Charles University

Faculty of Science

Study programme: Bioinformatics

Branch of study: Bioinformatics

Bc. Zuzana Halenková

Detection of structural variants in genomes of two nightingale species

Detekce strukturálních variant v genomech dvou druhů slavíků

Diploma thesis

Supervisor: RNDr. Radka Reifová, Ph.D.

Consultant: Mgr. Jakub Rídl, Ph.D.

Prague, 2021

Prohlášení

Prohlašuji, že jsem závěrečnou práci zpracovala samostatně a že jsem uvedla všechny použité informační zdroje a literaturu. Analyzovaná vstupní data (NGS sekvence a assembly genomů) byla připravena Mgr. Jakubem Rídlem, Ph.D. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

V Praze, 11. 8. 2021

…….

Zuzana Halenková

Acknowledgement

I would like to thank everyone who supported me in the preparation of this thesis. In the first place, to my supervisor Radka Reifová and to Jakub Rídl for their advice throughout conducting the analyses and writing this thesis. My biggest thanks go to my family, especially to my mum for her never-ending support during the course of my studies.

Analyses reported in this thesis were conducted using the computational resources of the Institute of Molecular Genetics of Czech Academy of Science.

Abstract

Structural variants are mutations in DNA sequence affecting the location, orientation, or the number of copies of regions longer than 50 bp. Although this type of variation has the potential to cause large phenotypic changes, structural variants remain largely understudied compared to other classes of variation (such as single nucleotide polymorphisms) due to the difficulties associated with their detection. Nevertheless, it was suggested that structural variants could play a profound role in the evolution of species. Inversions particularly are considered to be a potent mechanism for both adaptation and speciation due to their ability to suppress recombination. This thesis provides the first insight into the structural variation between two closely related naturally hybridizing species, the common nightingale (*Luscinia megarhynchos*) and the thrush nightingale (*Luscinia luscinia*). Structural variants were detected using long-read sequence data and high-quality *de novo* whole genome assemblies from one individual per species. High-confidence sets of structural variants were built by the intersection of results from several structural variant calling methods separately for each reference genome and included 18 839 variants for the common nightingale reference and 19 864 variants for the thrush nightingale reference. Among these, 9 candidate inversions polymorphic between the species were identified. These inversions could potentially play a role in nightingales' speciation.

Keywords: structural variants, speciation, adaptation, variant calling, inversions, common nightingale, thrush nightingale

Abstrakt

Strukturální varianty jsou mutace v sekvenci DNA ovlivňující pozici, orientaci nebo počet kopií úseků delších než 50 bp. Přestože má tento typ mutací potenciál k vytvoření velkých změn ve fenotypu, strukturální varianty jsou v porovnání s ostatními typy mutací (jako jsou například jednonukleotidové polymorfismy) z velké části neprobádané, neboť je poměrně obtížné je detekovat. Ukazuje se ale, že strukturální varianty by mohly hrát důležitou roli v evoluci druhů. Obzvláště inverze bývají považovány za účinný mechanismus pro adaptaci a speciaci kvůli jejich schopnosti potlačit rekombinaci. Tato práce poskytuje první vhled do strukturální variace mezi dvěma blízce příbuznými, přirozeně se křížícími druhy, slavíkem obecným (*Luscinia megarhynchos*) a slavíkem tmavým (*Luscinia luscinia*). Strukturální varianty byly detekovány pomocí dlouhých čtení a vysoce kvalitních *de novo* sestavených genomů z jednoho jedince od každého druhu. Finální výběr kandidátních strukturálních variant byl vytvořen pro každý referenční genom zvlášť jako průnik výsledků několika metod detekujících strukturální varianty a obsahoval 18 839 variant při referenci slavíka obecného a 19 864 variant při referenci slavíka tmavého. Mezi nimi bylo nalezeno 9 kandidátních inverzí polymorfních mezi druhy. Tyto inverze by potenciálně mohly hrát roli ve speciaci slavíků.

Klíčová slova: strukturální varianty, speciace, adaptace, variant calling, slavík obecný, slavík tmavý

Contents

List of abbreviations

BND	breakpoint
CN	common nightingale
DEL	deletion
DUP	duplication
FoSTeS	fork stalling and template switching
INS	insertion
INV	inversion
LD	linkage disequilibrium
MMBIR	microhomology-mediated break-induced replication
MMEJ	microhomology-mediated end joining
NAHR	non-allelic homologous recombination
NHEJ	non-homologous end joining
ONT	Oxford Nanopore Technologies sequencing
PAF	pairwise mapping format
QTL	quantitative trait loci
SSA	single-strand annealing
SV	structural variant
SNP	single nucleotide polymorphism
SRS	serial replication slippage
TN	thrush nightingale
VCF	variant calling format

1. Introduction

Although recent studies suggest that structural variants (SVs) are very common in a broad range of organisms (Catanach et al., 2019; Pang et al., 2010; Weissensteiner et al., 2020), they remain to be a largely understudied type of variation. That follows from the fact that until recently it was not possible to effectively analyze the entire range of their lengths and types. However, with the advancements of high throughput sequencing methods, this limitation was at least partially surmounted. Especially the development of long-read sequencing platforms made it possible to detect this variation class much more accurately. Long reads (i) facilitate highly continuous assembly of genomes, and (ii) can span the entire lengths of structural variants. The combination of these two benefits makes the investigation of structural variation a much more feasible task.

Since the discovery of the first inversions and duplications in the first half of the last century, structural variants were hypothesised to distinguish species (Bridges, 1936). Lately, a growing number of studies link structural variation with local adaptations and suggest its role in reproductive isolation. Particularly inversions were associated with variation in adaptive traits (Joron et al., 2011; Todesco et al., 2020). Recent evidence also suggests that sympatric species are more likely to differ in an inversion than allopatric species (Hooper & Price, 2017).

This thesis presents the first attempt to explore structural variation in the common nightingale (*Luscinia megarhynchos*) and the thrush nightingale (*L. luscinia*) from sequencing data. The species diverged recently (~1.8 Mya) (Storchová et al., 2010) and currently hybridize in their sympatric populations in central Europe, which makes this model an ideal candidate to study a potential role of structural variants and especially inversions (if found) in the evolution of species.

2. Structural variation and its role in the evolution

2.1 Structural variants

Structural variants (SVs) are structural changes in DNA that affect the position, orientation, or the presence/absence (or the number of copies) of long stretches of DNA. In comparison with previously more commonly studied single nucleotide polymorphisms (SNPs), SVs alternate larger regions of DNA. The threshold of what larger means, however, is in this case set slightly arbitrarily. Formerly it was usually set at longer than 1 000 bp (Feuk et al., 2006). Nevertheless, in more recent literature, we can commonly find SVs defined as regions longer than 50 bp (Sudmant et al., 2015; Mahmoud et al., 2019). In this thesis, the latter alternative will be used.

Structural variants are named in a way that describes the change in the DNA of an individual in comparison with a reference genome. There are five fundamental types of structural variants - insertions, duplications, inversions, deletions, and translocations. These simple structural variants are delimited by two breakpoints.

A SV is called an insertion (INS) if it introduces a novel stretch of sequence into a genome region (Figure 1A). Duplication (DUP) is an event in which an extra copy of an already present sequence region is incorporated. The two copies can either be located directly next to each other (in tandem - tandem duplications) (Figure 1B) or further away (interspersed duplications) (Figure 1C). A deletion (DEL) describes the disposal of a DNA region from the genome (Figure 1D). Together, deletions and duplications are called copy number variants (CNVs), as they describe the change in the number of copies of a locus in a genome. An inversion (INV) occurs when a sequence segment's orientation is turned around (Figure 1E). And finally, the term translocation (TRA) is used when a part of DNA changes its location within the genome. A reciprocal translocation is a type of translocation in which two chromosomal segments are reciprocally switched between two non-homologous chromosomes (Figure 1F). When such a translocation occurs between two acrocentric chromosomes and the exchanged segments span the entire length of long chromosome arms, the translocation is called a Robertsonian translocation (Figure 1G). This event can be also called a centric fusion, as the long chromosome arms are fused into one chromosome, while the short arms are usually lost.

