selection simulation, we used the linear mixed model used for the individual fitness association under different simulated sample sizes. In order to reach a reasonable statistical power, i.e. around 80%, it would be necessary to obtain more birds than what is available in our current data-set. The statistical power that is expected by the sample size of our current data-set (612 birds) is in median 69.50% (95% confidence interval of 1,000 simulations ranging from 66.54 to 72.32% in power). Extending inv_i by one level in the model, the power should be in median 100% (95% confidence interval of 100 simulations ranging from 99.63 to 100% in power). Thus, we extended the number of samples within each level of inv_i to accommodate from 300 (roughly the data-set available in this study) to 600 observations. By using a data-set of $\approx$730 birds would be possible to achieve a statistical power around 80%.

6.4 Discussion

The large inversion investigated in this study is located on the Chromosome 1A of the great tit genome and encompasses almost 1,000 genes. It is widespread over different European populations and has an observed carrier frequency of approximately 5% as well as high structural complexity, which evidences recombination in the center that supports that the inversion is more than 10^5 generations old (da Silva et al., 2019). Thus, this observed carrier frequency is likely stable, given that this inversion is not young enough to be still increasing towards its frequency plateau (i.e. it is unlikely that the inversion is younger than 400 generations, Figure 6.2a). A stable frequency far from fixation, after a number of generations, is expected for a deleterious variant under balancing selection Derks et al. (2018), which therefore may be the case for this inversion in the great tit genome.

In a model where the inversion can occur in homozygous state, a cross between two carriers should generate 25% of homozygous carriers. However, this was not observed in our data given that we obtained approximately 65% of heterozygous carriers and 35% of non-carriers from carrier-by-carrier pairs. Moreover, the number of hatched eggs is approximately 20% lower in carrier-by-carrier in comparison with other mating pairs, which is close to what is expected for a recessive lethal variant where at least 25% of the eggs in a clutch do not have a viable embryo. Thus, our results support that the lack of homozygotes in our population (da Silva et al., 2019) is because the inversion is actually lethal in homozygous state, precluding homozygote offspring in carrier-by-carriers. A comparable inversion in white-throated sparrows, which is also very large and comprises around 1,000 genes, may rarely happen in homozygous state Tuttle et al. (2016). However, in great tits it is still unclear if homozygous birds exist in nature in such extreme low rates. As the inversion encompasses a large number of genes, it can be challenging to find a candidate gene to explain the homozygous lethality of this inversion. However, based on sequencing data, there are three genes in the inversion downstream breakpoint of the inversion that could be potentially broken (da Silva et al., 2019) (i) Bisphosphoglycerate Mutase (*BPGM*), (ii) Caldesmon 1 *CALD1* and (iii) Phosphatidylinositol-4-Phosphate 3-Kinase Catalytic Subunit Type 2 Gamma (*PIK3C2G*). *BPGM* underlies oxygen sensing in blood cells Petousi et al. (2014) and the levels of oxygen have an important role on embryonic differentiation (Simon & Keith, 2008). Functional domains of the Caldesmon protein are necessary for the development of the early embryo as homozygous recessive mice do not develop (Deng et al., 2013). *PIK3C2G* gene could be also a candidate gene to explain the inversion lethality given that knockouts of other genes in the PI3K family lead to embryonic lethality in mouse (Bi et al., 1999). Thus, future studies focusing on these potential genes could clarify the actual molecular mechanism behind the homozygous lethality of the inversion.

It may be challenging to narrow down to the gene, or genes, which actually underlie the lethality of the inversion. However, the recessive lethality of the inversion is clearly reflected by the significant decrease in the number of hatched eggs in carrier-by-carriers. Thus, to surpass the disadvantage of being lethal, the inversion should confer a fitness advantage to the carriers or otherwise break Mendel's law. There are known examples of haplotypes harboring inversions in other species, such as the *t*-haplotype in mouse (Kelemen & Vicoso, 2018), which shows meiotic drive and therefore breaks Mendel's law (i.e. the transmission of the haplotype containing inversions is favoured and is therefore designated as a 'selfish gene'). Thus, it was important to determine if the inversion shows any sign of meiotic drive reflected on the offspring ratios. To answer this question, we analyzed 27 carrier-by-normal mating pairs, which have in total 105 birds profiled for the inversion. In carrier-by-normal pairs that have female as carrier, the offspring proportions support a normal Mendelian inheritance. By contrast, pairs in which the male is the carrier, the number of carrier offspring is approximately twice higher than expected by Mendelian inheritance, suggesting that the inversion can behave as a 'selfish' arrangement. Therefore, it is plausible to assume a sperm-related meiotic drive mechanism underlying the segregation distortion of this inversion.

There are a growing number of genetic variants that selfishly interfere on gamete production to increase their own rate of transmission (Lindholm et al., 2016), which can rely on a female- or male-specific biological mechanism. In males, segregation distortion can be achieved by a molecular mechanism that kills sperms lacking the selfish variant (Bravo Núñez et al., 2018). A truncated version of the *RanGAP* gene protein in drosophila, which is produced by gametes harboring a selfish gene, kills developing wild-type spermatids through an interaction with a wild-type specific satellite (Larracuente & Presgraves, 2012). Interestingly, the *RANGAP1* gene in great tits is located at the center of the Chromosome 1A, within the inversion genomic interval that is recombinant in 10% of the carriers da Silva et al. (2019). As the rate of extra-paternity pairs (EPP) in carrier (male)-by-normal pairs seems to be within the range that was previously reported for great tits (i.e. around 14%, (Blakey, 2008)), there is no evidence of sperm competition among the social carrier males and real sires (e.g. high extra-paternity rate could suggest lower semen quality in carriers). The actual mechanism of segregation distortion of this inversion still needs to be clarified, but our current results indicate that meiotic drive plays a central role in the maintenance of this recessive lethal inversion. However, the observed meiotic drive ratios may lead the inversion to a higher frequency than what is observed in our great tit population.

Drift-selection simulation (Wright, 1931) is a useful tool to understand the evolutionary dynamics of a genetic variant over time. By defining the fitness for each genotype as well the effective population size, drift should be taken into account to display the likely change in frequency over time. Specifically for the inversion, homozygotes should have fitness equal to zero, as they are unable to survive, and weighed gamete inheritance in each generation should be considered to account for the segregation distortion. We found that the inversion is expected to reach a stable frequency around 14.5% if no fitness disadvantage exists in heterozygotes. Thus, as the inversion is old enough to have reached its frequency plateau (da Silva et al., 2019), a fitness disadvantage should be present to explain the inversion frequency that is around 2.5% (da Silva et al., 2019).

Given that the 1A inversion may confer a fitness disadvantage to its carriers, a fitness component is expected to be associated with this variant. However, given that 25% of the carrier-by-carrier offspring is expected to be non-viable, the fitness associated to each of the possible mating pair combinations may be more informative than individual genotypes to understand the fitness advantage of this inversion. Therefore, we separately compared four fitness-related measurements (i.e. egg-laying dates, clutch size, number of hatched eggs and number of fledged chicks) among each of the three possible mating pairs (i.e. carrier-by-carrier, carrier-by-normal and normal-by-normal). However, excepting the number of hatched eggs, which exposes the inversion recessive lethality, all other fitness components were not significantly associated with mating pair inversion genotypes. In addition to mating pairs, we checked if our best proxy for fitness, the number of fledged chicks, was associated with being a carrier. The results were similarly negative as fledgling is not associated with the inversion, even in our model that considers the sex of the carrier. However, as our drift-selection scenario supports fitness disadvantage in heterozygotes, is important to understand the statistical power of our analyses to see if more birds would be required to find such an association.

To understand the effect of the inversion on heterozygotes, we used a linear mixed model to associate the carrier/non-carrier with the number of fledged chicks. However, the fixed effect that was observed for the inversion, on the number of fledged chicks, is more than four times lower than predicted by drift-selection simulation and clearly not significant. Assuming that the fixed effect predicted by the drift-selection simulation may be reflected on the number of fledged chicks, a reasonable statistical power may be achieved with a larger data-set. However, as drift-selection simulation considered a scenario where both sexes are present in comparable proportions, this difference between observed and expected fixed effect could be due to the limited sample size or to the unequal rate between male/female carriers in our association data-set. Otherwise, the expected fitness difference between carriers and non-carriers could be due to a completely different fitness component, which is not available in our population or was not captured by our experimental design. For example, a hypothetical higher probability to mate in a given breeding season in non-carriers could impose them a considerable fitness disadvantage, but the number of fledged birds would still not differ between carriers and non-carriers.

6.5 Conclusion

It is unclear if this inversion is associated with any phenotype related to mating behaviour, such as inversions linked with different morphs in ruffs and white-throated sparrows (Tuttle et al., 2016; Lamichhaney et al., 2016). Therefore, association studies other than number of fledged chicks as well as deeper understanding about the inversion sequences related to its respective meiotic drive system may assist in the discovery of the actual biological mechanism maintaining this large and complex inversion.

Chapter 7

General discussion

7.1 Introduction

Great tit (*Parus major*) is a songbird that has been widely used as a model species in ecology and evolution. Structural variants (SVs) have been increasingly explored in wild populations to better understand the evolution and ecology of different species. In this thesis, I have performed a detailed study of SVs in the great tit genome. However, the high technical and biological variation present in SVs posed challenges for their study. SVs can be complex due to the combination of different structural rearrangements (e.g. changes in copy number, inversions and translocations). However, changes in copy number (i.e. copy number variations - CNVs) are relatively easier to detect and may highlight more complex regions in the genome. Although easier to study, CNVs are also prone to technical variation, which can lead to a substantial number of false positive and negative CNV calls. Thus, I will discuss here how a detailed analysis of the genomic architecture underlying CNVs was used to deal with part of this technical variation. Moreover, I will discuss here how the ratio between expected and obtained CNV inheritance was also used to quantify and better classify the technical variation present in this CNV study. The understanding of technical variation in a CNV study is essential to perform subsequent analyses such as CNV-based genome-wide association studies (GWAS). However, the biological variation that is usually present in CNVs can be also challenging to deal with and interfere with the GWAS results. Overlapping CNVs can have different breakpoints and copy number states, which may complicate their classification into loci. Thus, an oversimplification of CNV loci may lead to wrong association results. I will discuss here how the CNV-based GWAS method proposed in this thesis was used to tackle part of this inherent biological variation in CNVs. Therefore, I will discuss here how this method was used to better understand the effect of CNVs on the seasonal timing in great tits. Moreover, I discuss here how the same CNV-based GWAS method can be used to study other fitness components and phenotypes in a species-independent manner. Although more accessible, CNVs represent only part of all SVs in a genome. Therefore, I discuss the identification and characterization of a large inversion in the great tit genome and how CNVs may be underlying it. Finally, I explore the recessive lethality and the selfish nature of this remarkable large inversion in the great tit genome.

7.2 Genomic architecture and inheritance reflects the confidence of CNV detection

The majority of the studies aiming to identify CNVs disregard the genomic architecture that is expected to be associated with these variants. Certain features in the genome are known to underlie CNV formation (Carvalho & Lupski, 2016) and are therefore expected to be enriched at CNV regions (CNVRs). Genomic features such as CpG islands, segmental duplications (LDs) and AT-rich segments can be generally defined as local genomic architecture, which underlies a region-specific replication efficiency. Replication-based mechanisms (RBMs) are less stable when replicating repetitive regions of the genome, which may promote the formation of new structural variants in these genomic regions (Carvalho & Lupski, 2016). For example, a non-allelic homologous recombination (NAHR) can occur between two intervals of the genome that have high sequence homology but are not alleles. Low copy repeats (LCRs) are highly homologous sequence elements that more often endure NAHR events, which in turn underlie the higher structural variability present within and in the vicinity LCRs. LCRs with lower similarity tend to be older than highly similar LCRs as each copy has longer been following an independent evolutionary path (Chaudhry et al., 2018). The higher incidence of CNVs at more recent LCRs is known in humans and was important to further understand the role of structural variants in the human-chimpanzee speciation (Perry, 2008). In great tits, LCRs that are enriched at or in the vicinity of CNVs show at least 98% identity (**Chapter 2**, Figure 2.4), confirming the expectation that recombination mechanisms, such as NAHR, may become less frequent as identity between sequences decreases. Therefore, a robust permutation overlap analysis between CNVs and the genomic features expected to be underlying their formation, i.e. such as LCRs, can be used to improve the knowledge on the molecular evolution of species as well as to assist in the assessment of false negative-positive assessment in CNVs.

By knowing the mechanisms underlying CNV formation, it is possible to tag genomic intervals that have a higher chance of harboring structural variations. If the CNVs identified in a study overlap such genomic features associated with CNV formation more than expected by chance, the CNV data-set under study may have an acceptable false-positive rate (i.e. *CNV calls* are not randomly distributed across the genome). In this thesis I have identified and compiled a collection of features associated with the formation of structural variants to understand their genomic colocalization with CNVs in the great tit genome. Apart from (i) LCRs, genomic features such as (ii) Interspersed repeats and low complexity DNA sequences, (iii) CpG sites, (iv) Transcription start sites (TSSs) and (v) AT-rich regions were also analyzed by overlap permutation to understand if CNVs usually colocate with these features. It was interesting to note that the overlap of CNVs with all the analyzed genomic features quite deviates from what is expected by chance.

The frequency of a CNV in a population is an important factor to consider before looking into their overlap with certain genomic architecture. For example, it is known that non-recurrent CNVs (i.e. rare arrangements) are enriched at genomic regions that are prone to break (Carvalho & Lupski, 2016), which is reflected by the inverse relationship between CNV frequency and overlap count with AT-rich sequences in great tits (**Chapter 2** of this thesis). Non-recurrent, or *de novo*, mutations have been shown to be functionally relevant (Veltman & Brunner, 2012), but CNVs usually follow Mendelian inheritance (Locke et al., 2006). Thus, if family information is available, analyses on the CNV inheritance can also confer more reliability to a CNV study by highlighting CNVs which are following the Mendelian law. We found a significant correlation between CNV inheritance ratio and their number of underlying SNP probes. As most of CNVs follow Mendelian inheritance, the higher proportion of inherited CNVs in calls supported by a higher number of probes show that a lower false negative-positive ratio may be achieved in regions with a higher SNP probe density. However, the use of independent platforms may be important to overcome this, and other, platform related bias and disentangle technical and biological variation (Li & Olivier, 2013).

Due to high variability and sometimes low resolution of the different methods and platforms that are able to detect CNVs (Li & Olivier, 2013), the use of more than one platform is desirable to better understand the false negative-positive ratio in a CNV data-set. This is sometimes denominated as ‘validation’ and can be accomplished by quantitative PCR (qPCR) (D’haene et al., 2010) or genome sequencing (Xie & Tammi, 2009a) when using a SNP array as the primary platform. The CNV data-set identified in our great tit population, with a species-specific high density SNP array (Kim et al., 2018), had a group of CNVs validated by qPCR that obtained a high validation rate (>90%, **Chapter 2**). However, as the inheritance patterns show that CNVs supported by a lower number of SNP probes tend to have an unexpected lower inheritance ratio, it is likely that the number of false negative CNVs is much higher than the false positives, at least in short CNVs. Although intraspecific genomic architecture is useful to define expected CNV distribution in a genome, existent interspecific genomic similarity (i.e. ‘synteny’ between species) can also tag genomic regions that are prone to harbor CNVs.

7.3 Interspecific evolutionary breakpoints are enriched in CNVs

Speciation, by which populations evolve by genetic selection into distinct species, is the evolutionary process responsible for the remarkable biodiversity on our planet. The genomic variation among species can then be explored by comparative genomics (Hardison, 2003), which shed a light on the association between phenotypic evolution (i.e. traits differing between species) and molecular evolution of the genome. The comparison among the genomes of different species can reveal genomic intervals, large gene-containing segments, that can be species-specific as well as intervals from a common ancestor. The genomic intervals that are conserved between species reflect the 'synteny' between their genomes (Sankoff, 2009). By contrast, genomic intervals flanking these syntenic regions harbor evolutionary breakpoints, which expose changes in the genome likely caused by speciation (Ruiz-Herrera et al., 2006). Repetitive elements are common at these evolutionary breakpoints (Longo et al., 2009), supporting that CNVs play a central role in speciation. Thus, the expected and observed overlap between CNVs and evolutionary footprints can be used to check the reliability of CNVs and their colocalization with evolutionary breakpoints in multiple pairwise comparisons between species.

The evolutionary breakpoints between great tit and chicken as well as zebra finch both overlap with CNVRs more than expected by chance, fitting the expected enrichment at these regions. In fact, homologous synteny blocks and evolutionary breakpoint regions reflect different evolutionary histories by harboring remarkably distinct types of genetic variation and gene profile (Larkin et al., 2009). Syntenic regions are enriched with conserved genes related to the development of the central nervous and other organ systems in mammals (Larkin et al., 2009). By contrast, evolutionary breakpoints may act as a major structural variability reservoir that underlies adaptive phenotypes (Larkin et al., 2009). Interestingly, evolutionary breakpoints in great tit, and consequently CNVs, are enriched with genes related to neuronal and cardiac processes. Therefore, phenotypic differences in the nervous system as well as in certain organs, such as the heart, may play a central role specifically in the bird speciation. Albeit selection drives speciation, the biodiversity within the same species is also propelled by selection. Thus, intraspecific genomic variation is also relevant to clarify the evolutionary history of a species. Thus, intraspecific genomic variation associated with certain traits cannot be detected by comparative genomics. Otherwise, genome-wide association studies (GWAS) are able to detect genetic variants that may underlie differences among individuals in a population (Visscher et al., 2017). Therefore, the study of phenotypes and fitness components, e.g. the egg-laying date in birds, might be able to reveal how changes in copy number underlie intraspecific biodiversity in great tit.