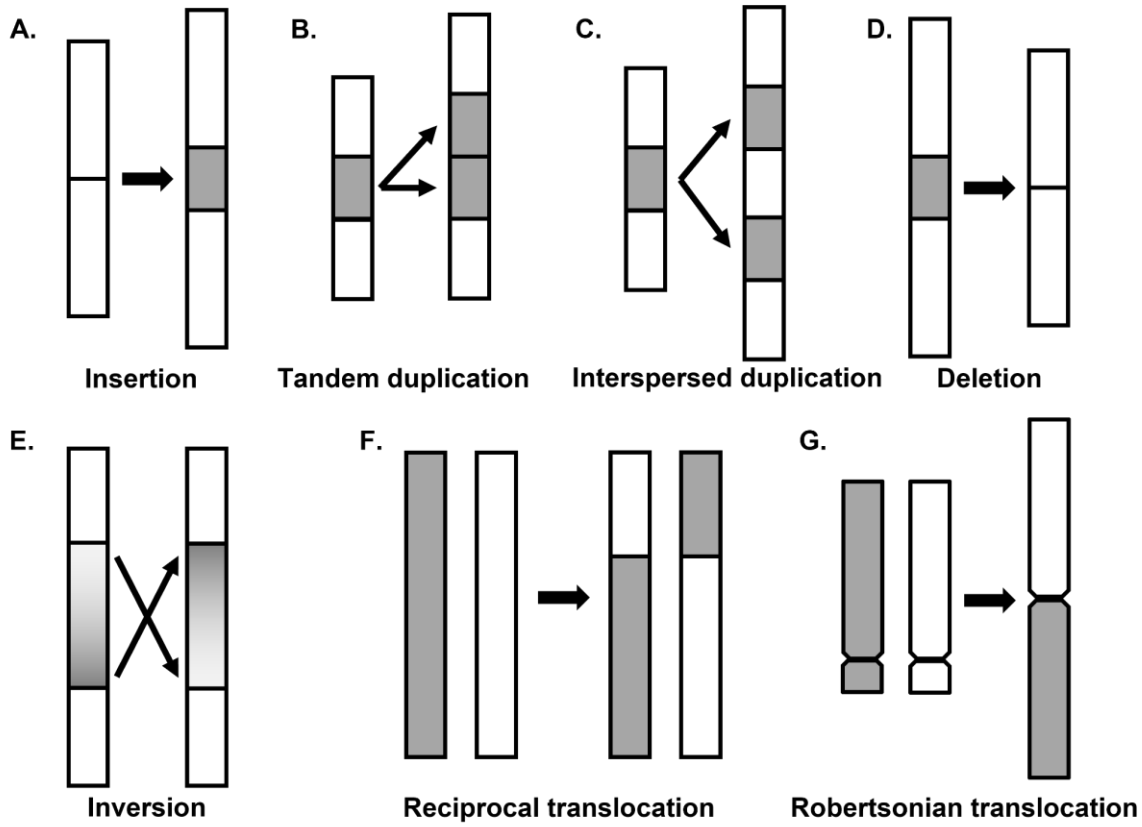


Figure 1: Types of structural variants.
A: Insertion. B: Tandem duplication. C: Interspersed duplication. D: Deletion. E: Inversion. F: Reciprocal translocation. G: Robertsonian translocation.

Various combinations of the five fundamental types of SVs form complex genomic rearrangements. They can include for example a triplication (combination of two duplications) (Zhang et al., 2009) or an inverted duplication (Allen et al., 2004; Hermetz et al., 2014).

2.2 How do SVs arise?

There are numerous ways how SVs can arise in a genome. They can be a result of an error during recombination (non-allelic homologous recombination), DNA replication (serial replication slippage, fork stalling and template switching, microhomology-mediated break-induced replication), or a break repair (microhomology-mediated end joining, non-homologous end joining, single-strand annealing). A great source of SVs also streams from the activity of transposable elements across the genome (mobile element insertions).

Non-allelic homologous recombination (NAHR) follows an ectopic crossing over event (i.e. crossing over in which two non-homologous regions are aligned). By this mechanism, recombination occurs between areas with high sequence identity (such as CNVs or *Alu* and L1 elements), which however lies in different parts of the genome. This then can introduce a deletion, a duplication, an inversion, or a translocation into the genome (Lupski, 1998; Robberecht et al., 2013) (Figure 2).

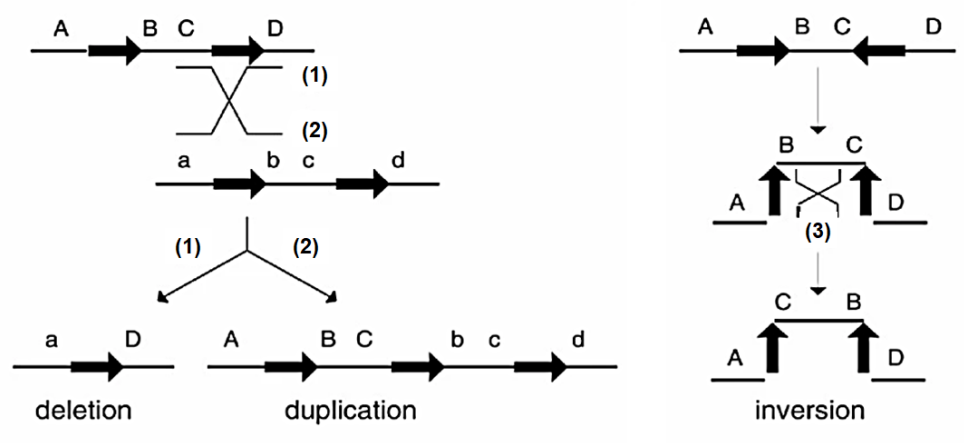


Figure 2: Schematic of the non-allelic homologous recombination.
Pairs of the same letters in uppercase and lowercase represent pairs of homologous sequences. Thick arrows represent regions with high sequence identity. Crossing lines represent a crossing-over event. After crossing over in non-homologous regions caused by the high sequence identity, part of the original sequence is (1) missing, (2) duplicated, or (3) inverted. Adapted from Gu et al., 2008.

Errors in DNA replication that can lead to the origin of smaller rearrangements are serial replication slippage (SRS), fork stalling and template switching (FoSTeS) and microhomology-mediated break-induced replication (MMBIR). SRS occurs in repetitive areas of a genome. DNA polymerase skips (forward slippage) or repeats the replication of a region (backward slippage) due to a brief uncoupling of the primer and template strands followed by their displaced realignment at a different copy of the repeated sequence (Viguera et al., 2001). This can result not only in simple deletions and duplications but also in tandem repetitive sequences such as microsatellites if the template slips forward or backward multiple times in a row (Chen et al., 2005). FoSTeS mechanism follows a single strand DNA lesion encounter that causes the replication fork to stall. This prompts the replication fork to switch to a template of another nearby replication fork (Lee et al., 2007). This model was later generalized into MMBIR (Hastings et al., 2009) (Figure 3), which

describes the rearrangements origin as a process similar to the well studied break-induced replication (double-strand break repair mechanism), only without the requirement of long homologous sequences between the lagging and the new template strand, leading to the annealing of microhomologous sites (Hastings et al., 2009). By these mechanisms, inversions, duplications, and translocations, as well as complex genomic rearrangements can be formed (Zhang et al., 2009).

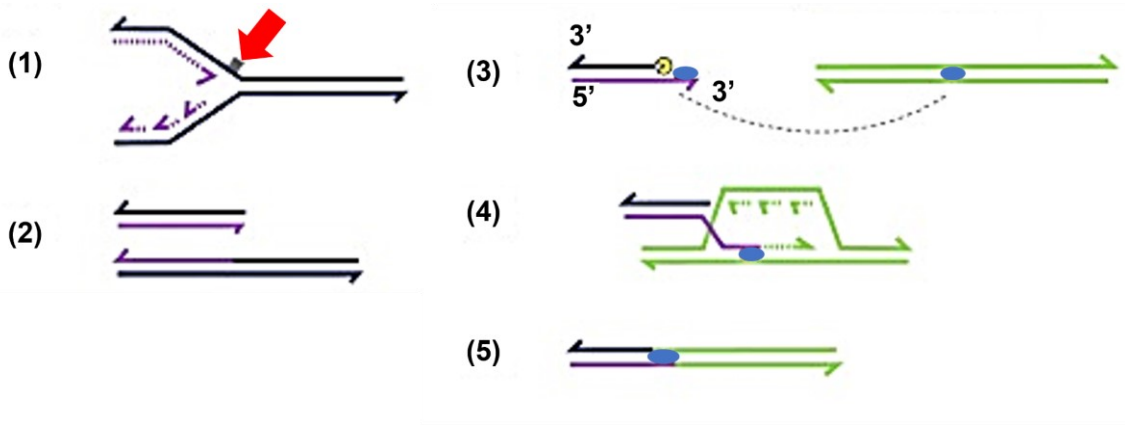


Figure 3: Schematic of microhomology-mediated break-induced replication.
(1) Replication fork is stalled by a lesion (red arrow). (2) Double-strand break arises. (3) 5' to 3' degradation uncovers a microhomologous region (blue). (4) Replication fork reassembles with different template strand. (5) Replication continues. Adapted from Ottaviani et al., 2014.

Structural variation can also be introduced due to a mistake in the DNA double-strand break repair process, such as non-homologous end joining (NHEJ), microhomology-mediated end joining (MMEJ), or single-strand annealing (SSA). All three mechanisms rely on a simple ligation of split DNA strands, however with a different level of required homology. While MMEJ requires only microhomologous regions (5-25 bp), SSA typically looks for longer homologous sequences (McVey & Lee, 2008). In addition, in NHEJ, the binding of factors to DNA ends prevents it from resection and guides it towards ligation with another DNA stretch with an end processed in this way (Moore & Haber, 1996). This makes NHEJ the least likely to introduce any rearrangements. The other two mechanisms, MMEJ and SSA, need to employ ends' degradation to uncover sequences which then can be used for homologous search of a counterpart to anneal to. This can be a source of deletions or translocations.

A rich source of insertions and deletions is in the activity of transposons across the

genome. Transposons can move in the genome by “cut and paste” or “copy and paste” mechanisms resulting in new mobile element insertions. Relative to the reference genome they manifest themselves as deletions or insertions (Stewart et al., 2011).

Complex genomic rearrangements can be a result of a single catastrophic mutation event. These are typical mainly for cancer cells. Such an event is for example chromothripsis, in which a genomic region is broken down into several pieces which are then rearranged back together in an aberrant order (Pellestor, 2019).

2.3 The history of structural variants research

The history of structural variants research goes back to the first half of the 20th century. In his comparison of *Drosophila simulans* and *D. melanogaster* chromosomes published in 1921, Alfred Sturtevant proposed the existence of an inversion between these two species' genomes on chromosome 3 (Sturtevant, 1921). He also foresaw this inversion as a recombination suppressor in the inverted area and its surroundings (Sturtevant & Mather, 1938). Other studies identifying large chromosomal rearrangements soon followed, for example, deletions, inversions and reciprocal translocations were described in maze (*Zea mays*) (McClintock, 1931), a gene duplication in *Drosophila melanogaster* (Bridges, 1936), and inverted regions in strains of *Drosophila pseudoobscura* (Dobzhansky & Sturtevant, 1938). Early on, chromosomal aberrations were also linked with several human diseases, such as Down's syndrome (Jacobs et al., 1959) or cancer (Nowell & Hungerford, 1960). The early cytogenetic studies of rearrangements in karyotypes were however somewhat overshadowed in the late 20th century by the research of other genetic markers (e.g. microsatellites, SNPs) and by the assumption that SNPs are the main source of genetic variation in populations.

Since the beginning of this century, it is becoming abundantly clear that SVs indeed play an important role in both human (Pang et al., 2010) and non-human (Hooper et al., 2019; Lowry & Willis, 2010; Manoukis et al., 2008) population genetics. This is linked with the development of high throughput techniques for SVs detection and genotyping. Variant calling from sequencing data produced by next-generation sequencing platforms makes it possible to explore all types of SVs across whole genomes at once (Catanach et al., 2019; Weissensteiner et al., 2020).

2.4 The role of structural variants in the evolution of species

SVs are theorized to play a significant role in the evolution of species, both in adaptive evolution of species and origin of reproductive isolation between species (speciation). SVs can directly create adaptive mutations (Kirkpatrick, 2010), or function as a mediator, providing a genomic background for other evolutionary processes to act on (Villoutreix et al., 2020).

Direct effects of structural variants include changes in gene expression caused by a rearrangement breakpoint getting in the way of an open reading frame (Guerrero et al., 2012) or by the introduction of a strong expression promoter in the gene vicinity by the insertion of a transposable element. Larger (spanning several Mb) SVs such as inversions have the ability to reduce recombination among multiple alleles of different genes. That can result in the formation of so-called "supergenes", which can underlie polymorphism in very complex phenotypes. Indirect effects include reduced fertility of heterozygotes for larger genomic rearrangements such as translocations.

SVs have been linked with adaptive traits in several studies. Formation of a "supergene" has been shown to underlie mimetic wing patterns in *Heliconius* butterflies (Joron et al., 2011) or behavioral and plumage morphs in ruff birds (*Philomachus pugnax*) (Küpper et al., 2016). Altered gene expression as a result of an insertion of a transposable element in the proximity of genes was observed in peppered moths (*Biston betularia*) (Van't Hof et al., 2016). And a large deletion of multiple genes is likely the underlying reason for cryptic coloration in stick insects (Villoutreix et al., 2020).

SVs could also play a large role in the establishment of reproductive isolation between species, although evidence for it is still scarce (Zhang et al., 2021). Below, several models of how SVs can contribute to reproductive isolation are described.

(1) SVs can directly cause mutations contributing to reproductive isolation

This was reported for example in two crow species (*Corvus corone* and *Corvus cornix*), where a retrotransposon insertion reduced expression of gene in its close proximity affecting the color of plumage, which contributed to premating isolation (Weissensteiner et al., 2020).

(2) Hybrid sterility models

The presence of a large structural variant, such as translocation or an inversion, in chromosome can lead to problems with meiotic pairing in individuals heterozygous in said SV. Possible mispairing might have a critical effect on gametogenesis and can lead to the production of defective gametes. That can cause sterility or subfertility of such an individual (Johnson, 2008). When considering the individual to be a hybrid between two populations (with the rearrangement fixed in one of them), this mechanism can promote reproductive isolation.

Such a mechanism was suggested to contribute to reproductive isolation for example between sunflower hybrid species *Helianthus anomalus*, *H. deserticola* and *H. paradoxus* and their parental species *H. annuus* and *H. petiolaris* (Lai et al., 2005). In this study, the authors described karyotypic differences between the parental species and hybrid species, identifying several translocations that emerged in the hybrid species *de novo* (i.e. were not present in neither of the parental species). Using quantitative trait loci (QTL) analyses, they showed the co-localization of rearrangements and pollen viability QTL in hybrids, which strongly suggest that SVs are directly responsible for the reduced fertility of hybrids.

(3) Reduced recombination models

Some types of structural variants, particularly inversions, have the ability to reduce the recombination rate in the rearranged genomic regions (Sturtevant, 1917). The suppression of recombination is caused by problems with meiotic pairing in heterozygotes or by the defective gametes containing deleterious mutations produced by recombination inside the inversion (Kirkpatrick, 2010). This phenomenon was observed in a wide range of species including yeast *Saccharomyces Cerevisiae* (Dresser et al., 1994) as well as insects (e.g. *Drosophila*, fire ants *Solenopsis invicta*) (Kulathinal et al., 2009; Wang et al., 2013) and plants (e.g. sunflowers *Helianthus petiolaris* and *H. annuus*, *Arabidopsis thaliana*) (Ederveen et al., 2015; Rieseberg et al., 1999).

As opposed to the hybrid-sterility models, reduced recombination models do not assume the rearrangements to have a direct effect on the fitness of the hybrids, but rather expect them to influence recombination. If speciation occurs in the face of ongoing gene flow (Rieseberg, 2001), inversions can help to prevent the recombination within the SV,

which may help to maintain the species-specific combinations of traits in the face of gene flow and thus facilitate speciation (Figure 4).

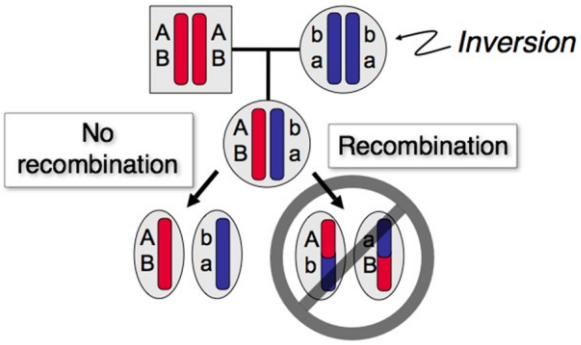


Figure 4: Inversion suppressing recombination.