7.4 Methods in CNV-based genome-wide association studies

CNV effects on phenotypes have been increasingly studied but open-source software to perform association analyses with CNVs are rare and mainly focused on case-control associations (Kim et al., 2012; Barnes & Plagnol, 2017; Larsen et al., 2018). Although most of the software available for CNV association are case-control based, there are a few options to associate quantitative traits with CNVs. A software implemented in Java that allows the analysis of quantitative phenotypes is CONAN (Forer et al., 2010), but the software focuses on the human genome and is only available upon request. R is a language and environment for statistical computing and graphics (R Core Team, 2019), which has been used to orchestrate high-throughput genomic analysis in large part by packages available at the Bioconductor repository (Huber et al., 2015). Thus, using Bioconductor packages and architecture as a foundation to construct a new R package, for a high-throughput genomic analysis, can improve the integration among currently available and future pipelines and their performance. The CNVasso (Subirana et al., 2011) R package allows quantitative phenotypes and includes good model flexibility. However, CNVasso currently does not discuss exiting methods to define CNV loci and does not make use of the Bioconductor architecture to deal with *CNV calls*, like is done by e.g. GenomicRanges (Lawrence et al., 2013) and RaggedExperiment (Morgan & Ramos, 2019).

In **Chapter 4** we further developed an existing CNV-based association strategy (da Silva et al., 2016; Geistlinger et al., 2018), which was used to perform the study presented in **Chapter 3**, into a R/Bioconductor package (**Chapter 4**) to allow the reproducibility of the observed results as well as provide a new freely available tool to the scientific community. The package was named after CNVRanger and provides a wide set of functions to deal with (i) concatenation of CNV loci and their association with (ii) phenotypes and (iii) gene expression. In addition, the CNVRanger package allows genome-wide association of raw intensity signals (i.e. Log R Ratios) with quantitative phenotypes, instead *CNV calls* directly, which can be used along the results from *CNV calls* to improve the reliability of the results in CNV data-sets containing a high number of false negatives (e.g. such as the great tit data-set explored in **Chapter 2**). Moreover, this first core version of the CNVRanger package described in **Chapter 4** lays the foundation to better translate well-established analyses in SNPs to CNVs in the future. For example, future versions of the package may allow linear mixed models in the association with phenotypes (already available in a development branch in github and used for the CNV-GWAS applied in **Chapter 3**) as well as relevant analyses in evolutionary studies such as Vs_t (i.e. to compare different populations, (Redon et al., 2006a)). Therefore, in a nutshell, CNVRanger aims to continuously integrate and standardize populational analysis of CNVs into the Bioconductor environment.

7.5 Association of CNVs with phenotypes

CNVs underlie a large proportion of the genetic variability in humans and different livestock and wild species (Zarrei et al., 2015; Upadhyay et al., 2017; Prunier et al., 2017). In fact, the percentage of the genome that is encompassed by CNVs is usually higher than SNPs (Shlien & Malkin, 2009). Thus, phenotypic variability coming from CNVs allows natural or artificial selection towards more adapted or intended traits. As it has been shown that CNVs can confer adaptability in rapidly changing environments (Simam et al., 2018; Chain et al., 2014; Prunier et al., 2017), genetic variation in seasonal timing of reproduction, which has been shifting under global warming (Kentie et al., 2018), may be also associated with CNVs. The association between CNVs and seasonal timing may assist the understanding of (i) how climate change could shape genomic diversity and (ii) possible genetic variants associated with phenotypic plasticity in timing.

In birds, seasonal timing of reproduction is recorded as egg-laying dates. Therefore, in **Chapter 3**, I have explored the association of CNVs with egg-laying dates in two different natural populations of great tits from the Netherlands (NL) and the United-Kingdom (UK) to understand how changes in copy number might affect breeding timing. In accordance with the expectation for a highly polygenic trait such as egg-laying dates, there was no strong association between a specific CNV and egg-laying dates in great tits. A similar result was found by an environment-dependent SNP-based GWAS in the same population from the Netherlands (Gienapp et al., 2017), in which the variation in egg-laying dates could not be explained by specific SNPs. Although both approaches (i.e. SNP- and CNV-based GWAS) support that timing is largely polygenic, the top associated regions are not coincidental. In fact, the linkage-disequilibrium (LD) between CNVs and SNPs in the great tit genome is low (**Chapter 3**), suggesting that each polymorphism type can underlie distinct phenotypic variability. Albeit no strong association between seasonal timing and genetic variants is known in great tit, **Chapter 3** describes few CNVs displaying a suggestive association with egg-laying dates are associated with circadian clock, reproductive success and mammalian pregnancy (**Chapter 3**). Thus, the colocalization of suggestive CNVs and interesting genes reveal regions to be further explored in the study of the genetic basis of seasonal timing in great tits. Moreover, CNVs represent only part of all structural variation present in a genome, thus other rearrangements as inversions also deserve further research.

7.6 Beyond CNVs: the inherent complexity of inversions

CNVs are widely explored because they can be more easily inferred in comparison with other structural variants in the genome. However, more complex structural rearrangements, such as inversions, have been increasingly associated with fitness components and speciation events (Hoffmann & Rieseberg, 2008; Knief et al., 2016). Inversions can be challenging to detect because contrary to a CNV, there is no change in signal intensity when a genomic interval is in a reverse orientation. Thus, the methods to enable the detection of inversions cannot be based on signal intensities but instead make use of the fact that inversions will follow a different evolutionary path in comparison with their collinear homologous regions (Faria et al., 2019). This is expected because the recombination between an inversion and its respective collinear arrangement is severely impaired, which, after enough generations, may lead to distinct allele frequencies at several SNPs encompassed by an inversion (Kirkpatrick, 2010). As inversions have a different allele profile, analysis such as principal component analysis (PCA) may assist in the identification of inversions (Kirkpatrick, 2010). In **Chapter 5**, we explored a large inversion on Chromosome 1A of great tits using PCA, SNP heterozygosity and LD patterns. As expected for an arrangement that is unable to perform recombination with its collinear homologous locus, the PCA, the heterozygosity and LD metrics clearly distinguished carriers and normal birds (**Chapter 5**, Figure 5.1).

Long-term suppression of recombination may lead to gene loss as demonstrated in the degenerated sexual Chromosome Y (Skaletsky et al., 2003), or in the case of birds the W Chromosome. However, young inversions tend to follow a process referred to as expansion degeneration, in which gene gain precedes gene loss (Stolle et al., 2018). The large and widespread inversion on Chromosome 1A, which encompasses almost 1,000 genes and is described in detail in the **Chapter 5**, is in agreement with the expansion degeneration hypothesis as it harbors a higher number of copies in at least two different intervals that are close to the downstream inversion breakpoint. Moreover, one of these CNV regions ('CNVR 2802', which was detected in the genome-wide CNV detection performed in **Chapter 2**) can reasonably tag the inversion as more than 95% of the carriers hold copy gains in that region. However, although relatively young when in comparison with a degenerated sexual chromosome, the inversion should be at least 10^5 generations old due to the evidence of a rare recombination event in the center. This recombination event is assumed to be responsible for the alternative inversion haplogroups in the inversion center as described in **Chapter 5**, which accounts for approximately 10% of the carriers identified in the great tit population analyzed. Even though the recombination between inversions and the collinear arrangement is rare, it is known to happen more frequently far from the breakpoints. However, the mechanisms underlying such a recombination event are poorly known and further research is needed.

In Drosophila, a cosmopolitan inversion shows gene exchange in the center (Hasson & Eanes, 1996) and its patterns of diversity and linkage disequilibrium at different inversion regions evidenced coadaptation for different geographical clines (Kennington et al., 2006). Therefore, distinct inversion ‘haplogroups’ can hold together favorable combinations of alleles that act together to lead to adaptive shifts. Low nucleotide diversity reflect genomic regions with low rates of meiotic crossing-over, as is the case around most inversion breakpoints. Interestingly, gene conversion exists within inversions of two Drosophila species hybrids even near inversion breakpoints (Korunes & Noor, 2018). Thus, nucleotide differences among ‘haplogroups’ as well their frequency in a population can unravel the evolutionary history of an inversion.

The existence of such a large and complex inversion, in approximately 5% of the great tits, posed questions about the possible phenotypic effects as well as biological mechanisms maintaining it in such a substantial frequency. The hypothesis that the inversion is the result of genetic drift is disputable (see **Chapter 5**) because (i) there is a high number of genes affected, increasing the chance of a phenotypic effect, (ii) homozygotes were not found, suggesting otherwise a recessive lethal variant. Moreover, apart from all the minor SNP alleles found to be close to fixation across the inversion, the CNV tagging the inversion (i.e. a CNV located within ‘CNVR 2802’) was shown to be partially overlapping three important genes. Therefore, these genes could be disrupted in carriers, which would lead to important phenotypic implications. Given possible phenotypic effects of the inversion, in **Chapter 6** I have investigated the association of the inversion with seasonal measurements (e.g. egg-laying dates and number of fledged chicks) to search for fitness advantage and deviations from Mendelian inheritance (i.e. indicating a selfish gene).

7.7 A recessive lethal and selfish inversion

A lack of homozygotes for the inversion was the first indication that it could be a recessive lethal arrangement. However, given the observed inversion allele frequency of ≈2.5%, the number observed homozygotes might be zero just due to the low likelihood of sampling these individuals. To properly identify recessive lethal variants by observed/expected genotype frequency ratios, the allele frequency of the variant needs to be considerable or the sample population needs be large. For example, in pigs more than 24,000 animals were used to scan for recessive lethal variants in the pig genome (Derks et al., 2017). Thus, to overcome the statistical limitation of using expected genotype proportions in a population with limited size, the offspring ratios and the number of hatched eggs in carrier-by-carrier mating pairs were instead explored. The homozygous lethality of the inversion in great tit was supported by the fact that no homozygotes from these carrier-by-carrier matings were found and the proportion of heterozygous was approximately 65% (i.e. fitting a model for a recessive lethal gene). Moreover, the number of hatched eggs in carrier-by-carriers is significantly lower, suggesting that homozygous embryos cannot be properly formed or have their development halted at some later stage (**Chapter 6**). However, to disclose the molecular mechanisms involved in the inversion lethality further studies on the development and gene expression of different embryonic stages in homozygotes should be performed.

In most of the cases, the function of a gene cannot be determined by simply identifying amino acid motifs in their proteins (Iredale, 1999) or by examining closely related family members (Hall et al., 2009). Alternatively, gene knockout can be used to uncover the phenotypic effects of a candidate gene mutation. Until recently, gene editing was a task that has considerable technical challenges involved. However, CRISPR-Cas9 has been shown to be a cost-effective and easy-to-use method to precisely and efficiently modify genomic loci of a wide array of cells and organisms (Doudna & Charpentier, 2014). Thus, CRISPR-Cas9 could be an alternative to generate modified bird embryos that are homozygous or heterozygous at a specific candidate gene (Paquet et al., 2016). By producing homozygous embryos, their development could be studied in detail. Otherwise, heterozygous embryos could be used to generate adult birds, which can be crossed to reveal if their offspring have viable homozygous or not. However, as the inversion encompasses almost 1,000 genes, the testing of all these genes can become costly and exhaustive. A gene from PI3K family that leads to embryonic lethality in mouse (Bi et al., 1999) is likely disrupted in the inversion (**Chapter 6**). As a preliminary test before designing a gene editing essay for this gene, the offspring ratio from pairs for which both parents are non-carriers and have a *CNV call* located at the 'CNVR 2802' could also be analyzed (i.e. such as was done for the carrier-by-carrier offspring in **Chapter 6**). If the offspring ratio of these pairs are similar to results observed in carrier-by-carriers, three likely disrupted genes within this CNVR, including *PIK3C2G* gene, will become the main candidates to explain the inversion recessive lethality. Additionally, the other 29 genes overlapped by 'CNVR 2802' can be also considered as candidates.

Lethal alleles tend to be purged from a population if their fitness is lower or similar to the homologous ancestral allele. By contrast, if a lethal allele has a fitness advantage, it could be maintained in the population by balancing selection (Derks et al., 2018). In addition, independently from fitness advantage, an allele can be maintained in a population by meiotic drive (Chevin & Hospital, 2006). Meiotic drive, or segregation distortion, is a phenomenon in which a given genetic variant is inherited more than expected by the Mendelian law (i.e. the chance to be inherited is higher than 50% and therefore labeled as a 'selfish gene'). Mechanisms underlying meiotic drive can include an unbalanced production of the lethal allele during the spermatogenesis or a motility advantage of the carrier sperm. Therefore, in **Chapter 6**, the offspring from carrier-by-normal mating pairs was analyzed to explore deviations from expected genotype ratios. In carrier-by-carrier pairs where the father was the carrier, the proportion of offspring carries was approximately 70% instead of the 50% that is expected in a variant following Mendelian law. Therefore, the maintenance of the inversion may be at least partially explained by its selfish nature, which can increase the inversion frequency even when a mild heterozygous fitness disadvantage is followed by a homozygous lethality.

There are known mechanisms of meiotic drive where the carrier gamete overcomes the competition by killing the alternative gametes, reviewed in (Bravo Núñez et al., 2018)). Alternatively, the meiotic element can confer motility advantage for gametes that harbor it, such is the case in zebra finch where the heterozygotes males for a supergene have the fastest and most successful sperm (Kim et al., 2017). The selfishness of the great tit inversion discussed in **Chapter 6** has probably a sperm-related mechanistic background because in carrier-by-normal pairs for which the mother is the carrier, the inversion inheritance simply follows Mendelian law. Therefore, the sperm quality and proportion of the sperms harboring the inversion allele may help to clarify which biological mechanism is underlying the meiotic drive of this inversion. Moreover, the analysis of the inversion inheritance pattern specifically for birds with the alternative 'haplogroups' in the center of the inversion could clarify if the gene underlying the meiotic drive is located in this regions. It is interesting to note that a gene underlying meiotic drive in Drosophila, i.e. *RANGAP1*, is also located in the center of the Chromosome 1A in the great tit genome. Albeit the mechanism or genes that are selfishly maintaining the inversion in the great tit is still unknown, therefore deserving further investigation, the results explored in **Chapter 6** strongly support that the inversion is indeed a selfish variant.

Although a selfish arrangement, the inversion selfishness is unable to solely explain its observed frequency (**Chapter 6**). A drift-selection simulation accounting for the inversion recessive lethality and selfishness obtained a stable frequency around 2.5% (i.e. observed frequency) only when heterozygotes had a fitness disadvantage around 12.7%. Therefore, apart from the obvious disadvantage of having 25% less offspring in carrier-by-carrier matings, the heterozygous may have a disadvantage in some fitness-related measurement such as the number of fledged birds. However, we could not find such an association between the inversion and lower number of fledged birds. Although it is true that our statistical power might be not sufficient to unravel such an association, it may be important to consider that this inversion might affect the fitness of the carriers through other fitness related measures or behaviours. For example, inversions in different bird species have been associated with their mating system (Tuttle et al., 2016; Küpper et al., 2015; Lamichhaney et al., 2016; Tuttle et al., 2016), which could be also affected at some extend by this inversion in great tit.

7.8 Structural variants are needed to understand biodiversity

Although SNPs are primarily used to show how the genetic variation is reflected by evolution (e.g. phylogenetic trees, (Morin et al., 2004; Leaché & Oaks, 2017)), structural variants have been proven to be responsible for a substantial part of this evolutionary history in several species (Wellenreuther et al., 2019). Therefore, a better understanding about all different classes of structural variants can be an useful tool to further understand biodiversity in nature. Consequently, ample genomic knowledge on the structural variants affecting the biodiversity on our planet can help in future conservation programs (Khan et al., 2016), as the pace of extinction in a number of species accelerates (Ceballos et al., 2015; Collins et al., 2018).

Under a world that is changing due to the climate change, structural variants such as CNVs have been shown to be responsible for adaptability to environments that are rapidly changing (Chain et al., 2014; Prunier et al., 2017; Simam et al., 2018). In great tit, the breeding timing has been shifting due to the global warming (Visser & Both, 2005), which makes such seasonal measurements good candidates to be associated with structural variants in the genome. Although it is still inconclusive if any CNV is associated with breeding timing (**Chapter 3**), a number of interesting genes overlap CNVs suggestively associated with egg-laying dates (i.e. p-value<0.1). For example, the *KPNB1* gene mediates the circadian clock function (Lee et al., 2015) and is therefore an obvious candidate to account for variation in breeding timing.

Inversions may have a central role on biodiversity by making use of non-canonical mechanisms during their evolution. Most of the regions of the genome can freely recombine during the pairing of homologous chromosomes but inversions are an exception (Sturtevant, 1921; Kirkpatrick, 2010). An inverted sequence is unable to perform recombination with its respective allelic homologous sequence as a different sequence order prevents proper pairing. This mechanism allows genetic variants within an inversion to work as an inheritance 'unit', which can be maintained unbroken across generations (Faria et al., 2019). The impaired recombination in inversions can allow a more complex biological system such as a selfish gene (Hammer et al., 1989), which may promote its own inheritance during the gametogenesis. Although the general importance of selfish genes for evolution and ecology is still not well known (Lindholm et al., 2016), an increasing number molecular mechanisms of segregation distortion have been reported in a different number of species (Lindholm et al., 2016; Bravo Núñez et al., 2018). Nevertheless, the identification of selfish rearrangement in the genome can be tricky or nearly impossible if they already reached fixation in a given species (Bravo Núñez et al., 2018).