Suppressed recombination caused by an inverted region spanning loci segregating for alleles *A/a* and *B/b* provides a mechanism to preserve combinations of alleles *AB* and *ab*. Adapted from Kirkpatrick, 2010.

A famous example suggesting the role of an inversion acting as a recombination suppressor in reproductive isolation was reported in yellow monkeyflower (*Mimulus guttatus*). Two ecotypes of this flower, annual and perennial, are adapted to their respective habitats, with the adaptive trait being among others flowering time, which creates an extrinsic prezygotic reproductive barrier. Using QTL mapping, an inversion was identified that involves loci responsible for much of the ecotypes' differentiation as well as loci contributing to reproductive isolation between the ecotypes (Lowry & Willis, 2010). In birds, inversions reducing gene flow on the Z chromosome were shown to differentiate in two subspecies of the long-tailed finch (*Poephila acuticauda*), however, in this case, a mechanism by which these affect reproductive isolation is unknown (Hooper et al., 2019).

(4) Models with gene duplication as a mechanism of intrinsic postzygotic isolation

Another type of structural variant with the potential to promote speciation is gene duplication. Gene duplication can in one of the copies either lead (i) to a loss of function by the accumulation of deleterious mutations or (ii) to the development of a new, original function by fixation of an advantageous allele, or (iii) the purpose of the original gene could be partitioned between the two copies accumulating complementary degenerative mutations

resulting in subfunctionalization of these two genes (Force et al., 1999; Lynch & Force, 2000). Both loss of function and subfunctionalization can cause hybrid sterility or inviability in backcross or F2 intercross hybrids, whose genome is a mosaic of the two parental species and may thus lack a functional copy of the respective duplicated gene.

This mechanism was observed between naturally hybridizing sister species of the yellow monkeyflower, *Mimulus nasutus* and *M. guttatus*. In this system, a subset of interspecific hybrids is unviable, because these individuals are missing a functional copy of an essential photosynthetic gene (Zuellig & Sweigart, 2018).

Also gene transpositions can possibly have a very similar effect on the reproductive barrier as the one explained here for duplications. That was proposed in *Drosophila melanogaster* and *D. simulans*. In this case, the absence of a gene crucial for male fertility causes sterility in some of the interspecific hybrids (Masly et al., 2006).

2.5 Detection of structural variants

Throughout history, a variety of methods for the identification of structural variants was presented. From detection of large (> 3 Mbp) rearrangements in patterns created by chromosomal banding techniques and hybridization approaches for characterization of copy number variants, the discovery of SVs from sequencing data emerged as a cost-effective high-throughput alternative facilitating characterization of rearrangements of all types and sizes at once while resolving their breakpoints at a single-nucleotide level (Balachandran & Beck, 2020).

Genomic sequencing techniques of SV calling are mostly inferring rearrangements from the mapping of short or long reads to a reference genome. Apart from that, some methodologies also include assembly of the reads into the process, either in the form of comparing a complete *de novo* assembly of a genome to the reference (i.e. instead of mapping the sequencing reads to the reference genome they directly map the assembled genome), or locally, by assembling reads mapping to the same area into contigs to achieve better results in that particular region (Rimmer et al., 2014).

SV callers vary in the types of SVs they can identify. Some of them specialize only in the detection of a specific type of SVs (e.g. Assemblytics focuses only on insertions,

deletions, and repeat expansions) (Nattestad & Schatz, 2016), while others do not interpret the type of called variant at all (e.g. GRIDSS) (Cameron et al., 2017). In that case, the SV caller outputs all recognized variants as a simple breakpoint (BND). While many SV callers can now assess all fundamental SV types at once (e.g. Manta, SVIM) (Chen et al., 2016; Heller & Vingron, 2019), detection of more complex rearrangements is more complicated and still a prerogative of a few programs, such as Sniffles (Sedlazeck, Rescheneder, et al., 2018).

2.5.1 Structural variant calling from short reads

For the detection of SV using short reads (usually produced by an Illumina instrument), paired-end sequencing data are employed to identify a rearrangement. In the presence of a SV, the distance between two mapped reads of a pair does not match the insert size of the sequencing library or the orientation of paired reads does not add up (Mahmoud et al., 2019). Furthermore, analysis of split reads (i.e. reads overlapping a rearrangement within themselves and therefore mapping by different segments to multiple different loci in the reference genome) can improve the resolution of the breakpoints (Balachandran & Beck, 2020). Finally, incorporating sequence coverage data into the process helps with the identification of copy number variants (Figure 5). On top of that, the aforementioned local assembly can further improve the results.

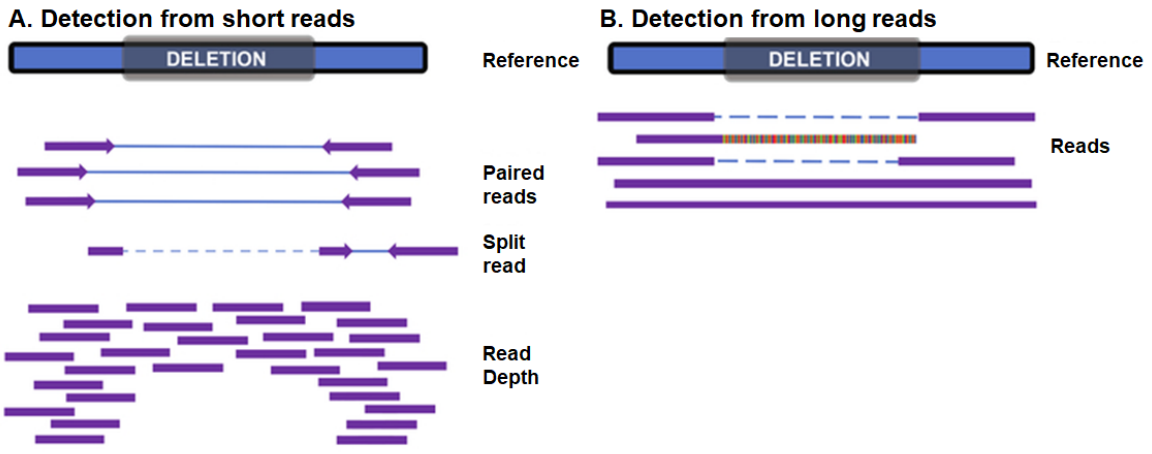


Figure 5: **Comparison of the short-read and long-read mapping approach to detection of structural variants.**
In the short-read approach (A), structural variants are elucidated from the patterns of alignments of paired reads and split reads and read depth. In the long-read approach (B), reads span the entire length of the variant. As an example, patterns for both approaches are shown here for a heterozygous deletion. Adapted from Balachandran & Beck, 2020.

Programs detecting SVs from short genomic reads are for example BreakDancer (Chen et al., 2009), Pindel (Ye et al., 2009), DELLY (Rausch et al., 2012), LUMPY (Layer et al., 2014), and MANTA (Chen et al., 2016). However, not every SV caller combines all of the methods described above. Different programs also tend to perform better with different SV types and sizes (Kosugi et al., 2019). Possible solution to that is offered by tools combining variants called by multiple methods, such as MetaSV (Mohiyuddin et al., 2015) and SURVIVOR (Jeffares et al., 2017).

2.5.2 Long-read mapping-based approach

With the development of long-read sequencing platforms such as Pacific Biosciences and Oxford Nanopore Technologies (ONT), methods to exploit data generated by these instruments to call the structural variants also emerged. The benefit of long reads for this task lies in the size of the genomic region that they can cover at once (several kb). They can often span the entire length of a structural variant and they also allow to identify SVs in more complex regions (e.g. repetitive elements) (Sedlazeck, Lee, et al., 2018). The drawback of long sequencing techniques is the high (over 10%) sequencing error rate (Goodwin et al., 2016), which needs to be taken into consideration not only in the SV calling step but also already during the alignment of reads to reference genome. Aligners tackling this issue are for example Minimap2 (Li, 2018) or NGMLR (Sedlazeck, Rescheneder, et al., 2018), which implements a convex gap scoring model that facilitates distinguishing sequencing errors from structural variation. These programs are capable of producing split alignments (i.e. split a read mapping by its different segments to separate parts of the genome and report all these mappings), which are crucial to SV identification.

Some of the SV callers employing long reads are sequencing-platform specific. These are for example PBHoney (English et al., 2014) and SMRT-SV (Huddleston et al., 2017) for data generated by Pacific Biosciences instruments. Other tools, such as Sniffles (Sedlazeck, Rescheneder, et al., 2018) and SVIM (Heller & Vingron, 2019), allow detection from both ONT and Pacific Biosciences sequencing reads. Both Sniffles and SVIM identify all five fundamental types of SVs and work similarly. They scan the alignments of single reads to the reference genome to identify signatures indicating the presence of a SV. These signatures are then clustered across all reads according to their features (position in the genome, span, number of supporting reads) and finally combined to identify the higher-order

variants (i.e. variants involving multiple regions in the genome, such as interspersed duplications and translocations).

In comparison with the short-read variant calling approach, detection of SVs using the long-read sequencing data is better at the identification of longer SVs (Mahmoud et al., 2019; Sedlazeck, Rescheneder, et al., 2018) (Figure 5). Comparative studies of SV callers also conclude that it is beneficial to employ multiple programs for the detection of SVs and combine their results to obtain more reliable results (of better sensitivity or better accuracy) (De Coster et al., 2019; Liu et al., 2020).

2.5.3 *De novo* assembly-based approach

The concept behind the de *novo* assembly-based approach is essentially very simple. Genome sequence is assembled out of the sequencing data and aligned to reference genome or another assembly. SVs are then detected directly from the differences between the two aligned sequences caused by the rearrangements (Mahmoud et al., 2019). The crucial step of this method is the computation of the whole-genome alignment. Aligners developed for this task are for example MUMmmer (Delcher et al., 1999) or Minimap2 (Li, 2018). Although this approach has the potential to detect all types of structural variants, the majority of available programs focus only on the detection of insertions and deletions - these are for example paftools.js (Li, 2018) and Assemblytics (Nattestad & Schatz, 2016). This is however not true for SVIM-asm SV caller, which identifies all fundamental types of SVs (Heller & Vingron, 2020).

2.6 Model system

This thesis investigates structural variation in genomes of two closely related bird species, the common nightingale (*Luscinia megarhynchos*) and the thrush nightingale (*Luscinia luscinia*), which belongs to the family *Muscicapidae* of the order *Passeriformes*. Representatives of both species are smaller migratory passerine birds, who share most of their morphological and ecological traits. In spite of the similarities, it is possible to distinguish these species by differences in their overall body size, plumage colour and wing characteristics. For the most part, they also differ in breeding areas, the common nightingale

nesting in western and southern Europe and the thrush nightingale in eastern and northern Europe. The two species diverged approximately 1.8 million years ago (Storchová et al., 2010). Their breeding grounds overlap in the secondary contact zone throughout central and eastern Europe where interspecific hybridization occasionally occurs (Kverek et al., 2008; Reifová et al., 2011). In concordance with Haldane's rule (Haldane, 1922), hybrids of the heterogametic sex (i.e. in this case females) are sterile (Reifová et al., 2011).

Birds have, compared to other vertebrates, relatively conserved karyotypes (Damas et al., 2019). However, studies which would identify SVs from genomic data are still scarce in birds. It is thus unclear how often closely related species differ in SVs and thus whether SVs can potentially contribute to speciation in birds. Previously reported studies focused mostly on inversions, indicating that they arise faster on the sex chromosomes than on the autosomes (Hooper et al., 2019; Hooper & Price, 2015) and that they are more prevalent in sympatric species than in allopatric species (Hooper & Price, 2017).

3. Aims of this thesis

The aim of this thesis was to examine the structural differences in genomes of two closely related species *Luscinia megarhynchos* and *L. luscinia,* that might have contributed to the reproductive isolation between these two species. To do so we have analysed sequencing data and assembled genomes from one individual of each species and identified possible genomic rearrangements using a combination of multiple SV callers. Focusing in greater detail on inversions, we have then tried to confirm their presence in the genome by closer inspection of the sequencing data underlying their identification by the SV callers. Specific aims included:

1) Identification of a high-confidence set of SVs.

   A high-confidence set of SVs was built from the overlap of rearrangements identified by three different programs, two long-read mapping-based and one of the *de novo* assembly-based approach.

2) More detailed inspection of candidate inversions from the high-confidence SVs set to assess their correctness and their potential role in the evolution of nightingale species.

4. Materials and Methods

4.1 Input data - reference genomes and sequencing reads

Genomes used for analyses in this thesis were *de novo* assembled from Oxford Nanopore Technologies (ONT) sequencing data of one female individual of the common nightingale (CN) and one female of the thrush nightingale (TN). Birds were captured in allopatric populations, specifically in the Czech republic (CN) and northern Poland (TN). For the common nightingale, over 3.25 million reads of average length 8 555 generated overall coverage of approximately 25x. For the thrush nightingale, ~27x coverage was achieved by ~6.87 million reads averaging in length at 4 451 bp. Assemblies were created using the Flye assembler (Kolmogorov et al., 2019), subsequently corrected with short Illumina reads and 10x Genomics linked reads obtained from the same individuals, and further improved by the chromosome conformation capture technique. Ultimately, 3 340 scaffolds were assembled for the common nightingale and 9 654 for the thrush nightingale, with the N50 statistics of 61.2 Mb and 62.4 Mb for the common nightingale and the thrush nightingale respectively (Table 1) (Rídl et al. unpublished results).

Table 1: Characteristics of the assemblies.

	common nightingale assembly	thrush nightingale assembly
Number of scaffolds	3 340	9 654
N50 (bp)	61 212 290	62 437 632
Largest scaffold (bp)	111 736 527	111 413 366
Total sequence length (bp)	1 098 093 286	1 102 405 890
Scaffold representing the Z chromosome	5	8

4.2 Structural variant calling

4.2.1 Analysis workflow

The diverse types of data available for two individuals allowed for the adaptation of multiple approaches to structural variant (SV) detection. Firstly, long-read mapping-based approaches were applied, relying on the alignment of ONT reads to a reference genome (reads-to-genome alignment and SV calling). And secondly, a *de novo* assembly-based approach was used, where SVs are called from the whole-genome alignment of both assemblies (genome-to-genome alignment and SV calling). Both types of analyses required selecting one genome as the reference sequence against which the SVs were called. All computations were therefore done in both directions (i.e. using both CN and TN as a reference) (Figure 6). Primarily, we were interested in detecting SVs between species (interspecific comparisons). However, as a control, we also conducted intraspecific comparisons, where SV calling from alignment of the sequencing data to the assembled genome of the same species was done. Such intraspecific analysis can detect within-species polymorphisms (if a variant is supported by approximately half of the mapped reads) and/or errors in the assembly (if a variant is supported by the entirety of mapped reads) and allow calls representing such events to be deciphered. For the sake of intelligibility of this thesis, this naming convention to identify the type of analysis will be further used: “species reference & species query” (i.e. CN reference & TN query stands for detection of variants in the thrush nightingale against the reference genome sequence of the common nightingale).