7.9 Thesis overview and future steps

Future efforts to improve CNV mapping in the great tit genome could make a broader use of more precise detection methods such as NGS. Although the CNVRs mapped with SNP array reflects genomic architecture as expected, their frequency is prone to be underestimated due to the apparent high number of false negatives. Thus, a CNV-dataset with more precise CNVR frequencies can facilitate future efforts to associate copy number change with phenotypes and/or fitness components in the great tit. Regarding the large inversion on Chromosome 1A, further characterization of 'haplogroups' might be essential in studies looking for the actual 'selfish' element that should be present in this inversion. For example, if the gene or genes underlying meiotic drive are located at the center of the inversion, it is likely that the alternative inversion 'haplogroup' is not a selfish arrangement. Moreover, the exploration of sperm morphology and motility in carriers, as well as the inversion quantification in their semen (by using e.g. quantitative Sanger or PCR), can shed light on which stage of the spermatogenesis the segregation distortion occurs.

References

Abyzov, A., Urban, A. E., Snyder, M., & Gerstein, M. (2011). CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*, *21*, 974–984. doi:doi: 10.1101/gr.114876.110.

Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nat. Rev. Genet.*, *12*, 363–376.

Andolfatto, P., Depaulis, F., & Navarro, A. (2001). Inversion polymorphisms and nucleotide variability in Drosophila. *Genet. Res. (Camb).*, *77*. doi:doi: 10.1017/S0016672301004955.

Antonarakis, S. E., Kazazian, H. H., & Tuddenham, E. G. D. (1995). Molecular etiology of factor VIII deficiency in hemophilia A. *Human Mutation*, *5*, 1–22. doi:doi: 10.1002/humu.1380050102.

Ayala, D., Fontaine, M. C., Cohuet, A., Fontenille, D., Vitalis, R., & Simard, F. (2011). Chromosomal Inversions, Natural Selection and Adaptation in the Malaria Vector Anopheles funestus. *Molecular Biology and Evolution*, *28*, 745–758. doi:doi: 10.1093/molbev/msq248.

Bailey, J. A., & Eichler, E. E. (2006). Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat. Rev. Genet.*, *7*, 552–564.

Bailey, J. A., Liu, G., & Eichler, E. E. (2003). An Alu Transposition Model for the Origin and Expansion of Human Segmental Duplications. *The American Journal of Human Genetics*, *73*, 823–834. doi:doi: 10.1086/378594.

van Balen, J. H. (2002). A Comparative Sudy of the Breeding Ecology of the Great Tit Parus major in Different Habitats. *Ardea*, *38-90*, 1–93. doi:doi: 10.5253/arde.v61.p1.

Barnes, C., & Plagnol, V. (2017). *CNVtools: a package to test genetic association with CNV data*. doi:doi: 10.18129/b9.bioc.cnvtools.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*. doi:doi: 10.18637/jss.v067.i01.

Batzer, M. A., & Deininger, P. L. (2002). Alu repeats and human genomic diversity. *Nature Reviews Genetics*, *3*, 370–379. doi:doi: 10.1038/nrg798.

Begum, N., Shen, W., & Manganiello, V. (2011). Role of PDE3A in regulation of cell cycle progression in mouse vascular smooth muscle cells and oocytes: implications in cardiovascular diseases and infertility. *Curr. Opin. Pharmacol.*, *11*, 725–729. doi:doi: 10.1016/j.coph.2011.10.006.

Benjamini, & Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc*, *57*, 289–300.

Bergero, R., Charlesworth, D., Filatov, D. A., & Moore, R. C. (2008). Defining Regions and Rearrangements of the Silene latifolia Y Chromosome. *Genetics*, *178*, 2045–2053. doi:doi: 10.1534/genetics.107.084566.

Bergero, R., Forrest, A., Kamau, E., & Charlesworth, D. (2007). Evolutionary Strata on the X Chromosomes of the Dioecious Plant Silene latifolia: Evidence From New Sex-Linked Genes. *Genetics*, *175*, 1945–1954. doi:doi: 10.1534/genetics.106.070110.

Bergero, R., Qiu, S., Forrest, A., Borthwick, H., & Charlesworth, D. (2013). Expansion of the Pseudo-autosomal Region and Ongoing Recombination Suppression in the Silene latifolia Sex Chromosomes. *Genetics*, *194*, 673–686. doi:doi: 10.1534/genetics.113.150755.

Beroukhim et al. (2007). Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proc Nat Acad Sci*, *104*, 20007–12.

Bi, L., Okabe, I., Bernard, D. J., Wynshaw-Boris, A., & Nussbaum, R. L. (1999). Proliferative Defect and Embryonic Lethality in Mice Homozygous for a Deletion in the p110α Subunit of Phosphoinositide 3-Kinase. *J. Biol. Chem.*, *274*, 10963–10968. doi:doi: 10.1074/jbc.274.16.10963.

BirdLife (2019). Birdlife international species factsheet: Parus major.

Bishop, R. (2010). Applications of fluorescence in situ hybridization (FISH) in detecting genetic aberrations of medical significance. *Bioscience Horizons*, *3*, 85–95. doi:doi: 10.1093/biohorizons/hzq009.

Blakey, J. K. (2008). Genetic evidence for extra-pair fertilizations in a monogamous passerine, the Great Tit Parus major. *Ibis*, *136*, 457–462. doi:doi: 10.1111/j.1474-919X.1994.tb01122.x.

Blaustein, M. P. (1988). Calcium transport and buffering in neurons. *Trends Neurosci.*, *11*, 438–443.

Bosse, M. et al. (2017). Recent natural selection causes adaptive evolution of an avian polygenic trait. *Science (80-.).*, *358*, 365–368. doi:doi: 10.1126/science.

aal3298.

Branca, A. et al. (2011). Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume medicago truncatula. *Proceedings of the National Academy of Sciences*, *108*, E864–E870.

Bravo Núñez, M. A., Nuckolls, N. L., & Zanders, S. E. (2018). Genetic Villains: Killer Meiotic Drivers. *Trends in Genetics*, *34*, 424–433. doi:doi: 10.1016/j.tig.2018.02.003.

Browning, S. R., & Browning, B. L. (2007). Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *Am. J. Hum. Genet.*, *81*, 1084–1097. doi:doi: 10.1086/521987.

Bryois, J., Buil, A., Evans, D. M., Kemp, J. P., Montgomery, S. B., Conrad, D. F., Ho, K. M., Ring, S., Hurles, M., Deloukas, P., Davey Smith, G., & Dermitzakis, E. T. (2014). Cis and Trans Effects of Human Genomic Variants on Gene Expression. *PLoS Genet.*, *10*, e1004461.

Buse, A., Dury, S. J., Woodburn, R. J. W., Perrins, C. M., & Good, J. E. G. (1999). Effects of elevated temperature on multi-species interactions: the case of Pedunculate Oak, Winter Moth and Tits. *Funct. Ecol.*, *13*, 74–82.

Calvete, O., Gonzalez, J., Betran, E., & Ruiz, A. (2012). Segmental Duplication, Microinversion, and Gene Loss Associated with a Complex Inversion Breakpoint Region in Drosophila. *Mol. Biol. Evol.*, *29*, 1875–1889. doi:doi: 10.1093/molbev/mss067.

Carlson, M. (2017). org.Hs.eg.db: Genome wide annotation for Human.

Carter, N. P. (2007). Methods and strategies for analyzing copy number variation using DNA microarrays. *Nature Genetics*, *39*, S16–S21. doi:doi: 10.1038/ng2028.

Carvalho, C. M. B., & Lupski, J. R. (2016). Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.*, *17*, 224–238. doi:doi: 10.1038/nrg.2015.25.

Carvalho, C. M. B., Pehlivan, D., Ramocki, M. B., Fang, P., Alleva, B., Franco, L. M., Belmont, J. W., Hastings, P. J., & Lupski, J. R. (2013). Replicative mechanisms for CNV formation are error prone. *Nat. Genet.*, *45*, 1319–1326.

Casci, T. (2010). SNPs that come in threes. *Nature Reviews Genetics*, *11*, 8–8. doi:doi: 10.1038/nrg2725.

Cauchoix, M., Hermer, E., Chaine, A. S., & Morand-Ferron, J. (2017). Cognition in the field: comparison of reversal learning performance in captive and wild passerines. *Scientific Reports*, *7*, 12945. doi:doi: 10.1038/s41598-017-13179-5.

Ceballos, G., Ehrlich, P. R., Barnosky, A. D., García, A., Pringle, R. M., & Palmer,

T. M. (2015). Accelerated modern human–induced species losses: Entering the sixth mass extinction. *Science Advances*, *1*, e1400253. doi:doi: 10.1126/sciadv.1400253.

Chain, F. J. J., Feulner, P. G. D., Panchal, M., Eizaguirre, C., Samonte, I. E., Kalbe, M., Lenz, T. L., Stoll, M., Bornberg-Bauer, E., Milinski, M., & Reusch, T. B. H. (2014). Extensive Copy-Number Variation of Young Genes across Stickleback Populations. *PLoS Genetics*, *10*, e1004830. doi:doi: 10.1371/journal.pgen.1004830.

Chambers, J., & Rabbitts, T. H. (2015). LMO2 at 25 years: a paradigm of chromosomal translocation proteins. *Open Biol.*, *5*, 150062. doi:doi: 10.1098/rsob.150062.

Chao, Y.-L., Chien, W.-H., Liao, H.-M., Fang, J.-S., & Chen, C.-H. (2009). Copy Number Variations and Psychiatric Disorders. *Tzu Chi Medical Journal*, *21*, 197–203. doi:doi: 10.1016/S1016-3190(09)60039-2.

Chari, A., Golas, M. M., Klingenhäger, M., Neuenkirchen, N., Sander, B., Englbrecht, C., Sickmann, A., Stark, H., & Fischer, U. (2008). An Assembly Chaperone Collaborates with the SMN Complex to Generate Spliceosomal SnRNPs. *Cell*, *135*, 497–509. doi:doi: 10.1016/j.cell.2008.09.020.

Chaudhry, S. R., Lwin, N., Phelan, D., Escalante, A. A., & Battistuzzi, F. U. (2018). Comparative analysis of low complexity regions in Plasmodia. *Scientific Reports*, *8*, 335. doi:doi: 10.1038/s41598-017-18695-y.

Chen, J.-M., Stenson, P. D., Cooper, D. N., & Férec, C. (2005). A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. *Human Genetics*, *117*, 411–427. doi:doi: 10.1007/s00439-005-1321-0.

Chen, K. et al. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods*, *6*, 677–681. doi:doi: 10.1038/nmeth.1363.

Chevin, L.-M., & Hospital, F. (2006). The Hitchhiking Effect of an Autosomal Meiotic Drive Gene: TABLE 1. *Genetics*, *173*, 1829–1832. doi:doi: 10.1534/genetics.105.052977.

Chiang, C., Layer, R. M., Faust, G. G., Lindberg, M. R., Rose, D. B., Garrison, E. P., Marth, G. T., Quinlan, A. R., & Hall, I. M. (2015). SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat. Methods*, *12*, 966–968. doi:doi: 10.1038/nmeth.3505.

Chmátal, L., Gabriel, S. I., Mitsainas, G. P., Martínez-Vargas, J., Ventura, J., Searle, J. B., Schultz, R. M., & Lampson, M. A. (2014). Centromere Strength Provides the Cell Biological Basis for Meiotic Drive and Karyotype Evolution in Mice. *Current Biology*, *24*, 2295–2300. doi:doi: 10.1016/j.cub.2014.08.017.

Clapham, D. E. (2007). Calcium Signaling. *Cell*, *131*, 1047–1058.

Clayton, D. (2015). snpStats: SnpMatrix and XSnpMatrix classes and methods.

Clop, a., Vidal, O., & Amills, M. (2012). Copy number variation in the genomes of domestic animals. *Anim. Genet.*, *43*, 503–17. doi:doi: 10.1111/j.1365-2052.2012.02317.x.

Coates, D. J., Byrne, M., & Moritz, C. (2018). Genetic Diversity and Conservation Units: Dealing With the Species-Population Continuum in the Age of Genomics. *Frontiers in Ecology and Evolution*, *6*. doi:doi: 10.3389/fevo.2018.00165.

Colicelli, J. (2004). Human RAS Superfamily Proteins and Related GTPases. *Sci. Signal.*, *2004*, re13–re13. doi:doi: 10.1126/stke.2502004re13.

Collins, K. S., Edie, S. M., Hunt, G., Roy, K., & Jablonski, D. (2018). Extinction risk in extant marine species integrating palaeontological and biodistributional data. *Proceedings of the Royal Society B: Biological Sciences*, *285*, 20181698. doi:doi: 10.1098/rspb.2018.1698.

Conover, C. A., Bale, L. K., & Nair, K. S. (2016). Comparative gene expression and phenotype analyses of skeletal muscle from aged wild-type and PAPP-A-deficient mice. *Exp. Gerontol.*, *80*, 36–42.

Conrad et al. (2010). Origins and functional impact of copy number variation in the human genome. *Nature*, *464*, 704–12.

Corsini, M., Dubiec, A., Marrot, P., & Szulkin, M. (2017). Humans and Tits in the City: Quantifying the Effects of Human Presence on Great Tit and Blue Tit Reproductive Trait Variation. *Frontiers in Ecology and Evolution*, *5*. doi:doi: 10.3389/fevo.2017.00082.

D'Angelo, C. S., & Koiffmann, C. P. (2012). Copy number variants in obesity-related syndromes: Review and perspectives on novel molecular approaches. *Journal of Obesity*, *2012*, 1–15. doi:doi: 10.1155/2012/845480.

Deem, A., Keszthelyi, A., Blackgrove, T., Vayl, A., Coffey, B., Mathur, R., Chabes, A., & Malkova, A. (2011). Break-Induced Replication Is Highly Inaccurate. *PLoS Biol.*, *9*, e1000594.

Deng, M., Boopathi, E., Hypolite, J. A., Raabe, T., Chang, S., Zderic, S., Wein, A. J., & Chacko, S. (2013). Amino acid mutations in the caldesmon COOH-terminal functional domain increase force generation in bladder smooth muscle. *American Journal of Physiology-Renal Physiology*, *305*, F1455–F1465. doi:doi: 10.1152/ajprenal.00174.2013.

Dennenmoser, S., Sedlazeck, F. J., Iwaszkiewicz, E., Li, X.-Y., Altmüller, J., & Nolte, A. W. (2017). Copy number increases of transposable elements and protein-coding genes in an invasive fish of hybrid origin. *Mol. Ecol.*, .

Derks, M. F. L., Lopes, M. S., Bosse, M., Madsen, O., Dibbits, B., Harlizius, B., Groenen, M. A. M., & Megens, H.-J. (2018). Balancing selection on a recessive lethal deletion with pleiotropic effects on two neighboring genes in the porcine genome. *PLOS Genetics*, *14*, e1007661. doi:doi: 10.1371/journal.pgen.1007661.

Derks, M. F. L., Megens, H.-J., Bosse, M., Lopes, M. S., Harlizius, B., & Groenen, M. A. M. (2017). A systematic survey to identify lethal recessive variation in highly managed pig populations. *BMC Genomics*, *18*, 858. doi:doi: 10.1186/s12864-017-4278-1.

Derks, M. F. L., Schachtschneider, K. M., Madsen, O., Schijlen, E., Verhoeven, K. J. F., & van Oers, K. (2016). Gene and transposable element methylation in great tit (Parus major) brain and blood. *BMC Genomics*, *17*, 332.

D'haene, B., Vandesompele, J., & Hellemans, J. (2010). Accurate and objective copy number profiling using real-time quantitative PCR. *Methods*, *50*, 262–70.

Didion, J. P. et al. (2015). A Multi-Megabase Copy Number Gain Causes Maternal Transmission Ratio Distortion on Mouse Chromosome 2. *PLOS Genetics*, *11*, e1004850. doi:doi: 10.1371/journal.pgen.1004850.

Diskin, S. J., Li, M., Hou, C., Yang, S., Glessner, J., Hakonarson, H., Bucan, M., Maris, J. M., & Wang, K. (2008). Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res.*, *36*, e126.

Doe, J., & Smith, R. (2016). Title of the paper. *Title of the journal*, *2*, 10–30.

Doudna, J. A., & Charpentier, E. (2014). The new frontier of genome engineering with CRISPR-Cas9. *Science*, *346*, 1258096–1258096. doi:doi: 10.1126/science.1258096.

Ekblom, R. (2016). A bird's eye view of a deleterious recessive allele. *Journal of Animal Ecology*, *85*, 855–856. doi:doi: 10.1111/1365-2656.12514.

Excoffier, L., & Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.*, *12*, 921–7.

Fadista, J., Thomsen, B., Holm, L.-E., & Bendixen, C. (2010). Copy number variation in the bovine genome. *BMC Genomics*, *11*, 284.

Faria, R., Johannesson, K., Butlin, R. K., & Westram, A. M. (2019). Evolving Inversions. *Trends in Ecology & Evolution*, *34*, 239–248. doi:doi: 10.1016/j.tree.2018.12.005.

Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. *Nature Reviews Genetics*, *7*, 85–97. doi:doi: 10.1038/nrg1767.

Fidler, A. E., van Oers, K., Drent, P. J., Kuhn, S., Mueller, J. C., & Kempenaers, B. (2007). Drd4 gene polymorphisms are associated with personality variation in a passerine bird. *Proc. R. Soc. B Biol. Sci.*, *274*, 1685–1691.

Fishman, L., & Kelly, J. K. (2015). Centromere-associated meiotic drive and female fitness variation in Mimulus. *Evolution*, *69*, 1208–1218. doi:doi: 10.1111/evo.12661.

Forer, L., Schönherr, S., Weissensteiner, H. et al. (2010). CONAN: copy number variation analysis software for genome-wide association studies. *BMC Bioinformatics*, *11*, 318.

Fox, J., & Weisberg, S. (2011). *An R Companion to Applied Regression.* (2nd ed.). Thousand Oaks CA: Sage.

Franchitto, A. (2013). Genome Instability at Common Fragile Sites: Searching for the Cause of Their Instability. *Biomed Res. Int.*, *2013*, 1–9.

Frankham, R., Ballou, J. D., & Briscoe, D. A. (2009). *Introduction to Conservation Genetics*. Cambridge University Press. doi:doi: 10.1017/cbo9780511809002.

Fungtammasan, A., Walsh, E., Chiaromonte, F., Eckert, K. A., & Makova, K. D. (2012). A genome-wide analysis of common fragile sites: What features determine chromosomal instability in the human genome? *Genome Res.*, *22*, 993–1005.

Furuta, Y., Kawai, M., Yahara, K., Takahashi, N., Handa, N., Tsuru, T., Oshima, K., Yoshida, M., Azuma, T., Hattori, M., Uchiyama, I., & Kobayashi, I. (2011). Birth and death of genes linked to chromosomal inversion. *Proc. Natl. Acad. Sci.*, *108*, 1501–1506. doi:doi: 10.1073/pnas.1012579108.

Geistlinger, L., & da Silva, V. H. (2019). Cnvranger. doi:doi: 10.18129/b9.bioc.cnvranger.

Geistlinger, L., da Silva, V. H., Cesar, A. S. M., Tizioto, P. C., Waldron, L., Zimmer, R., Regitano, L. C. d. A., & Coutinho, L. L. (2018). Widespread modulation of gene expression by copy number variation in skeletal muscle. *Sci. Rep.*, *8*, 1399. doi:doi: 10.1038/s41598-018-19782-4.

Gel et al. (2015). regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics*, *32(2)*, 289–91.

Gienapp, P., & Bregnballe, T. (2012). Fitness Consequences of Timing of Migration and Breeding in Cormorants. *PLoS One*, *7*, e46165. doi:doi: 10.1371/journal.pone.0046165.

Gienapp, P., Hemerik, L., & Visser, M. E. (2005). A new statistical tool to predict phenology under climate change scenarios. *Glob. Chang. Biol.*, *11*, 600–606. doi:doi: 10.1111/j.1365-2486.2005.00925.x.

Gienapp, P., Laine, V. N., Mateman, A. C., van Oers, K., & Visser, M. E. (2017). Environment-Dependent Genotype-Phenotype Associations in Avian Breeding Time. *Front. Genet.*, *8*. doi:doi: 10.3389/fgene.2017.00102.

Golbabapour, S., Majid, N. A., Hassandarvish, P., Hajrezaie, M., Abdulla, M. A., & Hadi, A. H. A. (2013). Gene Silencing and Polycomb Group Proteins: An Overview of their Structure, Mechanisms and Phylogenetics. *Omi. A J. Integr. Biol.*, *17*, 283–296. doi:doi: 10.1089/omi.2012.0105.

Gonzalez, E. (2005). The Influence of CCL3L1 Gene-Containing Segmental Duplications on HIV-1/AIDS Susceptibility. *Science*, *307*, 1434–1440. doi:doi: 10.1126/science.1101160.

Green, P., & MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*, 493–498. doi:doi: 10.1111/2041-210X.12504.

Grijzenhout, A., Godwin, J., Koseki, H., Gdula, M. R., Szumska, D., McGouran, J. F., Bhattacharya, S., Kessler, B. M., Brockdorff, N., & Cooper, S. (2016). Functional analysis of AEBP2, a PRC2 Polycomb protein, reveals a Trithorax phenotype in embryonic development and in ESCs. *Development*, *143*, 2716–2723. doi:doi: 10.1242/dev.123935.

Grüebler, M. U., & Naef-Daenzer, B. (2010). Fitness consequences of timing of breeding in birds: date effects in the course of a reproductive episode. *Journal of Avian Biology*, *41*, 282–291. doi:doi: 10.1111/j.1600-048X.2009.04865.x.

Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, *32*, 2847–2849. doi:doi: 10.1093/bioinformatics/btw313.

Hall, B., Limaye, A., & Kulkarni, A. B. (2009). Overview: Generation of Gene Knockout Mice. In *Current Protocols in Cell Biology*. Hoboken, NJ, USA: John Wiley & Sons, Inc. doi:doi: 10.1002/0471143030.cb1912s44.

Hammer, M. F., Schimenti, J., & Silver, L. M. (1989). Evolution of mouse chromosome 17 and the origin of inversions associated with t haplotypes. *Proceedings of the National Academy of Sciences*, *86*, 3261–3265. doi:doi: 10.1073/pnas.86.9.3261.

Hardison, R. C. (2003). Comparative Genomics. *PLoS Biology*, *1*, e58. doi:doi: 10.1371/journal.pbio.0000058.

Harewood, L., Kishore, K., Eldridge, M. D., Wingett, S., Pearson, D., Schoenfelder, S., Collins, V. P., & Fraser, P. (2017). Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. *Genome Biology*, *18*, 125. doi:doi: 10.1186/s13059-017-1253-8.

Harris, R. A. et al. (2013). Confounding by Repetitive Elements and CpG Islands Does Not Explain the Association between Hypomethylation and Genomic Instability. *PLoS Genet.*, *9*, e1003333.

Hasson, E., & Eanes, W. F. (1996). Contrasting histories of three gene regions associated with In(3L)Payne of Drosophila melanogaster. *Genetics*, *144*, 1565–75.

Hastings, P. J., Ira, G., & Lupski, J. R. (2009). A Microhomology-Mediated Break-Induced Replication Model for the Origin of Human Copy Number Variation. *PLoS Genetics*, *5*, e1000327. doi:doi: 10.1371/journal.pgen.1000327.

Hau, M. (2001). Timing of Breeding in Variable Environments: Tropical Birds as Model Systems. *Hormones and Behavior*, *40*, 281–290. doi:doi: 10.1006/hbeh.2001.1673.

Hedrick, P. W. (1987). Gametic disequilibrium measures: proceed with caution. *Genetics*, *117*, 331–41.

Hellen, E. H. (2015). Inversions and Evolution of the Human Genome. In *eLS* (pp. 1–6). Chichester, UK: John Wiley & Sons, Ltd. doi:doi: 10.1002/9780470015902.a0026320.

Helm, B., & Visser, M. E. (2010). Heritable circadian period length in a wild bird population. *Proc. R. Soc. B Biol. Sci.*, *277*, 3335–3342. doi:doi: 10.1098/rspb.2010.0871.

Hillier, L. W. et al. (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, *432*, 695–716.

Hoffmann, A. A., & Rieseberg, L. H. (2008). Revisiting the Impact of Inversions in Evolution: From Population Genetic Markers to Drivers of Adaptive Shifts and Speciation? *Annu. Rev. Ecol. Evol. Syst.*, *39*, 21–42. doi:doi: 10.1146/annurev.ecolsys.39.110707.173532.

Hoglund, P. J., Nordstrom, K. J. V., Schioth, H. B., & Fredriksson, R. (2011). The Solute Carrier Families Have a Remarkably Long Evolutionary History with the Majority of the Human Families Present before Divergence of Bilaterian Species. *Mol. Biol. Evol.*, *28*, 1531–1541. doi:doi: 10.1093/molbev/msq350.

Hooper, D. M., & Price, T. D. (2017). Chromosomal inversion differences correlate with range overlap in passerine birds. *Nat. Ecol. Evol.*, *1*, 1526–1534. doi:doi: 10.1038/s41559-017-0284-6.

Hsu, L. Y. F., Benn, P. A., Tannenbaum, H. L., Perlis, T. E., Carlson, A. D., Opitz, J. M., & Reynolds, J. F. (1987). Chromosomal polymorphisms of 1, 9, 16, and Y in 4 major ethnic groups: A large prenatal study. *American Journal of Medical Genetics*, *26*, 95–101. doi:doi: 10.1002/ajmg.1320260116.

Huang, Y.-C., Dang, V. D., Chang, N.-C., & Wang, J. (2018). Multiple large inversions and breakpoint rewiring of gene expression in the evolution of the fire ant social supergene. *Proceedings of the Royal Society B: Biological Sciences*, *285*, 20180221. doi:doi: 10.1098/rspb.2018.0221.

Huber et al. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods*, *12*, 115–21.

Husby, A., Visser, M. E., & Kruuk, L. E. B. (2011). Speeding Up Microevolution: The Effects of Increasing Temperature on Selection and Genetic Variance in a Wild Bird Population. *PLoS Biol.*, *9*, e1000585. doi:doi: 10.1371/journal.pbio.1000585.

Ionita-Laza, I., Rogers, A. J., Lange, C., Raby, B. A., & Lee, C. (2009). Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. *Genomics*, *93*, 22–26. doi:doi: 10.1016/j.ygeno.2008.08.012.

Iredale, J. P. (1999). Demystified ... gene knockouts. *Molecular pathology : MP*, *52*, 111–6. doi:doi: 10.1136/mp.52.3.111.

Itsara, A. et al. (2009). Population Analysis of Large Copy Number Variants and Hotspots of Human Genetic Disease. *Am. J. Hum. Genet.*, *84*, 148–161. doi:doi: 10.1016/j.ajhg.2008.12.014.

Ji, H., Long, V., Briody, V., & Chien, E. K. (2011). Progesterone Modulates Integrin α2 (ITGA2) and α11 (ITGA11) in the Pregnant Cervix. *Reproductive Sciences*, *18*, 156–163. doi:doi: 10.1177/1933719110382305.

Jiang, W., Wei, M., Liu, M., Pan, Y., Cao, D., Yang, X., & Zhang, C. (2017). Identification of Protein Tyrosine Phosphatase Receptor Type O (PTPRO) as a Synaptic Adhesion Molecule that Promotes Synapse Formation. *J. Neurosci.*, *37*, 9828–9843. doi:doi: 10.1523/JNEUROSCI.0729-17.2017.

Jones, F. C. et al. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, *484*, 55–61. doi:doi: 10.1038/nature10944.

Joober, R., & Boksa, P. (2009). A new wave in the genetics of psychiatric disorders: the copy number variant tsunami. *Journal of psychiatry & neuroscience : JPN*, *34*, 55–9.

Kapun, M., Fabian, D. K., Goudet, J., & Flatt, T. (2016). Genomic Evidence for Adaptive Inversion Clines in Drosophila melanogaster. *Molecular Biology and Evolution*, *33*, 1317–1336. doi:doi: 10.1093/molbev/msw016.

Kapusta, A., & Suh, A. (2016). Evolution of bird genomes-a transposon's-eye view. *Ann. N. Y. Acad. Sci.*, .

Katju, V., & Bergthorsson, U. (2013). Copy-number changes in evolution: rates, fitness effects and adaptive significance. *Frontiers in Genetics*, *4*. doi:doi: 10.3389/fgene.2013.00273.

Keel, B. N., Lindholm-Perry, A. K., & Snelling, W. M. (2016). Evolutionary and Functional Features of Copy Number Variation in the Cattle Genome1. *Frontiers in Genetics*, *7*. doi:doi: 10.3389/fgene.2016.00207.

Kehrer-Sawatzki, H., & Cooper, D. N. (2008). Molecular mechanisms of chromosomal rearrangement during primate evolution. *Chromosom. Res.*, *16*, 41–56. doi:doi: 10.1007/s10577-007-1207-1.

Kelemen, R. K., & Vicoso, B. (2018). Complex History and Differentiation Patterns of the t -Haplotype, a Mouse Meiotic Driver. *Genetics*, *208*, 365–375. doi:doi: 10.1534/genetics.117.300513.

Keller, L., & Ross, K. G. (1998). Selfish genes: a green beard in the red fire ant. *Nature*, *394*, 573–575. doi:doi: 10.1038/29064.

Kendall, K. M., Rees, E., Escott-Price, V., Einon, M., Thomas, R., Hewitt, J., O'Donovan, M. C., Owen, M. J., Walters, J. T., & Kirov, G. (2017). Cognitive Performance Among Carriers of Pathogenic Copy Number Variants: Analysis of 152,000 UK Biobank Subjects. *Biological Psychiatry*, *82*, 103–110. doi:doi: 10.1016/j.biopsych.2016.08.014.

Kennington, W. J., Partridge, L., & Hoffmann, A. A. (2006). Patterns of Diversity and Linkage Disequilibrium Within the Cosmopolitan Inversion In(3R)Payne in Drosophila melanogaster Are Indicative of Coadaptation. *Genetics*, *172*, 1655–1663. doi:doi: 10.1534/genetics.105.053173.

Kentie, R., Coulson, T., Hooijmeijer, J. C. E. W., Howison, R. A., Loonstra, A. H. J., Verhoeven, M. A., Both, C., & Piersma, T. (2018). Warming springs and habitat alteration interact to impact timing of breeding and population dynamics in a migratory bird. *Global Change Biology*, *24*, 5292–5303. doi:doi: 10.1111/gcb.14406.

Khaja, R., MacDonald, J. R., Zhang, J., & Scherer, S. W. (2006). Methods for Identifying and Mapping Recent Segmental and Gene Duplications in Eukaryotic Genomes. In *Gene Mapping, Discov. Expr.* (pp. 9–20). Humana Press.

Khan, S., Nabi, G., Ullah, M. W., Yousaf, M., Manan, S., Siddique, R., & Hou, H. (2016). Overview on the Role of Advance Genomics in Conservation Biology of Endangered Species. *International Journal of Genomics*, *2016*, 1–8. doi:doi: 10.1155/2016/3460416.

Khurana, E., Lam, H. Y. K., Cheng, C., Carriero, N., Cayting, P., & Gerstein, M. B. (2010). Segmental duplications in the human genome reveal details of pseudogene formation. *Nucleic Acids Res.*, *38*, 6997–7007.

Kim et al. (2012). CNVRuler: a copy number variation-based case-control association analysis tool. *Bioinformatics*, *28*, 1790–2.

Kim, H., Kang, K., Ekram, M. B., Roh, T.-Y., & Kim, J. (2011). Aebp2 as an Epigenetic Regulator for Neural Crest Cells. *PLoS One*, *6*, e25174. doi:doi: 10.1371/journal.pone.0025174.

Kim, J.-M., Santure, A. W., Barton, H. J., Quinn, J. L., Cole, E. F., Visser, M. E., Sheldon, B. C., Groenen, M. A. M., van Oers, K., & Slate, J. (2018). A high-density SNP chip for genotyping great tit (*Parus major*) populations and its application to studying the genetic architecture of exploration behaviour. *Molecular Ecology Resources*, *18*, 877–891. doi:doi: 10.1111/1755-0998.12778.

Kim, K.-W., Bennison, C., Hemmings, N., Brookes, L., Hurley, L. L., Griffith, S. C., Burke, T., Birkhead, T. R., & Slate, J. (2017). A sex-linked supergene controls sperm morphology and swimming speed in a songbird. *Nat. Ecol. Evol.*, *1*, 1168–1176. doi:doi: 10.1038/s41559-017-0235-2.

Kirkpatrick, M. (2006). Chromosome Inversions, Local Adaptation and Speciation. *Genetics*, *173*, 419–434. doi:doi: 10.1534/genetics.105.047985.

Kirkpatrick, M. (2010). How and Why Chromosome Inversions Evolve. *PLoS Biol.*, *8*, e1000501. doi:doi: 10.1371/journal.pbio.1000501.

Kirov, G. (2015). CNVs in neuropsychiatric disorders. *Human Molecular Genetics*, *24*, R45–R49. doi:doi: 10.1093/hmg/ddv253.

Knief, U., Hemmrich-Stanisak, G., Wittig, M., Franke, A., Griffith, S. C., Kempenaers, B., & Forstmeier, W. (2016). Fitness consequences of polymorphic inversions in the zebra finch genome. *Genome Biology*, *17*, 199. doi:doi: 10.1186/s13059-016-1056-3.

Kondrashov, F. A. (2012). Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proc. R. Soc. B Biol. Sci.*, *279*, 5048–5057.

Koneswaran, G., & Nierenberg, D. (2008). Global Farm Animal Production and Global Warming: Impacting and Mitigating Climate Change. *Environmental Health Perspectives*, *116*, 578–582. doi:doi: 10.1289/ehp.11034.

Korunes, K. L., & Noor, M. A. F. (2018). Pervasive gene conversion in chromosomal inversion heterozygotes. *Molecular Ecology*, (p. mec.14921). doi:doi: 10.1111/mec.14921.

Küpper, C. et al. (2015). A supergene determines highly divergent male reproductive morphs in the ruff. *Nat. Genet.*, *48*, 79–83. doi:doi: 10.1038/ng.3443.