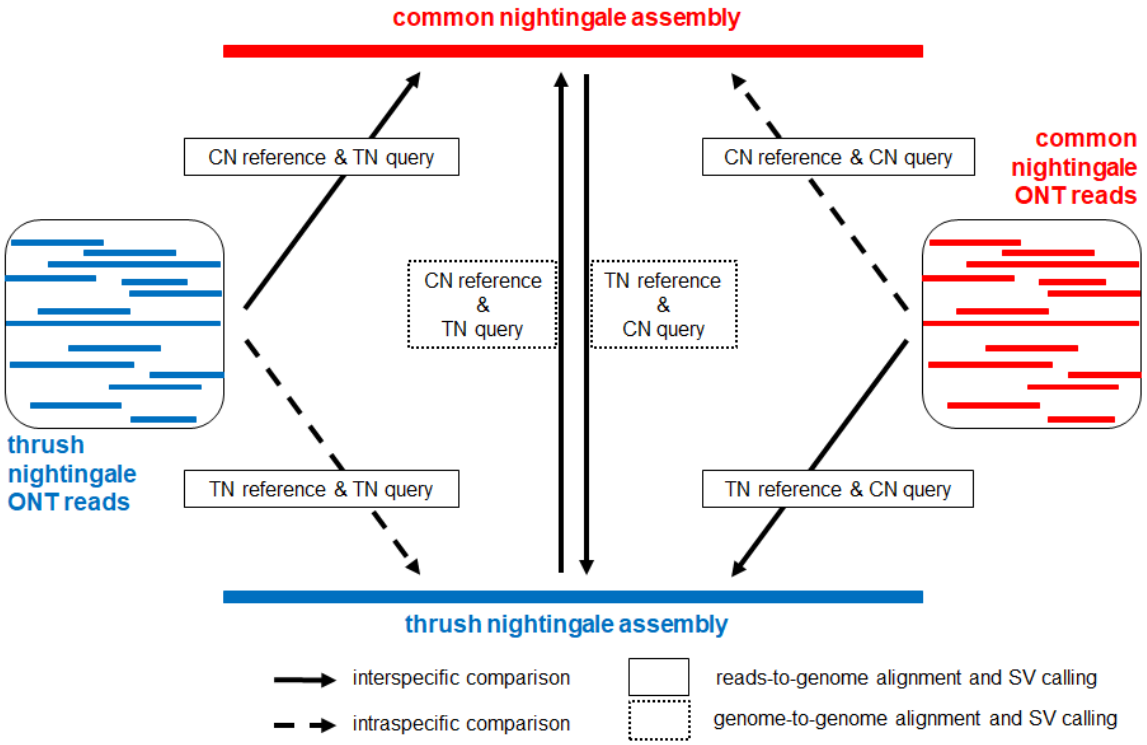


Figure 6: Diagram of the analysis workflow. Input data for the common nightingale (CN) are presented in red, for the thrush nightingale (TN) in blue. Arrows depict the process of alignment and SV calling and are oriented in the direction from the query (i.e. the species in which the SVs are detected) to the reference (i.e. the species against whose genome are the SVs called).

4.2.1 Reads-to-genome and genome-to-genome alignments

Reads from ONT sequencing were aligned to the respective assembled genome using Minimap2 aligner, version 2.20 (Li, 2018), with the following options: -a to generate output in sam format, -x map-ont, a preset option for the alignment of long ONT reads, and --MD adding the optional MD tag containing information about mismatched and deleted reference bases into the output SAM file, as required by the variant callers.

Whole genome alignments were produced also using the Minimap2 aligner (Li, 2018). The following options were used: -a to generate output in sam format and -x asm10, which is a preset option for alignment of an assembly to a reference with up to 10% sequence divergence.

All computed alignments were subsequently sorted and indexed using samtools sort and samtools index functions (SAMtools version 1.10) (Daněček et al., 2021). Whole

genome alignments were also converted to pairwise mapping format (PAF) using paftools.js script (part of the Minimap2 software package) (Li, 2018). PAF-converted alignments were used for the conversion of genomic positions between the two species.

4.2.2 Structural variant calling

Variant calling from generated reads-to-genome alignments was performed by two structural variant callers, Sniffles (Sedlazeck, Rescheneder, et al., 2018) and SVIM (Heller & Vingron, 2019). Both Sniffles (version 1.0.11) and SVIM (version 1.4.2) were run with the default parameters generating an output in the variant calling format (VCF). One important characteristic distinguishing the outputs of these two programs is the way they treat extremely long variants. SVIM outputs variants longer than 100 kb as translocation breakpoints irrespective of whether they represent translocations or other types of SVs (i.e. a breakpoint call in results from SVIM can either be a translocation or any type of variant spanning more than 100 kb), while Sniffles does not have an upper bound on length of a SV (i.e. even variants longer than 100 kb are called with their specific types). Another difference is in the final filtering of the variants. Sniffles outputs only variants supported by at least 10 sequencing reads, whereas SVIM includes even hits supported by as little as one read, but accompanies them by a score expressing the weak experimental support. While the score for most variant types reflects primarily the number of reads supporting a variant call, the score of inversions might be underestimated if one inversion breakpoint is supported better than the other one. To avoid this and also to evade setting a fixed threshold for the score, unfiltered outputs from SVIM were used with the assumption that low-scoring poorly supported hits will be filtered out in the intersection step (see below).

From the genome-to-genome alignments, rearrangements were identified using SVIM-asm (Heller & Vingron, 2020). SVIM-asm (version 1.0.2) was run in the haploid mode with the default settings. Notably, with these parameters, rearrangements longer than 100 kb are called translocation breakpoints, i.e. in the same way as by SVIM (Heller & Vingron, 2019).

4.3 Creating a high confidence set of structural variants

A high confidence set of putative structural variants was created using the SURVIVOR toolset (Jeffares et al., 2017). Firstly, from each VCF file, variants not longer than 50 bp were filtered out using the SURVIVOR filter command. Secondly, an intersection of results from individual callers was created for each direction of interspecific analysis using the SURVIVOR merge tool. SURVIVOR merge command combines calls from different programs based on their agreement in breakpoints locations and type. Currently, there is no standardized value for breakpoints distance to use as a threshold for merging (De Coster et al., 2021). SURVIVOR allows using either a fixed distance in a number of base pairs or, alternatively, it can assess this parameter relatively to a variant's length. To better reflect the wide range of rearrangements' sizes, the latter approach was adopted. The merging process was carried out multiple times with this parameter iterating over values between 0 and 1, expressing the maximal distance between corresponding breakpoints in the percentage of respective variant's length (i.e. for a 100 bp variant and parameter value of 0.1, variants with breakpoints within 10 bp distance from the particular variant's breakpoints are examined). To account for the different behaviour of SV callers in case of long rearrangements (e.g. a deletion longer than 100 kb might be called as deletion by Sniffles, while SVIM and SVIM-asm would output it as a translocation breakpoint due to their length restriction), SURVIVOR merge was not set up to consider variant type. Results of this merging were plotted in R studio (RStudio Team, 2020) using package ggplot2 (Wickham, 2016).

4.4 Closer inspection of inversions polymorphic between species

From the high-confidence set of SVs, inversions were selected to undergo further scrutiny. Calls from both interspecific comparisons (variants called against both reference genomes) were compared and inspected in Integrative Genomics Viewer (IGV, version 2.8.0) (Robinson et al., 2011). For each reference genome, the following tracks were loaded into the IGV: reads-to-genome alignment of the query species' reads to the reference, reads-to-genome alignment of the reference species' reads to the reference, genome-to-genome alignment of the two genomes, and variant calls from all interspecific and intraspecific comparisons with the selected reference. Rearrangements discovered by intraspecific

comparisons were used to distinguish putative true variants from errors in assemblies. For the purposes of visualization of inversions, alignments were colored by the orientation in which they map.

5. Results

5.1 Structural variant calling from reads-to-genome alignments

5.1.1 Interspecific comparisons

Results of variant calling with Sniffles from the alignment of long ONT reads to the reference genomes in the interspecific comparisons are summarized in Table 2. In total, Sniffles identified 47 047 and 46 430 variants longer than 50 bp in analyses with the common nightingale and the thrush nightingale as a reference, respectively. The majority of the discovered variants are insertions and deletions shorter than 1 000 bp (87.7% and 84.4% for the CN and the TN reference, respectively). Out of 121 and 148 inversions detected using CN reference and TN reference analyses, respectively, more than 60% were predicted to span more than 10 kb. Both analyses identified approximately 100 duplications (103 using CN reference and 100 using TN reference) of which more than a half falls into the largest size category (over 10 kb). The number of translocation breakpoints called using CN reference & TN query analysis was lower than the number of breakpoints using the TN reference & CN query analysis (3 869 vs. 5 325, respectively) (Table 2).

Table 2: Number of SV calls by Sniffles with the common nightingale genome as the reference and the thrush nightingale sequencing data as the query (CN reference & TN query) and with the thrush nightingale as the reference and the common nightingale as the query (TN reference & CN query) split up by type and size. In the CN reference & TN query, callset were two variants with an unresolved type between a deletion and an inversion that are reported here as deletions. Two variants with an unresolved type between a deletion and an inversion from the TN reference & CN query dataset are reported here as deletions and one variant of type undecided between a duplication and an insertions as a duplication. DEL stands for deletions, DUP for duplications, INV for inversions, INS for insertions and BND for translocation breakpoints.

	CN reference & TN query					**TN reference & CN query**				
Length	DEL	DUP	INV	INS	BND	DEL	DUP	INV	INS	BND
50-100 bp	8 872	0	0	9 047	3 869	7 620	0	0	8 823	5 325
100-1000 bp	13 938	14	10	9 403		12 570	6	19	10 155	
1-10 kb	1 019	35	34	587		937	38	28	669	
> 10 kb	80	54	77	8		80	56	101	3	
Total	**23 909**	**103**	**121**	**19 045**	**3 869**	**21 207**	**100**	**148**	**19 650**	**5 325**

The sum of SVs contained in the initial output from SVIM variant calling goes to several hundreds of thousands (Table 3). It is important to stress that these results are unfiltered and therefore certainly contain false positive hits. Once again, shorter insertions and deletions mostly prevail, although especially in the CN reference & TN query output, breakpoints are also quite ubiquitous. By contrast, inversions are the least occurring variants with only a few hundreds of candidates.

Table 3: Number of structural variant calls by SVIM with the common nightingale genome as the reference and the thrush nightingale sequencing data as the query (CN reference & TN query) and with the thrush nightingale genome as the reference and the common nightingale sequencing data as the query (TN reference & CN query) split up by type and size. DEL stands for deletions, DUP for duplications, INV for inversions, INS for insertions and BND for translocation breakpoints and SVs longer than 100 kb.

	CN reference & TN query					**TN reference & CN query**				
Length	DEL	DUP	INV	INS	BND	DEL	DUP	INV	INS	BND
50-100 bp	102 983	171	6	113 375	217 726	193 531	155	4	87 097	28 694
100-1 000 bp	131 730	2 203	196	122 120		209 400	1 874	150	112 141	
1-10 kb	6 080	882	223	4 157		4 886	784	132	4 180	
> 10 kb	238	620	173	99		105	473	101	47	
Total	**241 031**	**3 876**	**598**	**239 751**	**217 726**	**407 922**	**3 286**	**387**	**203 465**	**28 694**

5.1.2 Intraspecific comparisons

SV calling by Sniffles in the intraspecific comparison, which can detect variant polymorphisms within the species and/or possible errors in the assembly, identified 15 593 variants by the CN reference & CN query procedure and 19 716 variants by the TN reference & TN query procedure (Supplementary Table S1). The majority of these consists of shorter (length below 1 kb) insertions and deletions (68.9% and 62.3% in the CN and TN, respectively), much like in the interspecific comparison, although here the number of insertions exceeds the number of deletions. These analyses also called 99 putative inversion polymorphisms in the CN and 120 in the TN, with the biggest size category spanning regions over 10 kb long having superiority in numbers (77 in the CN and 90 in the TN). Intraspecific comparisons by SVIM yielded hundreds of thousands of candidates of potential

heterozygous or erroneous sites. However, more than 90% of these can be filtered out by applying a moderate filter eliminating variants with the value of score lower than 5 (Supplementary Table S2). The overall numbers of called variants in the intraspecific comparisons were lower than in the interspecific comparisons, as was expected.

5.2 Variant calling from genome-to-genome alignments

Variant calling from genome to genome alignments of both assembled sequences by SVIM-asm yielded a total of 71 782 putative rearrangements in the comparison of the TN assembly against the CN reference and 70 893 in the comparison of the CN against the TN reference (Table 4). Insertions and deletions of length not exceeding 1 000 bp make up a vast majority of these candidates (approximately 95% in both analyses). While all 28 duplications identified with the CN reference & TN query approach were predicted as tandem, 4 out of 41 duplications called in the TN reference & CN query tactic were marked as interspersed.

Table 4: Number of SV calls by SVIM-asm with the common nightingale genome as the reference and the thrush nightingale genome as the query (CN reference & TN query) and with the thrush nightingale genome as the reference and the common nightingale genome as the query (TN reference & CN query) split up by type and size. DEL stands for deletions, DUP for duplications, INV for inversions, INS for insertions and BND for translocation breakpoints and SVs longer than 100 kb.

	CN reference & TN query					**TN reference & CN query**				
Length	DEL	DUP	INV	INS	BND	DEL	DUP	INV	INS	BND
50-100 bp	15 078	4	0	14 236	490	13 850	13	0	14 805	1 158
100-1000 bp	20 424	19	10	19 132		18 593	14	12	20 035	
1-10 kb	1 256	2	10	1 024		1 042	10	16	1 246	
> 10 kb	64	3	0	30		35	4	1	59	
Total	**36 822**	**28**	**20**	**34 422**	**490**	**33 520**	**41**	**29**	**36 145**	**1 158**

5.3 Creating a high confidence set of structural variants

5.3.1 Choosing the merging parameter value

Parameter denoting the distance between two breakpoints for variants called by different programs to be merged in the percentage of length was iterated over values between 0 and 1 and the total numbers of variants in the final datasets were plotted. From these results, a value of 0.25 (25%) was selected as a threshold for SVs merging. By this process, it was ensured that potential true variants were not neglected by selecting a threshold in the area of the steep cline (Figure 7).

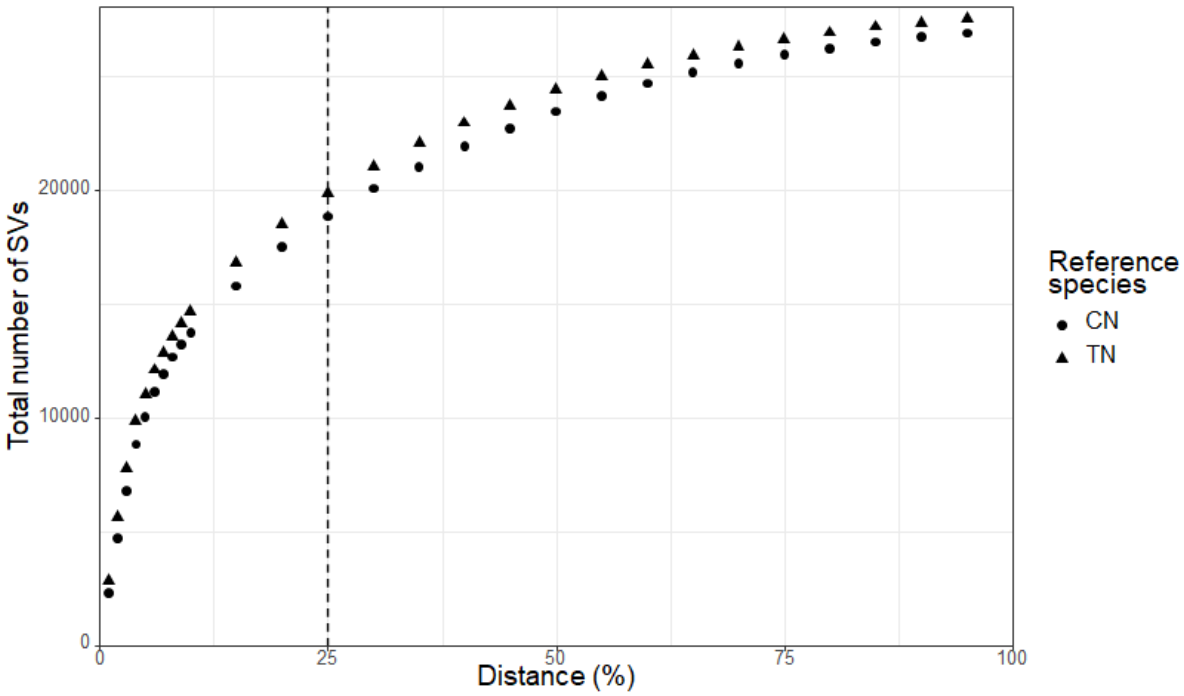


Figure 7: Graph of the total number of merged variants called against the common nightingale reference genome and against the thrush nightingale reference genome for the variable value of merging parameter.

5.3.2 High-confidence set of variants

For each reference assembly in interspecific comparisons, a high-confidence set of variants was created by merging the outputs of all three aforementioned SV callers. The numbers of consensus variant calls are summed up in Table 5. In the CN reference & TN query analysis, a total number of 18 839 overlapping variant calls was detected. In the TN reference & CN query analysis, it was about 1 000 variants more (19 863). Regardless of the

direction of the analysis, deletions were the most prevalent SV type with 10 819 hits in the CN reference and 10 346 hits in the TN reference, closely followed by insertions with 7 967 and 9 414 hits, respectively. 7 and 19 inversions were identified in the CN reference & TN query and TN reference & CN query comparison, respectively. In comparison with individual SV callers' results, the number of translocation breakpoints greatly decreased to 36 candidates in the CN reference & TN query and 76 in the TN reference & CN query.