Kvist, L., Martens, J., Higuchi, H., Nazarenko, A. A., Valchuk, O. P., & Orell, M. (2003). Evolution and genetic structure of the great tit (Parus major) complex. *Proc. R. Soc. B Biol. Sci.*, *270*, 1447–1454.

Laine, V. N. et al. (2016). Evolutionary signals of selection on cognition from the great tit genome and methylome. *Nat. Commun.*, *7*, 10474.

Lamichhaney, S. et al. (2016). Structural genomic changes underlie alternative reproductive strategies in the ruff (Philomachus pugnax). *Nature Genetics*, *48*, 84–88. doi:doi: 10.1038/ng.3430.

Lanzetti, L., Rybin, V., Malabarba, M. G., Christoforidis, S., Scita, G., Zerial, M., & Di Fiore, P. P. (2000). The Eps8 protein coordinates EGF receptor signalling through Rac and trafficking through Rab5. *Nature*, *408*, 374–377. doi:doi: 10.1038/35042605.

Larkin, D. M., Pape, G., Donthu, R., Auvil, L., Welge, M., & Lewin, H. A. (2009). Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. *Genome Research*, *19*, 770–777. doi:doi: 10.1101/gr.086546.108.

Larracuente, A. M., & Presgraves, D. C. (2012). The Selfish Segregation Distorter Gene Complex of Drosophila melanogaster. *Genetics*, *192*, 33–53. doi:doi: 10.1534/genetics.112.141390.

Larsen, S. J., do Canto, L. M., Rogatto, S. R., & Baumbach, J. (2018). CoNVaQ: a web tool for copy number variation-based association studies. *BMC Genomics*, *19*, 369.

Lawrence et al. (2013). Software for computing and annotating genomic ranges. *PLoS Comput Biol*, *9*, e1003118.

Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.*, *15*, R84. doi:doi: 10.1186/gb-2014-15-6-r84.

Leaché, A. D., & Oaks, J. R. (2017). The Utility of Single Nucleotide Polymorphism (SNP) Data in Phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, *48*, 69–84. doi:doi: 10.1146/annurev-ecolsys-110316-022645.

Lee, J. A., & Lupski, J. R. (2006). Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. *Neuron*, *52*, 103–121. doi:doi: 10.1016/j.neuron.2006.09.027.

Lee, Y., Jang, A. R., Francey, L. J., Sehgal, A., & Hogenesch, J. B. (2015). KPNB1 mediates PER/CRY nuclear translocation and circadian clock function. *eLife*, *4*, e08647. doi:doi: 10.7554/eLife.08647.

Lenth, R. (2019). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.3.3.

Levy, R. J., Xu, B., Gogos, J. A., & Karayiorgou, M. (2012). Copy Number Variation and Psychiatric Disease Risk. In *Genomic Structural Variants* (pp. 97–113). Springer, New York, NY. doi:doi: 10.1007/978-1-61779-507-7_4.

Lewontin, R. C. (1964). The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics*, *49*, 49–67.

Lewontin, R. C., & Kojima, K.-i. (1960). The Evolutionary Dynamics of Complex Polymorphisms. *Evolution (N. Y).*, *14*, 458. doi:doi: 10.2307/2405995.

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., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*, 2078–2079. doi:doi: 10.1093/bioinformatics/btp352.

Li, J. et al. (2012). Genomic Hypomethylation in the Human Germline Associates with Selective Structural Mutability in the Human Genome. *PLoS Genet.*, *8*, e1002692.

Li, W., & Olivier, M. (2013). Current analysis platforms and methods for detecting copy number variation. *Physiological genomics*, *45*, 1–16. doi:doi: 10.1152/physiolgenomics.00082.2012.

Liang, C., Wang, X., Hu, J., Lian, X., Zhu, T., Zhang, H., Gu, N., & Li, J. (2017). PTPRO Promotes Oxidized Low-Density Lipoprotein Induced Oxidative Stress and Cell Apoptosis through Toll-Like Receptor 4/Nuclear Factor κB Pathway. *Cell. Physiol. Biochem.*, *42*, 495–505. doi:doi: 10.1159/000477596.

Lindholm, A. K. et al. (2016). The Ecology and Evolutionary Dynamics of Meiotic Drive. *Trends in Ecology & Evolution*, *31*, 315–326. doi:doi: 10.1016/j.tree.2016.02.001.

Littlejohn, M. D. et al. (2016). Sequence-based Association Analysis Reveals an MGST1 eQTL with Pleiotropic Effects on Bovine Milk Composition. *Sci. Rep.*, *6*, 25376. doi:doi: 10.1038/srep25376.

Liu, G. E., Brown, T., Hebert, D. a., Cardone, M. F., Hou, Y., Choudhary, R. K., Shaffer, J., Amazu, C., Connor, E. E., Ventura, M., & Gasbarre, L. C. (2011). Initial analysis of copy number variations in cattle selected for resistance or susceptibility to intestinal nematodes. *Mamm. Genome*, *22*, 111–21.

Liu, P. et al. (2014). Mechanism, Prevalence, and More Severe Neuropathy Phenotype of the Charcot-Marie-Tooth Type 1A Triplication. *Am. J. Hum. Genet.*, *94*, 462–469.

Liu, S., Yao, L., Ding, D., & Zhu, H. (2010). CCL3L1 Copy Number Variation and Susceptibility to HIV-1 Infection: A Meta-Analysis. *PLoS ONE*, *5*, e15778. doi:doi: 10.1371/journal.pone.0015778.

Locke, D. P., Sharp, A. J., McCarroll, S. a., McGrath, S. D., Newman, T. L., Cheng, Z., Schwartz, S., Albertson, D. G., Pinkel, D., Altshuler, D. M., & Eichler, E. E. (2006). Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am. J. Hum. Genet.*, *79*, 275–90.

Longo, M. S., Carone, D. M., Green, E. D., O'Neill, M. J., & O'Neill, R. J.

(2009). Distinct retroelement classes define evolutionary breakpoints demarcating sites of evolutionary novelty. *BMC Genomics*, *10*, 334. doi:doi: 10.1186/1471-2164-10-334.

Lucas, C., Nicolas, M., & Keller, L. (2015). Expression of foraging and Gp-9 are associated with social organization in the fire ant Solenopsis invicta. *Insect Mol. Biol.*, *24*, 93–104. doi:doi: 10.1111/imb.12137.

Lynch, M., Ackerman, M. S., Gout, J.-F., Long, H., Sung, W., Thomas, W. K., & Foster, P. L. (2016). Genetic drift, selection and the evolution of the mutation rate. *Nat. Rev. Genet.*, *17*, 704–714.

Lyon, M. F. (2003). Transmission Ratio Distortion in Mice. *Annual Review of Genetics*, *37*, 393–408.

Ma, J., & Amos, C. I. (2012). Investigation of Inversion Polymorphisms in the Human Genome Using Principal Components Analysis. *PLoS One*, *7*, e40224. doi:doi: 10.1371/journal.pone.0040224.

Mayle, R., Campbell, I. M., Beck, C. R., Yu, Y., Wilson, M., Shaw, C. A., Bjergbaek, L., Lupski, J. R., & Ira, G. (2015). Mus81 and converging forks limit the mutagenicity of replication fork breakage. *Science*, *349*, 742–747. doi:doi: 10.1126/science.aaa8391.

McGaugh, S. E., Heil, C. S. S., Manzano-Winkler, B., Loewe, L., Goldstein, S., Himmel, T. L., & Noor, M. A. F. (2012). Recombination Modulates How Selection Affects Linked Sites in Drosophila. *PLoS Biol.*, *10*, e1001422.

van der Meer, E., & van Oers, K. (2015). Gender and Personality Differences in Response to Social Stressors in Great Tits (Parus major). *PLoS One*, *10*, e0127984.

Meinshausen, M., Meinshausen, N., Hare, W., Raper, S. C. B., Frieler, K., Knutti, R., Frame, D. J., & Allen, M. R. (2009). Greenhouse-gas emission targets for limiting global warming to 2 °C. *Nature*, *458*, 1158–1162. doi:doi: 10.1038/nature08017.

Merrill, C. (1999). Truncated RanGAP Encoded by the Segregation Distorter Locus of Drosophila. *Science*, *283*, 1742–1745. doi:doi: 10.1126/science.283.5408.1742.

Meyers, S. N., McDaneld, T. G., Swist, S. L., Marron, B. M., Steffen, D. J., O'Toole, D., O'Connell, J. R., Beever, J. E., Sonstegard, T. S., & Smith, T. P. L. (2010). A deletion mutation in bovine SLC4A2 is associated with osteopetrosis in Red Angus cattle. *BMC genomics*, *11*, 337. doi:doi: 10.1186/1471-2164-11-337.

Mgbemene, C. A., Nnaji, C. C., & Nwozor, C. (2016). Industrialization and its Backlash: Focus on Climate Change and its Consequences. *Journal of Environmental Science and Technology*, *9*, 301–316. doi:doi: 10.3923/jest.2016.301.316.

Mitchem, K. L., Hibbard, E., Beyer, L. A., Bosom, K., Dootz, G. A., Dolan, D. F., Johnson, K. R., Raphael, Y., & Kohrman, D. C. (2002). Mutation of the novel gene Tmie results in sensory cell defects in the inner ear of spinner, a mouse model of human hearing loss DFNB6. *Hum. Mol. Genet.*, *11*, 1887–98.

Morgan, & Ramos (2017). RaggedExperiment: Representation of sparse experiments and assays across samples. DOI: 10.18129/b9.bioc.RaggedExperiment, http://bioconductor.org/packages/RaggedExperiment, .

Morgan, M., & Ramos, M. (2019). *RaggedExperiment: Representation of Sparse Experiments and Assays Across Samples*. R package version 1.7.5.

Morin, P. A., Luikart, G., Wayne, R. K., & the SNP workshop group (2004). SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution*, *19*, 208–216. doi:doi: 10.1016/j.tree.2004.01.009.

Moro, C., Cornette, R., Vieaud, A., Bruneau, N., Gourichon, D., Bed'hom, B., & Tixier-Boichard, M. (2015). Quantitative Effect of a CNV on a Morphological Trait in Chickens. *PLoS One*, *10*, e0118706.

Morrow, E. M. (2010). Genomic Copy Number Variation in Disorders of Cognitive Development. *Journal of the American Academy of Child & Adolescent Psychiatry*, *49*, 1091–1104. doi:doi: 10.1016/j.jaac.2010.08.009.

Mulder, H., Ahrén, B., & Sundler, F. (1996). Islet amyloid polypeptide and insulin gene expression are regulated in parallel by glucose in vivo in rats. *Am. J. Physiol.*, *271*, E1008–14.

Munguía-Rosas, M. A., Ollerton, J., Parra-Tabla, V., & De-Nova, J. A. (2011). Meta-analysis of phenotypic selection on flowering phenology suggests that early flowering plants are favoured. *Ecology Letters*, *14*, 511–521. doi:doi: 10.1111/j.1461-0248.2011.01601.x.

Nadeau, N. J. et al. (2016). The gene cortex controls mimicry and crypsis in butterflies and moths. *Nature*, *534*, 106–110.

Natri, H. M., Shikano, T., & Merilä, J. (2013). Progressive Recombination Suppression and Differentiation in Recently Evolved Neo-sex Chromosomes. *Molecular Biology and Evolution*, *30*, 1131–1144. doi:doi: 10.1093/molbev/mst035.

Naz, S. et al. (2002). Mutations in a Novel Gene, TMIE, Are Associated with Hearing Loss Linked to the DFNB6 Locus. *Am. J. Hum. Genet.*, *71*, 632–636. doi:doi: 10.1086/342193.

Nguyen, H. T., Merriman, T. R., & Black, M. A. (2014). The CNVrd2 package: measurement of copy number at complex loci using high-throughput sequencing data. *Front. Genet.*, *5*. doi:doi: 10.3389/fgene.2014.00248.

Nicolazzi, E. L., Iamartino, D., & Williams, J. L. (2014). AffyPipe: an open-source pipeline for Affymetrix Axiom genotyping workflow. *Bioinformatics*, *30*, 3118–

3119.

Nipitwattanaphon, M., Wang, J., Dijkstra, M. B., & Keller, L. (2013). A simple genetic basis for complex social behaviour mediates widespread gene expression differences. *Molecular Ecology*, *22*, 3797–3813. doi:doi: 10.1111/mec.12346.

Noordwijk, A. J. V., Van Balen, J. H., & Scharloo, W. (1980). Heritability of Ecologically Important Traits in the Great Tit. *Ardea*, *55(1–2)*, 193–204. doi:doi: 10.5253/arde.v68.p193.

Pagès, H., Aboyoun, P., Gentleman, R., & DebRoy, S. (2017). Biostrings: String objects representing biological sequences, and matching algorithms.

Pailhoux, E., Vigier, B., Chaffaux, S., Servel, N., Taourit, S., Furet, J. P., Fellous, M., Grosclaude, F., Cribiu, E. P., Cotinot, C., & Vaiman, D. (2001). A 11.7-kb deletion triggers intersexuality and polledness in goats. *Nature genetics*, *29*, 453–8. doi:doi: 10.1038/ng769.

Palacios, R., Palacios-Flores, K., Flores, M., & Dávila, G. (2017). Rearrangements . In *Reference Module in Life Sciences*. Elsevier. doi:doi: 10.1016/B978-0-12-809633-8.06999-5.

Paquet, D., Kwart, D., Chen, A., Sproul, A., Jacob, S., Teo, S., Olsen, K. M., Gregg, A., Noggle, S., & Tessier-Lavigne, M. (2016). Efficient introduction of specific homozygous and heterozygous mutations using CRISPR/Cas9. *Nature*, *533*, 125–129. doi:doi: 10.1038/nature17664.

Pastinen, T. (2006). Influence of human genome polymorphism on gene expression. *Hum. Mol. Genet.*, *15*, R9–R16.

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet.*, *2*, e190. doi:doi: 10.1371/journal.pgen.0020190.

Paudel, Y., Madsen, O., Megens, H.-J., Frantz, L. A. F., Bosse, M., Crooijmans, R. P. M. A., & Groenen, M. A. M. (2015). Copy number variation in the speciation of pigs: a possible prominent role for olfactory receptors. *BMC Genomics*, *16*, 330.

Peiffer, D. A. (2006). High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res.*, *16*, 1136–1148.

Perkel, J. (2008). SNP genotyping: six technologies that keyed a revolution. *Nat. Methods*, *5*, 447–453.

Perrins, C. M. (1970). THE TIMING OF BIRDS‘ BREEDING SEASONS. *Ibis*, *112*, 242–255. doi:doi: 10.1111/j.1474-919X.1970.tb00096.x.

Perry, G. (2008). The evolutionary significance of copy number variation in the human genome. *Cytogenet. Genome Res.*, *123*, 283–287. doi:doi: 10.1159/000184719.

Perry, G. H., Tchinda, J., McGrath, S. D., Zhang, J., Picker, S. R., Caceres, A. M., Iafrate, A. J., Tyler-Smith, C., Scherer, S. W., Eichler, E. E., Stone, A. C., & Lee, C. (2006). Hotspots for copy number variation in chimpanzees and humans. *Proc. Natl. Acad. Sci.*, *103*, 8006–8011.

Perry, G. H. et al. (2008). Copy number variation and evolution in humans and chimpanzees. *Genome Res.*, *18*, 1698–1710.

Pértille, F., Da Silva, V. H., Johansson, A. M., Lindström, T., Wright, D., Coutinho, L. L., Jensen, P., & Guerrero-Bosagna, C. (2019). Mutation dynamics of CpG dinucleotides during a recent event of vertebrate diversification. *Epigenetics*, (pp. 1–23). doi:doi: 10.1080/15592294.2019.1609868.

Petousi, N. et al. (2014). Erythrocytosis associated with a novel missense mutation in the BPGM gene. *Haematologica*, *99*, e201–e204. doi:doi: 10.3324/haematol.2014.109306.

Pezer, Ž., Harr, B., Teschke, M., Babiker, H., & Tautz, D. (2015). Divergence patterns of genic copy number variation in natural populations of the house mouse (Mus musculus domesticus) reveal three conserved genes with major population-specific expansions. *Genome Research*, *25*, 1114–1124. doi:doi: 10.1101/gr.187187.114.

Pittman, A. M., Fung, H.-C., & de Silva, R. (2006). Untangling the tau gene association with neurodegenerative disorders. *Human Molecular Genetics*, *15*, R188–R195. doi:doi: 10.1093/hmg/ddl190.

Portenko, L. &. K., & Wunderlich (1984). Parus major L. – Atlas Verbr. palaearkt. *Vögel*, *12*, 1–9.

Prinsen, R., Rossoni, A., Gredler, B., Bieber, A., Bagnato, A., & Strillacci, M. (2017). A genome wide association study between CNVs and quantitative traits in Brown Swiss cattle. *Livest. Sci.*, *202*, 7–12.

Prunier, J., Caron, S., & MacKay, J. (2017). CNVs into the wild: screening the genomes of conifer trees (Picea spp.) reveals fewer gene copy number variations in hybrids and links to adaptation. *BMC Genomics*, *18*, 97. doi:doi: 10.1186/s12864-016-3458-8.

Puig, M., Caceres, M., & Ruiz, A. (2004). Silencing of a gene adjacent to the breakpoint of a widespread Drosophila inversion by a transposon-induced antisense RNA. *Proc. Natl. Acad. Sci.*, *101*, 9013–9018. doi:doi: 10.1073/pnas.0403090101.

Puig, M., Casillas, S., Villatoro, S., & Cáceres, M. (2015). Human inversions and their functional consequences. *Briefings in Functional Genomics*, *14*, 369–379. doi:doi: 10.1093/bfgp/elv020.

Purcell et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, *81*, 559–75.

Qian, W., & Zhang, J. (2014). Genomic evidence for adaptation by gene duplication. *Genome Res.*, *24*, 1356–1362.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*, 841–2.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria.

Redon et al. (2006a). Global variation in copy number in the human genome. *Nature*, *444*, 444–54.

Redon, R. et al. (2006b). Global variation in copy number in the human genome. *Nature*, *444*, 444–54. doi:doi: 10.1038/nature05329.

Robinson et al. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*, 139–40.

Rozycka, M., Lu, Y.-J., Brown, R. A., Lau, M. R., Shipley, J. M., & Fry, M. J. (1998). cDNA Cloning of a Third Human C2-Domain-Containing Class II Phosphoinositide 3-Kinase, PI3K-C2γ, and Chromosomal Assignment of This Gene (PIK3C2G) to 12p12. *Genomics*, *54*, 569–574. doi:doi: 10.1006/geno.1998.5621.

Ruiz-Herrera, A., Castresana, J., & Robinson, T. J. (2006). Is mammalian chromosomal evolution driven by regions of genome fragility? *Genome Biol.*, *7*, R115.

Saladin, V., Bonfils, D., Binz, T., & Richner, H. (2003). Isolation and characterization of 16 microsatellite loci in the Great Tit Parus major. *Molecular Ecology Notes*, *3*, 520–522. doi:doi: 10.1046/j.1471-8286.2003.00498.x.

Salleron, L., Magistrelli, G., Mary, C., Fischer, N., Bairoch, A., & Lane, L. (2014). DERA is the human deoxyribose phosphate aldolase and is involved in stress response. *Biochim. Biophys. Acta - Mol. Cell Res.*, *1843*, 2913–2925. doi:doi: 10.1016/j.bbamcr.2014.09.007.

Sandler, L., & Golic, K. (1985). Segregation distortion in drosophila. *Trends in Genetics*, *1*, 181–185. doi:doi: 10.1016/0168-9525(85)90074-5.

Sandler, L., & Novitski, E. (1957). Meiotic Drive as an Evolutionary Force. *The American Naturalist*, *91*, 105–110. doi:doi: 10.1086/281969.

Sankoff, D. (2009). The where and wherefore of evolutionary breakpoints. *J. Biol.*, *8*, 66. doi:doi: 10.1186/jbiol162.

Schaper, S. V., Rueda, C., Sharp, P. J., Dawson, A., & Visser, M. E. (2011). Spring phenology does not affect timing of reproduction in the great tit (parus major). *Journal of Experimental Biology*, *214*, 3664–3671. doi:doi: 10.1242/jeb.059543.

Schmidt, J. M., Good, R. T., Appleton, B., Sherrard, J., Raymant, G. C., Bogwitz,

M. R., Martin, J., Daborn, P. J., Goddard, M. E., Batterham, P., & Robin, C. (2010). Copy Number Variation and Transposable Elements Feature in Recent, Ongoing Adaptation at the Cyp6g1 Locus. *PLoS Genet.*, *6*, e1000998.

Schrider, D. R., & Hahn, M. W. (2010). Lower Linkage Disequilibrium at CNVs is due to Both Recurrent Mutation and Transposing Duplications. *Mol. Biol. Evol.*, *27*, 103–111. doi:doi: 10.1093/molbev/msp210.

Schrider, D. R., Navarro, F. C. P., Galante, P. A. F., Parmigiani, R. B., Camargo, A. A., Hahn, M. W., & de Souza, S. J. (2013). Gene Copy-Number Polymorphism Caused by Retrotransposition in Humans. *PLoS Genet.*, *9*, e1003242.

Sehn, J. K. (2015). Insertions and Deletions (Indels). In *Clinical Genomics* (pp. 129–150). Elsevier. doi:doi: 10.1016/B978-0-12-404748-8.00009-5.

Sen, S. K., Han, K., Wang, J., Lee, J., Wang, H., Callinan, P. A., Dyer, M., Cordaux, R., Liang, P., & Batzer, M. A. (2006). Human Genomic Deletions Mediated by Recombination between Alu Elements. *The American Journal of Human Genetics*, *79*, 41–53. doi:doi: 10.1086/504600.

Seong, H.-A., Jung, H., Choi, H.-S., Kim, K.-T., & Ha, H. (2005). Regulation of Transforming Growth Factor-β Signaling and PDK1 Kinase Activity by Physical Interaction between PDK1 and Serine-Threonine Kinase Receptor-associated Protein. *J. Biol. Chem.*, *280*, 42897–42908. doi:doi: 10.1074/jbc.M507539200.

Seutin, G., White, B. N., & Boag, P. T. (1991). Preservation of avian blood and tissue samples for DNA analyses. *Can. J. Zool.*, *69*, 82–90.

Shao, H., Ganesamoorthy, D., Duarte, T., Cao, M. D., Hoggart, C. J., & Coin, L. J. M. (2018). npInv: accurate detection and genotyping of inversions using long read sub-alignment. *BMC Bioinformatics*, *19*, 261. doi:doi: 10.1186/s12859-018-2252-9.

Sharakhov, I. V., White, B. J., Sharakhova, M. V., Kayondo, J., Lobo, N. F., Santolamazza, F., della Torre, A., Simard, F., Collins, F. H., & Besansky, N. J. (2006). Breakpoint structure reveals the unique origin of an interspecific chromosomal inversion (2La) in the Anopheles gambiae complex. *Proc. Natl. Acad. Sci.*, *103*, 6258–6262. doi:doi: 10.1073/pnas.0509683103.

Sharp, A. J. et al. (2005). Segmental Duplications and Copy-Number Variation in the Human Genome. *Am. J. Hum. Genet.*, *77*, 78–88.

Shin, J.-H., Blay, S., McNeney, B., & Graham, J. (2006). Ldheatmap: An r function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *J Stat Soft*, *16*.

Shlien, A., & Malkin, D. (2009). Copy number variations and cancer. *Genome Medicine*, *1*, 62. doi:doi: 10.1186/gm62.

Shlien, A., & Malkin, D. (2010). Copy number variations and cancer susceptibility. *Current Opinion in Oncology*, *22*, 55–63. doi:doi: 10.1097/cco.0b013e328333dca4.

da Silva et al. (2018). CNVs are associated with genomic architecture in a songbird. *BMC Genomics*, *19*, 195.

da Silva, V. H., Laine, V. N., Bosse, M., Spurgin, L. G., Derks, M. F. L., van Oers, K., Dibbits, B., Slate, J., Crooijmans, R. P. M. A., Visser, M. E., & Groenen, M. A. M. (2019). The genomic complexity of a large inversion in great tits. *Genome Biology and Evolution*, . doi:doi: 10.1093/gbe/evz106.

da Silva, V. H., Regitano, L. C. d. A., Geistlinger, L., Pértille, F., Giachetto, P. F., Brassaloti, R. A., Morosini, N. S., Zimmer, R., & Coutinho, L. L. (2016). Genome-Wide Detection of CNVs and Their Association with Meat Tenderness in Nelore Cattle. *PLOS ONE*, *11*, e0157711. doi:doi: 10.1371/journal.pone.0157711.

Simam, J., Rono, M., Ngoi, J., Nyonda, M., Mok, S., Marsh, K., Bozdech, Z., & Mackinnon, M. (2018). Gene copy number variation in natural populations of Plasmodium falciparum in Eastern Africa. *BMC Genomics*, *19*, 372. doi:doi: 10.1186/s12864-018-4689-7.

Simon, M. C., & Keith, B. (2008). The role of oxygen availability in embryonic development and stem cell function. *Nature Reviews Molecular Cell Biology*, *9*, 285–296. doi:doi: 10.1038/nrm2354.

Singhal, S. et al. (2015). Stable recombination hotspots in birds. *Science (80-.).*, *350*, 928–932.

Skaletsky, H. et al. (2003). The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*, *423*, 825–837. doi:doi: 10.1038/nature01722.

Skinner, B. M., Al Mutery, A., Smith, D., Völker, M., Hojjat, N., Raja, S., Trim, S., Houde, P., Boecklen, W. J., & Griffin, D. K. (2014). Global patterns of apparent copy number variation in birds revealed by cross-species comparative genomic hybridization. *Chromosom. Res.*, *22*, 59–70.

Smit, A. F. A., Hubley, R., & Green, P. (2013-2015). RepeatMasker Open-4.0.

Smith, H. G., Kallander, H., & Nilsson, J.-A. (1989). The Trade-Off Between Offspring Number and Quality in the Great Tit Parus major. *The Journal of Animal Ecology*, *58*, 383. doi:doi: 10.2307/4837.

Stevison, L. S., Hoehn, K. B., & Noor, M. A. F. (2011). Effects of Inversions on Within- and Between-Species Recombination and Divergence. *Genome Biol. Evol.*, *3*, 830–841. doi:doi: 10.1093/gbe/evr081.

Stewart, J. M., & Levy, H. M. (1970). The Role of the Calcium-Troponin-Tropomyosin Complex in the Activation of Contraction. *J. Biol. Chem.*, *245*, 5764–5772.

Stolle, E., Pracana, R., Howard, P., Paris, C. I., Brown, S. J., Castillo-Carrillo, C., Rossiter, S. J., & Wurm, Y. (2018). Degenerative expansion of a young supergene. *Molecular Biology and Evolution*, . doi:doi: 10.1093/molbev/msy236.

Sturtevant, A. H. (1921). A Case of Rearrangement of Genes in Drosophila. *Proceedings of the National Academy of Sciences*, *7*, 235–237. doi:doi: 10.1073/pnas.7.8.235.

Subirana, I., Diaz-Uriarte, R., Lucas, G., & Gonzalez, J. R. (2011). CNVassoc: Association analysis of CNV data using R. *BMC Med Genomics*, *4*, 47.

Sutter, A., & Lindholm, A. K. (2016). Meiotic drive changes sperm precedence patterns in house mice: potential for male alternative mating tactics? *BMC Evolutionary Biology*, *16*, 133. doi:doi: 10.1186/s12862-016-0710-4.

Šťovíček, V., Váchová, L., Begany, M., Wilkinson, D., & Palková, Z. (2014). Global changes in gene expression associated with phenotypic switching of wild yeast. *BMC Genomics*, *15*, 136.

Tattini, L., D'Aurizio, R., & Magi, A. (2015). Detection of genomic structural variants from next-generation sequencing data. *Frontiers in Bioengineering and Biotechnology*, *3*. doi:doi: 10.3389/fbioe.2015.00092.

Thackeray, S. J. et al. (2016). Phenological sensitivity to climate across taxa and trophic levels. *Nature*, *535*, 241–245. doi:doi: 10.1038/nature18608.

Thompson, M. J., & Jiggins, C. D. (2014). Supergenes and their role in evolution. *Heredity*, *113*, 1–8. doi:doi: 10.1038/hdy.2014.20.

Trask, B. (1998). Large multi-chromosomal duplications encompass many members of the olfactory receptor gene family in the human genome. *Human Molecular Genetics*, *7*, 2007–2020. doi:doi: 10.1093/hmg/7.13.2007.

Tukey, J. W. (1949). Comparing Individual Means in the Analysis of Variance. *Biometrics*, *5*, 99. doi:doi: 10.2307/3001913.

Tuttle, E. M., Bergland, A. O., Korody, M. L., Brewer, M. S., Newhouse, D. J., Minx, P., Stager, M., Betuel, A., Cheviron, Z. A., Warren, W. C., Gonser, R. A., & Balakrishnan, C. N. (2016). Divergence and Functional Degradation of a Sex Chromosome-like Supergene. *Curr. Biol.*, *26*, 344–350. doi:doi: 10.1016/j.cub.2015.11.069.

Untergasser, A., Nijveen, H., Rao, X., Bisseling, T., Geurts, R., & Leunissen, J. A. M. (2007). Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Res.*, *35*, W71–4.

Upadhyay, M., da Silva, V. H., Megens, H.-J., Visker, M. H. P. W., Ajmone-Marsan, P., Bâlteanu, V. A., Dunner, S., Garcia, J. F., Ginja, C., Kantanen, J., Groenen,

M. A. M., & Crooijmans, R. P. M. A. (2017). Distribution and Functionality of Copy Number Variation across European Cattle Populations. *Frontiers in Genetics*, *8*. doi:doi: 10.3389/fgene.2017.00108.

Veltman, J. A., & Brunner, H. G. (2012). De novo mutations in human genetic disease. *Nat. Rev. Genet.*, *13*, 565–575.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics*, *101*, 5–22. doi:doi: 10.1016/j.ajhg.2017.06.005.

Visser, M. E., & Both, C. (2005). Shifts in phenology due to global climate change: the need for a yardstick. *Proceedings of the Royal Society B: Biological Sciences*, *272*, 2561–2569. doi:doi: 10.1098/rspb.2005.3356.

Visser, M. E., Holleman, L. J. M., & Gienapp, P. (2006). Shifts in caterpillar biomass phenology due to climate change and its impact on the breeding biology of an insectivorous bird. *Oecologia*, *147*, 164–172. doi:doi: 10.1007/s00442-005-0299-6.

Visser, M. E., te Marvelde, L., & Lof, M. E. (2012). Adaptive phenological mismatches of birds and their food in a warming world. *Journal of Ornithology*, *153*, 75–84. doi:doi: 10.1007/s10336-011-0770-6.

Visser, M. E., Noordwijk, A. J. v., Tinbergen, J. M., & Lessells, C. M. (1998). Warmer springs lead to mistimed reproduction in great tits (Parus major). *Proc. R. Soc. B Biol. Sci.*, *265*, 1867–1870.

Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013). Detecting Natural Selection in Genomic Data. *Annu. Rev. Genet.*, *47*, 97–120.

Volker, M., Backstrom, N., Skinner, B. M., Langley, E. J., Bunzey, S. K., Ellegren, H., & Griffin, D. K. (). Copy number variation, chromosome rearrangement, and their association with recombination during avian evolution. *Genome Res.*, (pp. 503–511).

Vu, V., Verster, A. J., Schertzberg, M., Chuluunbaatar, T., Spensley, M., Pajkic, D., Hart, G. T., Moffat, J., & Fraser, A. G. (2015). Natural Variation in Gene Expression Modulates the Severity of Mutant Phenotypes. *Cell*, *162*, 391–402.

Walsh, M. P. (1994). Calmodulin and the regulation of smooth muscle contraction. *Mol. Cell. Biochem.*, *135*, 21–41.

Wang et al. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res*, *17*, 1665–74.

Wang, J., Ross, K. G., & Keller, L. (2008). Genome-Wide Expression Patterns and the Genetic Architecture of a Fundamental Social Trait. *PLoS Genetics*, *4*, e1000127. doi:doi: 10.1371/journal.pgen.1000127.

Wang, J., Wurm, Y., Nipitwattanaphon, M., Riba-Grognuz, O., Huang, Y.-C., Shoemaker, D., & Keller, L. (2013). A Y-like social chromosome causes alternative colony organization in fire ants. *Nature*, *493*, 664–668. doi:doi: 10.1038/nature11832.

Warren, W. C. et al. (2010). The genome of a songbird. *Nature*, *464*, 757–762.

Weischenfeldt, J., Symmons, O., Spitz, F., & Korbel, J. O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.*, *14*, 125–138.

Wellenreuther, M., Mérot, C., Berdan, E., & Bernatchez, L. (2019). Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification. *Molecular Ecology*, *28*, 1203–1209. doi:doi: 10.1111/mec.15066.

Wilkin, T. A., Perrins, C. M., & Sheldon, B. C. (2007). The use of GIS in estimating spatial variation in habitat quality: a case study of lay-date in the Great Tit Parus major. *Ibis (Lond. 1859).*, *149*, 110–118. doi:doi: 10.1111/j.1474-919X.2007.00757.x.

Williams, R. B., Chan, E. K., Cowley, M. J., & Little, P. F. (2007). The influence of genetic variation on gene expression. *Genome Res.*, *17*, 1707–1716.

Winchester, L., Yau, C., & Ragoussis, J. (2009). Comparing CNV detection methods for SNP arrays. *Briefings in functional genomics & proteomics*, *8*, 353–66. doi:doi: 10.1093/bfgp/elp017.

Wray, N. R. (2005). Allele Frequencies and the r2 Measure of Linkage Disequilibrium: Impact on Design and Interpretation of Association Studies. *Twin Res. Hum. Genet.*, *8*, 87–94. doi:doi: 10.1375/twin.8.2.87.

Wright, D., Boije, H., Meadows, J. R. S., Bed'hom, B., Gourichon, D., Vieaud, A., Tixier-Boichard, M., Rubin, C.-J., Imsland, F., Hallböök, F., & Andersson, L. (2009). Copy Number Variation in Intron 1 of SOX5 Causes the Pea-comb Phenotype in Chickens. *PLoS Genet.*, *5*, e1000512.

Wright, E. (2016). Using decipher v2.0 to analyze big biological sequence data in r. *The R Journal*, *8*.

Wright, S. (1931). Evolution in Mendelian Populations. *Genetics*, *16*, 97–159.

Wu, C. I., Lyttle, T. W., Wu, M. L., & Lin, G. F. (1988). Association between a satellite DNA sequence and the Responder of Segregation Distorter in D. melanogaster. *Cell*, *54*, 179–89.

Xie, C., & Tammi, M. T. (2009a). CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics*, *10*, 80. doi:doi: 10.1186/1471-2105-10-80.

Xie, C., & Tammi, M. T. (2009b). CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics*, *10*, 80. doi:doi: 10.1186/1471-2105-10-80.

Xu, L., Cole, J. B., Bickhart, D. M., Hou, Y., Song, J., VanRaden, P. M., Sonstegard, T. S., Van Tassell, C. P., & Liu, G. E. (2014). Genome wide CNV analysis reveals additional variants associated with milk production traits in Holsteins. *BMC genomics*, *15*, 683. doi:doi: 10.1186/1471-2164-15-683.

Yalcin, B., Wong, K., Bhomra, A., Goodson, M., Keane, T. M., Adams, D. J., & Flint, J. (2012). The fine-scale architecture of structural variants in 17 mouse genomes. *Genome Biol.*, *13*, R18.

Yamada, K., Nomura, N., Yamano, A., Yamada, Y., & Wakamatsu, N. (2012). Identification and characterization of splicing variants of PLEKHA5 (Plekha5) during brain development. *Gene*, *492*, 270–275. doi:doi: 10.1016/j.gene.2011.10.018.

Yau, C., & Holmes, C. C. (2008). CNV discovery using SNP genotyping arrays. *Cytogenetic and genome research*, *123*, 307–12. doi:doi: 10.1159/000184722.

Ye, K., Hall, G., & Ning, Z. (2016). Structural Variation Detection from Next Generation Sequencing. *Journal of Next Generation Sequencing & Applications*, *01*. doi:doi: 10.4172/2469-9853.S1-007.

Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *Omi. A J. Integr. Biol.*, *16*, 284–287.

Zarrei, M., MacDonald, J. R., Merico, D., & Scherer, S. W. (2015). A copy number variation map of the human genome. *Nature Reviews Genetics*, *16*, 172–183. doi:doi: 10.1038/nrg3871.

Zaykin, D. V., Pudovkin, A., & Weir, B. S. (2008). Correlation-Based Inference for Linkage Disequilibrium With Multiple Alleles. *Genetics*, *180*, 533–545.

Zemanova, M. A., Perotto-Baldivieso, H. L., Dickins, E. L., Gill, A. B., Leonard, J. P., & Wester, D. B. (2017). Impact of deforestation on habitat connectivity thresholds for large carnivores in tropical forests. *Ecological Processes*, *6*, 21. doi:doi: 10.1186/s13717-017-0089-1.

Zhang, F., Khajavi, M., Connolly, A. M., Towne, C. F., Batish, S. D., & Lupski, J. R. (2009). The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nature Genetics*, *41*, 849–853. doi:doi: 10.1038/ng.399.

Zhang, G. et al. (2014a). Comparative genomics reveals insights into avian genome evolution and adaptation. *Science (80-.).*, *346*, 1311–1320.

Zhang, H., & Freudenreich, C. H. (2007). An AT-Rich Sequence in Human Common Fragile Site FRA16D Causes Fork Stalling and Chromosome Breakage in S. cerevisiae. *Mol. Cell*, *27*, 367–379.

Zhang, X., Du, R., Li, S., Zhang, F., Jin, L., & Wang, H. (2014b). Evaluation of copy number variation detection for a SNP array platform. *BMC bioinformatics*, *15*, 50. doi:doi: 10.1186/1471-2105-15-50.

Zhang, Z., Schwartz, S., Wagner, L., & Miller, W. (2000). A Greedy Algorithm for Aligning DNA Sequences. *J. Comput. Biol.*, *7*, 203–214.

Zhao, M., Wang, Q., Wang, Q., Jia, P., & Zhao, Z. (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*, *14*, S1. doi:doi: 10.1186/1471-2105-14-S11-S1.

Zhao, W., Ma, N., Wang, S., Mo, Y., Zhang, Z., Huang, G., Midorikawa, K., Hiraku, Y., Oikawa, S., Murata, M., & Takeuchi, K. (2017). RERG suppresses cell proliferation, migration and angiogenesis through ERK/NF-κB signaling pathway in nasopharyngeal carcinoma. *J. Exp. Clin. Cancer Res.*, *36*, 88. doi:doi: 10.1186/s13046-017-0554-9.

Zhao, X., Emery, S. B., Myers, B., Kidd, J. M., & Mills, R. E. (2016). Resolving complex structural genomic rearrangements using a randomized approach. *Genome Biol.*, *17*, 126.

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, *28*, 3326–3328. doi:doi: 10.1093/bioinformatics/bts606.

Ziyatdinov, A., Vázquez-Santiago, M., Brunel, H., Martinez-Perez, A., Aschard, H., & Soria, J. M. (2018). lme4qtl: linear mixed models with flexible covariance structure for genetic studies of related individuals. *BMC Bioinformatics*, *19*, 68. doi:doi: 10.1186/s12859-018-2057-x.

Summary

Knowledge on the evolutionary ecology of wild species allows insights on how ongoing climate change is affecting the biodiversity on our planet. A key effect of climate change is that species at different trophic levels shift their phenology at a significantly different pace. These differential shifts lead to selection for earlier timing at the species higher up in the food chain, such as insectivorous birds. The great tit (*Parus major*) is an insectivorous songbird that has been used as a model species for ecology and evolution. To understand how this species may genetically respond to the selection on timing we need to understand the degree and nature of its genomic variation. Recently, the reference great tit genome has been developed and explored using a variety of high-throughput platforms, allowing a detailed characterization of genomic structural variations as presented in this thesis that goes beyond the already explored SNP variation.

Chapter 2 presents the results of a genome-wide copy number variation (CNV) detection strategy based on a species-specific high-density SNP array, which generated a CNV-map for the great tit genome, which was partially validated by quantitative PCR (qPCR). By using the available family structure (i.e. mother-offspring) the inheritance patterns were analyzed showing that in particular larger CNVs are inherited as expected by Mendel's law. CNVs are expected to follow Mendelian inheritance, and therefore the general deviation from the Mendelian inheritance in shorter CNVs show that CNV confidence is length-dependent in our data-set (i.e. shorter CNVs identified in mother have a higher chance be missed by our method in the offspring). However, as qPCR showed a high validation rate for distinct CNV lengths, it is likely that CNVs identified in this study have a higher rate of false negatives than false positives. CNVs are frequency-dependently associated with a number of genomic features that underlie their formation. Overlap with genomic features such as CpG sites and transcription start sites (TSSs) confirmed a non-random distribution of the identified CNVs. Moreover, CNVs may have a crucial role in evolution as they are enriched around evolutionary breakpoints in different species, including in the great tit genome.

By knowing the limitations of the CNV map developed in **Chapter 2**, in **Chapter 3** CNVs were used to better understand the genetic contribution of structural variations in seasonal timing. Breeding timing can be studied by the egg-laying date in a breeding season, which is a relevant fitness-related seasonal measurement commonly used as a proxy for timing. Because great tit CNVs are likely to contain a high rate of false negatives, several CNV regions (CNVRs) may have their frequency underestimated. To overcome this problem to some extent, a hybrid CNV-GWAS approach was used in which CNVs and raw signal intensities (i.e. log R ratio - LRR) were jointly associated with egg-laying date measurements. As expected, egg-laying dates are largely polygenic and therefore not strongly associated with any CNV in particular. However, suggestively associated regions harboring genes related to circadian clock and in mammals to pregnancy, highlight relevant candidates for future research.

Methods to associate CNVs with quantitative phenotypes are not extensively developed and documented in the literature. In **Chapter 4** we therefore present a R/Bioconductor package that integrates the methods used in **Chapter 3** to allow higher reproducibility of our results through an open-source CNV-GWAS software that is freely available for the scientific community. Moreover, CNVRanger also implements a list of functions to allow CNV summarization and association with gene expression.

In **Chapter 5** all somatic chromosomes were further explored by PCA, F_{ST} and heterozygosity. A large inversion located at Chromosome 1A, which overlaps 90% of its size, was identified and explored in detail. In agreement with other inversions reported in the literature, this inversion is structurally complex and has signals of degeneration expansion, which is common in young supergenes, such as a high incidence of CNVs close to the breakpoints. The inversion is widespread across different European populations at a frequency of $\approx$5%. Furthermore, the inversion can be divided into at least two different haplogroups, which are distinguishable by their completely different genotype distribution around the center of the chromosome. Finally, the lack of homozygotes among all birds (>2,000) explored in **Chapter 5** suggested a possible recessive lethal effect for this inversion that was then investigated in **Chapter 6**.

Chapter 6 explores the inheritance patterns and fitness effects of the inversion described in **Chapter 5**. Offspring ratios as well as number of hatched eggs in carrier-by-carrier mating pairs support that the inversion is indeed lethal in homozygous state. Thus, a deviation from Mendel's law (e.g. segregation distortion) or a fitness advantage may exist that result in maintaining such a lethal variant (i.e. balancing selection). Segregation distortion was explored by examining the difference between expected and observed offspring ratios in carrier-by-normal pairs. In pairs where the male is the inversion carrier, the inversion is inherited twice more often than the normal variant, suggesting a male related segregation distortion. However, a drift-selection simulation indicates that a fitness disadvantage should be present in carriers to explain both the observed inversion frequency and the segregation distortion. None of the fitness components explored here is associated with the inversion, which suggests that the fitness component affected by this inversion might not be captured by the experimental design used in this thesis.

Structural variants can reveal important genes in speciation and intraspecific selection, but they can be extremely challenging to explore (**Chapters 2 and 5**). Although clearly polygenic, the association of CNVs with seasonal measurements should be further investigated, mainly in genes within suggestive CNVs (**Chapter**

3). Future studies should also clarify the molecular mechanisms maintaining the segregation distortion as well as which fitness-related measurements can explain the expected fitness disadvantage of such an exceptional structural rearrangement on Chromosome 1A of great tit.

Samenvatting

Om inzicht te krijgen in hoe klimaatsverandering de biodiversiteit op onze aarde beïnvloedt moeten we de evolutionaire ecologie van wilde populaties bestuderen. Eén van de belangrijkste gevolgen van klimaatsverandering is dat soorten hun fenotype aanpassen, maar dat soorten op de verschillende trofische niveaus dat met een verschillende snelheid doen. Dit verschil in snelheid leidt tot selectie voor vroegere seizoenstiming bij soorten hoger in de voedsel keten, zoals insecten etende vogels. De koolmees (*Parus major*) is zo'n insecten etende zangvogel en is een modelsoort voor ecologische en evolutionaire studies. Om te begrijpen hoe soorten genetisch reageren op selectie van timing moeten we begrijpen wat de mate en soort genomische variatie is die hieraan is gekoppeld. Recent is het referentie genoom van de koolmees beschreven waarbij gebruik gemaakt is van een variëteit aan geavanceerde platformen. Dit heeft het mogelijk gemaakt om een gedetailleerde karakterisering van genomische structurele variatie in koolmees te bestuderen, zoals in dit proefschrift gedaan wordt, die verder gaat dan de reeds onderzochte SNP variatie.

In **Hoofdstuk 2** worden de copy nummer variatie (CNV) resultaten gepresenteerd van een genoomwijde detectie strategie gebaseerd op een hoge dichtheid SNP array voor de koolmees. Dit heeft geresulteerd in een CNV kaart van het koolmeesgenoom die gevalideerd is met behulp van een kwantitatieve PCR (qPCR). Door gebruik te maken van de familiestructuur (o.a. moeder-nakomeling) kon de overerving van CNVs worden bestudeerd waarbij hoofdzakelijk grote CNVs overerven in overeenstemming met de wet van Mendel. Omdat alle CNVs de Mendeliaanse overerving zouden moeten volgen geeft de afwijking hiervan bij kleinere CNVs aan dat de betrouwbaarheid van het detecteren van CNV lengte afhankelijk is in onze dataset (dus dat met onze methode kleinere CNVs gevonden in de moeders een hogere kans hebben om gemist te worden in de nakomeling). Alhoewel de qPCR resultaten een hoge mate van validatie lieten zien, is het aannemelijk dat de gevonden CNVs in deze studie een hoger percentage vals negatieve dan vals positieve laten zien.. Overlap met deze kenmerken zoals CpG eilanden en transcriptie startplaatsen (TSSs) bevestigt een niet random distributie op het genoom van de gevonden CNVs. Ook kunnen CNVs een belangrijke evolutionaire rol spelen omdat deze verhoogd aanwezig zijn in de buurt van evolutionaire breekpunten in het genoom van verschillende soorten waaronder ook de koolmees.

In **Hoofdstuk 3** wordt de in **Hoofdstuk 2** ontwikkelde CNV kaart gebruikt, met in achtneming van de beperkingen van deze kaart, om de genetische bijdrage aan de variatie in seizoensgebonden timing bij koolmezen beter te begrijpen. Doordat de gevonden CNVs in het koolmeesgenoom mogelijk een hoge mate van vals negatieve bevatten zijn de CNV regio's (CNVRs) mogelijk in hun frequentie onderschat. Om dit probleem te omzeilen is een hybride CNV-GWAS procedure toegepast waarbij de CNV en de ruwe data (Log R ratio - LRR) samen geassocieerd werden met de ei-legdatum. Zoals verwacht is de ei-legdatum een eigenschap die door een groot aantal genen bepaald wordt en is dan ook niet sterk geassocieerd met een specifieke CNV. De gebieden met een suggestieve associatie bevatten genen gerelateerd aan de biologische klok en zwangerschap bij zoogdieren. Dit zijn daarmee mogelijk kandidaten voor toekomstig onderzoek.

De methoden om CNVs en kwantitatieve kenmerken te associëren zijn beperkt of slecht gedocumenteerd in de literatuur. **In Hoofdstuk 4** presenteren we een R/Bioconductor pakket om de methoden die gebruikt zijn in **Hoofdstuk 3** te integreren om zo tot een betere reproduceerbaarheid van de resultaten te komen. Dit open-source CNV-GWAS software pakket, genaamd CNVRanger, is vrij toegankelijk. CNV opsomming en associatie met genexpressie is een van de geïmplementeerde functies van CNVRanger.

In **Hoofdstuk 5** worden de chromosomen verder bestudeerd met behulp van PCA, FST en heterozygotie. Een grote inversie op chromosoom 1A, die 90% van het chromosoom omvat, is geïdentificeerd en beschreven. In overeenstemming met de inversies beschreven in de literatuur is deze inversie structureel complex en bevat signalen van degeneratieve expansie, iets dat gebruikelijk is bij jonge supergenen, zoals een verhoogde incidentie van CNVs dicht bij de breekpunten. De inversie komt verspreid voor binnen de Europese populaties met een frequentie van ongeveer 5%. Bovendien kan de inversie verdeeld worden in tenminste twee verschillende haplogroepen, die te onderscheiden zijn op basis van een compleet verschillende genotype distributie rond het midden van het chromosoom. Binnen alle geteste vogels (>2000) werden geen homozygoten aangetroffen. Dit suggereert een mogelijk recessief lethaal effect voor deze inversie.

In **Hoofdstuk 6** worden het overervingspatroon en de fitness effecten van de inversie zoals beschreven in **Hoofdstuk 5** verder onderzocht. Het percentage eieren dat uitkomt bij een kruising tussen twee dragers, is duidelijk verlaagd wat aangeeft dat de inversie in homozygote staat lethaal is. Voor een lethale variant om in de populatie te blijven bestaan moeten de heterozygoten een fitnessvoordelel hebben wat dan leidt tot gebalanceerde selectie. Segregatie vervorming werd verder bestudeerd door het verschil te bepalen in verwacht versus geobserveerde ratio van nakomelingen met en zonder de inversie in kruisingen tussen een drager en een normale wild type ouder. In de paringen waar het mannetje drager is van de inversie werd de inversie tweemaal zo vaak overgeërfd dan de normale variant. Dit suggereert een mannelijk gerelateerde segregatie vervorming. Een gesimuleerd drift-selectie scenario geeft aan dat er een fitness nadeel aanwezig moet zijn in dragers om zowel de geobserveerde inversie frequentie als de segregatie vervorming te verklaren. Maar geen van de fitness componenten die bestudeerd zijn, is geassocieerd met de inversie. Structurele varianten kunnen genen bevatten die belangrijk zijn voor soortvorming en intraspecifieke selectie (**Hoofdstuk 2 en 5**). De associatie van CNVs en de gekoppelde genen met seizoensmetingen, is duidelijk polygeen (**Hoofdstuk 3**). Toekomstige studies zijn noodzakelijk om meer inzicht te krijgen in het moleculair mechanisme verantwoordelijk voor de waargenomen segregatie vervorming en de reden voor het verwachte fitness nadeel van deze enorme structurele herschikking op chromosoom 1A van koolmees.

Curriculum vitae

About the author

Vinicius was born on 31st March 1988 in the north of Paraná state of Brazil, in a city with an indigenous name, *Apucarana* (meaning 'like a big forest'). Vinicius was always interested in computers and science fiction, but the widespread livestock production in Brazil had a strong influence in his professional aspirations. Thus, after reading about animal breeding, and how the rapid evolving molecular tools could be used for that, he decided to apply for a bachelor's degree in Animal Science. In 2007, he was approved as undergraduate student at the State University of Londrina (UEL), where he had received different scholarships to pursue extracurricular internal and external trainees on genome research. Vinicius obtained the title of animal scientist in 2011, in a public graduation ceremony where he was the representative speaker. His undergraduate work on a nucleotide change associated with carcass traits in chicken, as a trainee at the Brazilian Agricultural Research Corporation (EMBRAPA), received in 2012 a student travel award bursary from the XXIV World's Poultry Congress that was held in Salvador, Brazil.

In 2012, he was accepted by the MSc programme in Animal Science and Pastures provided by the Luiz de Queiroz College of Agriculture (ESALQ), which a campus of the University of São Paulo (USP). There, he developed his MSc project on the structural variants in a Brazilian cattle genome at the laboratory of Animal Biotechnology, in which he also collaborated with molecular studies involving distinct livestock species. Thus, Vinicius developed a 'sandwich' extension to his MSc project within a six-month stay at the Institute of Bioinformatics, which is part of the Ludwig Maximilian University (LMU) of Munich, Germany. His MSc project received a travel award bursary from the 34th International Society of Animal Genetics (ISAG 2014) that was held in Xi´an, China.

In 2015, he was selected by the European Graduate School in Animal Breeding and Genetics (EGS-ABG) joint PhD programme. His PhD project was a joint effort between (i) Wageningen University & Research, (ii) Swedish University of Agricultural Sciences (SLU) and (iii) Netherlands Institute of Ecology (NIOO), which generated a detailed study on the structural variants in the genome of a model species, the great tit (*Parus major*). The results of his PhD project are presented in this thesis entitled 'Structural variants in the great tit genome and their effect on seasonal timing'.

Peer-reviewed journal publications

1 - **da Silva**, Vinicius, Marcel Ramos, Martien Groenen, Richard Crooijmans, Anna Johansson, Luciana Regitano, Luiz Coutinho, Ralf Zimmer, Levi Waldron, and Ludwig Geistlinger. 2019. *CNVRanger: Association Analysis of CNVs with Gene Expression and Quantitative Phenotypes.* Bioinformatics btz632:1–2.

2 - Pértille, Fábio, Vinicius H. **da Silva**, Anna M. Johansson, Tom Lindström, Dominic Wright, Luiz L. Coutinho, Per Jensen, and Carlos Guerrero-Bosagna. 2019. *Mutation Dynamics of CpG Dinucleotides during a Recent Event of Vertebrate Diversification.* Epigenetics 14(7):685–707.

3 - **da Silva**, Vinicius H., Veronika N. Laine, Mirte Bosse, Lewis G. Spurgin, Martijn F. L. Derks, Kees van Oers, Bert Dibbits, Jon Slate, Richard P. M. A. Crooijmans, Marcel E. Visser, and Martien A. M. Groenen. 2019. *The Genomic Complexity of a Large Inversion in Great Tits.* Genome Biology and Evolution 11(7):1870–81.

4 - **da Silva**, Vinicius H., Veronika N. Laine, Mirte Bosse, Kees van Oers, Bert Dibbits, Marcel E. Visser, Richard P. M. A. Crooijmans, and Martien A. M. Groenen. 2018. *CNVs Are Associated with Genomic Architecture in a Songbird.* BMC Genomics 19(S2):195.

5 - Schurink, Anouk, Vinicius H. **da Silva**, Brandon D. Velie, Bert W. Dibbits, Richard P. M. A. Crooijmans, Liesbeth Franois, Steven Janssens, Anneleen Stinckens, Sarah Blott, Nadine Buys, Gabriella Lindgren, and Bart J. Ducro. 2018. *Copy Number Variations in Friesian Horses and Genetic Risk Factors for Insect Bite Hypersensitivity.* BMC Genetics 19(1):49.

6 - Geistlinger, Ludwig, Vinicius Henrique **da Silva**, Aline Silva Mello Cesar, Polyana Cristine Tizioto, Levi Waldron, Ralf Zimmer, Luciana Correia de Almeida Regitano, and Luiz Lehmann Coutinho. 2018. *Widespread Modulation of Gene Expression by Copy Number Variation in Skeletal Muscle.* Scientific Reports 8(1):1399.

7 - Upadhyay, Maulik, Vinicus H. **da Silva**, Hendrik-Jan Megens, Marleen H. P. W. Visker, Paolo Ajmone-Marsan, Valentin A. Bâlteanu, Susana Dunner, Jose F. Garcia, Catarina Ginja, Juha Kantanen, Martien A. M. Groenen, and Richard P. M. A. Crooijmans. 2017. *Distribution and Functionality of Copy Number Variation across European Cattle Populations.* Frontiers in Genetics 8.

8 - **da Silva**, Vinicius Henrique, Luciana Correia de Almeida Regitano, Ludwig Geistlinger, Fábio Pértille, Poliana Fernanda Giachetto, Ricardo Augusto Brassaloti, Natália Silva Morosini, Ralf Zimmer, and Luiz Lehmann Coutinho. 2016. *Genome-Wide Detection of CNVs and Their Association with Meat Tenderness in*

Nelore Cattle. PLOS ONE 11(6):e0157711.

9 - Pértille, Fábio, Carlos Guerrero-Bosagna, Vinicius Henrique **da Silva**, Clarissa Boschiero, José de Ribamar da Silva Nunes, Mônica Corrêa Ledur, Per Jensen, and Luiz Lehmann Coutinho. 2016. *High-Throughput and Cost-Effective Chicken Genotyping Using Next-Generation Sequencing.* Scientific Reports 6:26929.

Individual training plan

European Graduate School in Animal Breeding and Genetics

Individual training and supervision plan EGS-ABG

First □ Mid-term □ Final X

BASIC INFORMATION	
Name of PhD student	Vinicius Henrique da Silva
First institution	Wageningen University (WU)
Second institution	Swedish University of Agricultural Sciences (SLU)
Principal supervisor (1st)	Dr Richard Crooijmans
Co-supervisor(s) (1st)	Prof. Dr Martien Groenen and Prof. Dr Marcel E. Visser
Principal supervisor (2nd)	Dr Anna Maria Johansson
Co-supervisor(s) (2nd)	Prof. Dr Dirk-Jan de Koning
Project title	Genetics of seasonal timing in the great tit (Parus major)
Date of enrolment	1st of September, 2015
Expected date of submission/defense?	August, 2019/January, 2020

	TRAINING (30 ECTS minimum)		
Mandatory courses*)	**Where/when**	**First (ECTS)**	**Second (ECTS)**
WIAS Introduction Day	Wageningen/October 1, 2015	0.3	0.3
Welcome to EGS-ABG	Uppsala/October 13-17, 2015	1.5	1.5
EGS-ABG Fall Research School 2015:	Uppsala/October 18-23, 2015	2.0	2.0
EGS-ABG Summer Research School 2017: Emerging technologies in animal breeding	Wageningen/ 13-17 February 2017	2.0	2.0
Research Integrity & Ethics in Animal Sciences	Wageningen/June 30 & July 4-5	1.5	1.5
Advanced scientific courses (≥18 ECTS)			
SSB-30806 Systems Biology from Gene to Ecosystem	Wageningen/ October 31 until December 20, 2016	6.0	6.0
Bioconductor for genomic data science	On-line Johns Hopkins/November 2-29, 2015	1.3	1.3
Orientation on mathematical modelling in biology	Wageningen/ February 8-12, 2016	1.5	1.5
Getting started in ASReml	Wageningen/February 2, 2016	0.3	0.3
Phylogenetic analysis using R	Barcelona/ 6-10 March 2017	3.5	1.5
Linear models in animal breeding	Orsa Grönklitt / Jun 17-21, 2018	3.0	3.0
Introduction to programming in R	Uppsala/ April 16-20, 2018	2.0	2.0
Analysis of High Throughput Sequencing Data	Cambridge/ October 22-26, 2018	1.5	1.5
EGS-ABG Summer course Paris	Paris / May 28 to June 1st, 2018	2.0	2.0
Professional Skills support courses (≥6 ECTS)			
Communication with the Media and the General Public	Wageningen/ November 9, 10 & 30, 2015	1.0	1.0
WIAS Course on Essential Skills	Wageningen/ April 26-29, 2016	1.2	1.2

	TRAINING (30 ECTS minimum)		
Scientific Writing	Wageningen/ February 9 to April 19	1.8	1.8
WIAS course High Impact Writing in Science	Wageningen / 29 May 29 - Jun 1 2017	1.3	1.3
Presenting with impact	Wageningen/ December 3-17, 2018	2	2
Total credits (≥30 ECTS)		**35.7**	**33.7**

	DISSEMINATION OF KNOWLEDGE*)		
Teaching/MSc supervision	**Where/When**	**First (ECTS)**	**Second (ECTS)**
Assistant in Genomics course	Wageningen/2017	1.0	1.0
Supervising BsC student	Wageningen/2016-17	0.5	0.5
Assistant in introduction to R course	*Uppsala/2018*		
International conferences (minimum of 3)			
Congress of the European Society for Evolutionary Biology (ESEB 2017)	Groningen / August 2017	1.0	
The Plant and Animal Genome XXV Conference (PAG)	San Diego/14-18 January, 2018	1.0	
ZOOLOGY 2018 (Royal Belgian Zoological Society)	Antwerpen/14-15 December 2018	1.0	
Seminars and workshop (minimum 1)			
NIOO Research Days	Heeze / November 11-12, 2015	0.6	
WIAS day 2016	Wageningen / February 4 2016	0.3	
WGS PhD Workshop Carousel 2016	Wageningen / April 8 2016	0.3	
WIAS day 2017	Wageningen / February 6, 2017	0.3	
NIOO Research Days	Soesterberg / November 8-9 , 2017	0.6	

	DISSEMINATION OF KNOWLEDGE*)		
Presentations (Minimum 4 original presentations of which at least 1 oral)			
The copy number variations at genes related to neuronal functions under selection in great tit - http://edepot.wur.nl/378719	Wageningen / February 4 2016	1.0	
Repetitive nature of genes underlying neurons in great tit	Groningen / August 2017	1.0	
Chromosome-wide inversion in great tit genome	San Diego / January 2018	1.0	
A large and wide-spread inversion is probably lethal in homozygous state in the great tit	Antwerpen / December 2018	1.0	

Acknowledgements

It is hard to me to disentangle my previous personal and professional history from the period dedicated for this thesis. It feels like the end of long journey as a student and the birth of an independent researcher. I have always been more focused on the power of science than the main institutions that actually produce it, i.e. Universities. The inquisitive and self-critic nature of the science was always my way to face cultural and religious dogmas. Therefore, my first acknowledgement goes to science and all past and present scientists that fought/fight for a world in which research and knowledge is the main driver.

I have learned quick enough that our lifetime is limited, and to be able to live most of it close to science and research was necessary to follow an academic path as yearly as possible. Thus, I dedicated myself to improve my skills on molecular research during my Bachelor's degree. However, dedication is usually not enough when the investment in science is scarce. That's when luck can play a role. I was luck enough to find inspiring professors and colleagues during this process, which gave me opportunities and courage to continue close to the scientific research. I said that is hard to disentangle my history because after my MsC degree, I was able to get a PhD position at ESALQ (University of São Paulo). However, before registering as an ESALQ PhD candidate, I received the approval letter for this joint PhD in the European Graduate School in Animal Breeding and Genetics (EGS-ABG), an Erasmus Mundus programme. The joint PhD candidate position in EGS-ABG would involve the collaboration between two Universities; Wageningen University & Research (WUR) and Swedish University of Agricultural Sciences (SLU), and a research institute, Netherlands Institute of Ecology (NIOO). The PhD position was an exciting project on the genomics of the seasonal timing of great tits. At that time, I was extremely happy to be approved in such a competitive process, but I was cautious in accepting the EGS-ABG and deny the ESALQ position without asking for advice. Although my MsC supervisor, Prof. Coutinho, would lose me as a PhD candidate, he was very proud of my achievement and promptly suggested the EGS-ABG position. I am grateful for such a suggestion, which gave me self-confidence to face a big and new challenge abroad.

After these four years abroad, I am the one feeling proud for being part of the EGS-ABG PhD programme, which involved so many outstanding institutions and professionals. Thus, I would like first to thank Erasmus Mundus, which provided funding for the EGS-ABG PhD programme. Also, I would like to express my gratitude to all the members and secretary of the EGS-ABG PhD programme, mainly to Prof. Dr Etienne Verrier from AgroParisTech, which is the EGS-ABG coordinator. Moreover, I would like thank all the three institutions involved in this project. WUR was the first university in my PhD journey in the beautiful city of life sciences, Wageningen, which provided all structure and resources for the development of my project. Also in Wageningen, working part time at Netherlands Institute of Ecology

(NIOO) was a magnifying experience. Thus, I would like to thank NIOO for all the support (and data!) for my project. Last, but not least, I would like to express my gratitude to Swedish University of Agricultural Sciences (SLU), located in Uppsala, which was my second university and provided all the assistance during my time in Sweden.

As my EGS-ABG joint PhD project involved different institutions, I had different promotors and co-promotors playing an cohesive collaboration on my project development. I would like to express my gratitude to Dr Martien Groenen, my promotor at WUR, which was an active collaborator during all these four years being very supportive and accessible. Martien was always sharp, inquisitive and detail oriented, which was essential to my improvement as a researcher. My gratitude also goes to Dr Marcel Visser, which was also my promotor at WUR, but also at NIOO. Marcel was a fundamental support mainly for the ecological aspects in my project. However, he was also extremely supportive during the design and improvement of my entire thesis. I also express my gratitude to Richard, which was my co-promotor and daily advisor at WUR. Richard was always present in the project development as well as a big help on a wide range of different scientific and common issues. Wrong hotel reservation in US? Richard can help! In Sweden I was also granted with helpful supervisors. In SLU, my gratitude goes to Anna, which was my co-promotor and daily advisor during my time in Sweden. Dirk-Jan (DJ), which was my promotor at SLU, also deserves my gratitude for his great general support and for improving my modelling strategies.

On my first day at WUR, Prof. Martien was supportive and showed the whole department to me. Funny enough, he showed the only microwave placed at the coffee corner explaining that some people may use it to warm their food. I could not really understand why just some people would need it. After all, everyone need to warm up their food, right? Well, not if your lunch is bread and cheese! There are a lot of Dutch cultural differences which I was able to cope, understand and even incorporate, but sorry to say that I can't let go my warm lunch! Needless to say tough that I will miss Golda cheese pretty much. Jokes apart, I appreciated my live in the Netherlands and I am grateful for such an experience. Moving to the Netherlands was not easy, but it was greatly supported by the helpful secretary of the Animal Breeding and Genomics (ABG) department, at the WUR. In special, I would like to express my gratitude to Lisette and Ada, which were extremely supportive during my moving to the Netherlands. In addition, I would like to further thanks Lisette for her prompt assistance in a wide range of different issues during my whole period at WUR and even during my moving to Sweden, and the moving back to the Netherlands, during my PhD project.

In addition to all my promotors and co-promotors, some other people also deserve

my acknowledgements for being part of my scientific improvement. Thank you Veronika, which was always involved on my project development and acted as my daily supervisor at NIOO. Mirte for the brilliant insights and nice discussions. Martijn for all the bioinformatic discussions and assistance. Maulik, thank you for very nice scientific and political (why not!?) discussions. Moreover, I will be always grateful for the dinner you and Nehal provided after Mitra's birth. That was really important. Other dinners were also fun! Thank you all other colleagues from WUR, which were always helpful and good companies. Andre, Lim, Maria, Zhou, Langging, Chiara, Hendrik-Jan, Bart, Bert, Ole, Harmen, Piter, Kimberley, Rosilde, Tom, Pascal, Claudia, Kasper, Shuwen and others that colleagues that I sadly forgot to include in here.

Thank you all other colleagues from SLU, which made my stay in Sweden more fun. Juan, Julie, Gabriela, Patricia, Juliette, Elisabeth, Tomas, Cano, Suvi, Christos, Lotta, Joanna, Getinet, Erling, Patrik, Sandrine and many other people for the warm treatment under the freezing Swedish weather (though the summer was brilliant!).

Thank you all other colleagues from NIOO, which were receptive and patient when ecology discussions were on play. Judith, Kees, Barbara, Jip and many other people at NIOO, which directly or indirectly collaborated to this thesis.

This whole time in Europe would have been much harder alone. Luckily enough, I was able to bring my beloved Carolina with me. She is my partner, friend, flower and, since 2016, the mother of my chubby and smart little daughter Mitra that loves to pretend to be a small pig. Roinc roinc chips! Thank you both for making all the challenges and difficulties during this time abroad way smoother. Love you guys! Gratitude to all my uncles and aunts that were always supporting and sometimes visiting! A special thanks to Silvana, which was my friend way before becoming my mother in law. Thank you for your daughter and all the help by coming to Netherlands several times. Finally, I would like to thank my mom, father and brother, which always supported my decisions and provided energy to continue. Also my gratitude should be extended to my grandpa and grandma, which never really understood my career but were nevertheless supportive and proud. Unfortunately, in the last year of my PhD, I lost my father and both my grandpa and grandma. However, I know they would be glad to witness this work done, which I therefore dedicate to them.

The research described in this thesis was supported by the European Commission (within the framework of the Erasmus-Mundus joint doctorate "EGS-ABG") and Wageningen University & Research

The thesis was printed by Digiforce, Proefschriftmaken.nl, De Limiet 26, 4131NC, Vianen, the Netherlands

Cover design by Leonardo Rock and Vinicius Henrique da Silva