Table 5: Number of SV calls in the high-confidence dataset called from the thrush nightingale data against the common nightingale reference genome (CN reference & TN query) and from the common nightingale data against the thrush nightingale genome (TN reference & CN query) split up by type and size. DEL stands for deletions, DUP for duplications, INV for inversions, INS for insertions and BND for translocation breakpoints.

	CN reference & TN query					**TN reference & CN query**				
Length	DEL	DUP	INV	INS	BND	DEL	DUP	INV	INS	BND
50-100 bp	3 954	0	0	3 046	36	3 599	0	0	3 557	76
100-1000 bp	6 303	5	2	4 528		6 192	4	6	5 408	
1-10 kb	553	4	5	392		553	2	10	449	
> 10 kb	9	1	0	1		2	2	3	0	
Total	**10 819**	**10**	**7**	**7 967**	**36**	**10 346**	**8**	**19**	**9 414**	**76**

5.4 Closer inspection of inversions polymorphic between species

The high-confidence set of structural variant calls from SURVIVOR contained 7 variants identified as inversions for the CN reference & TN query analysis (Table 6) and 19 for the TN reference & CN query analysis (Table 7). Variants were assigned IDs A-G (the CN reference & TN query callset) and 1-19 (the TN reference & CN query callset). Apart from variants marked as inversions by SURVIVOR, also translocation breakpoints were checked for the presence of an intrachromosomal rearrangement (which might have occurred as a result of SVIM and SVIM-asm calling variants longer than 100 kb as breakpoints), however, no extra putative inversions were found. None of the identified putative inversions from either merged dataset was localized on the Z sex chromosome (scaffold 5 in the common nightingale and scaffold 8 in the thrush nightingale) (Table 1, Table 6, Table 7).

Closer inspection of high-confidence inversions using IGV and assessment of concordance between individual merged calls revealed that out of the 19 inversions in the TN reference & CN query, SURVIVOR ranked inversions 2 and 9 by mistake due to a low-support SVIM inversion call sharing the same position as a universally merged deletion call. Additionally, variants 1, 6 and 17 were ruled out as false positives due to their properties. Variant 1 overlaps a low coverage region at the end of scaffold 1 (Supplementary Figure 1) and variants 6 and 17 are delimited by immense coverage differences. For all of these three variants, there are no reads whose mapping would support the existence of an inversion at their breakpoints. The rest of the variants seemed like true inversions (in the context of reads and genome alignments with TN reference).

By expectation, true inversions fixed between species should be present in both datasets. Coordinates of the inversions from one species were converted to the coordinate system of the other species using the generated PAF alignments. This conversion was straightforward from the alignment except for 3 problematic variants 1, 6 and 17 (Supplementary Table 3). Based on the translated coordinates, the following variants were paired between the two callsets: A and 5, D and 10, E and 11, F and 12, and G and 14. For the location of each variant in one of the final datasets, the presence/absence of an inversion in each of the intermediate interspecific and intraspecific comparisons was assessed (Supplementary Table S3). These inversions could potentially play a role in nightingale speciation.

Table 6: Variants from the high confidence dataset called against the common nightingale assembly identified as inversions by SURVIVOR. Variants identified as true inversions by closer examination are highlighted in bold.

ID	Scaffold	Start position (bp)[1]	Interval of start position (bp)[2]	End position (bp)[1]	Interval of end position (bp)[2]	Average predicted length (bp)
A	**1**	**73 416 546**	**[-15, 5]**	**73 417 896**	**[-9, 0]**	**1 350**
B	1	97 990 313	[-5, 0]	97 991 186	[-2, 6]	876
C	8	19 077 482	[0, 1]	19 082 345	[0, 0]	4 863
D	**9**	**22 146 872**	**[0, 3]**	**22 148 043**	**[-17, 0]**	**1 149**
E	**9**	**24 244 431**	**[-3, 0]**	**24 246 871**	**[-1, 1]**	**2 441**
F	**10**	**5 859 092**	**[-10, 0]**	**5 859 398**	**[-1, 9]**	**313**
G	**17**	**13 690 667**	**[0, 2]**	**13 692 688**	**[-3, 0]**	**2 441**

[1] As called by SURVIVOR from the breakpoint positions in three intermediate callsets (1-based).
[2] Range of the identified breakpoint positions by different SV callers relative to the value in Start/End position column.

Table 7: Variants from the high confidence dataset called against the thrush nightingale assembly identified as inversions by SURVIVOR. Variants in italics are filtered out false positives. Variants identified as true inversions by closer examination are highlighted in bold.

ID	Scaffold	Start position (bp)[1]	Interval of start position (bp)[2]	End position (bp)[1]	Interval of end position (bp)[2]	Average predicted length (bp)
1	*1*	*111 922*	*[0, 678]*	*115 525*	*[0, 750]*	*3 672*
2	*1*	*52 925 041*	*[-3, 717]*	*52 937 288*	*[-167, 0]*	*11 933*
3	1	53 554 411	[-1, 1]	53 554 970	[0, 17]	567
4	1	101 212 465	[0, 1]	101 217 146	[-1, 0]	4 680
5	**2**	**72 702 377**	**[-3, 0]**	**72 703 704**	**[0, 3]**	**1 329**
6	*3*	*1 010 068*	*[0, 2]*	*12 126 194*	*[0, 1]*	*11 116 126*
7	**3**	**81 944 194**	**[-74, 0]**	**81 945 366**	**[0, 143]**	**1 279**
8	**5**	**45 861 823**	**[-6, 3]**	**45 863 398**	**[0, 5]**	**1 579**
9	*7*	*27 989 240*	*[0, 320]*	*27 994 295*	*[-486, 0]*	*4 518*
10	**9**	**22 293 947**	**[0, 1]**	**22 295 102**	**[-55, 0]**	**1 136**
11	**9**	**24 390 642**	**[-2, 0]**	**24 393 088**	**[0, 6]**	**2 450**
12	**10**	**29 606 047**	**[-13, 0]**	**29 606 364**	**[-2, 0]**	**323**
13	**14**	**7 960 175**	**[0, 0]**	**7 960 645**	**[-2, 0]**	**469**
14	**17**	**6 422 555**	**[-4, 0]**	**6 424 573**	**[-1, 1]**	**2 019**
15	24	2 505 543	[-1, 8]	2 505 837	[-2, 0]	291
16	**27**	**3 917 603**	**[-2, 9]**	**3 918 936**	**[-16, 0]**	**1 325**
17	*28*	*2 163 137*	*[-1, 248]*	*3 424 984*	*[-232 354, 3]*	*1 261 850*
18	30	138 279	[-6, 1]	138 731	[-31, 0]	443
19	42	45 589	[-6, 4]	46 311	[-7, 4]	722

[1] As called by SURVIVOR from the breakpoint positions in three intermediate callsets (1-based).
[2] Range of the identified breakpoint positions by different SV callers relative to the value in Start/End position column.

Variant A(5) is probably heterozygous in the TN, as it is supported by approximately half of the reads in both the intraspecific TN comparisons and the interspecific comparisons against the CN reference. Results for variant D(10) suggest that it is a true fixed inversion, as it is not called in any of the intraspecific comparisons and the entirety of reads supports this rearrangement in all the interspecific reads-to-genome analyses. Variant E(11) is likely polymorphic in the CN, as it was identified as a homozygous inversion in the CN reference & TN query part of the analysis, while the numbers of supporting reads in the TN reference & CN query and CN reference & CN query suggest heterozygosity. And finally, variants F(12) and G(14) were not called by any intraspecific comparison and sequencing data support their existence well, hinting that these are also likely fixed inversions between species.

As for the unpaired variants B and C in the CN reference & TN query part of the analysis (i.e. variants identified by all three SV callers in the CN reference & TN query analysis, but not in the TN reference & CN query analysis), variant B was detected also in the CN intraspecific comparison by both SV callers. As it is localized towards the end of the scaffold (the entire length of assembled scaffold 1 is approximately 112 Mbp), i.e. in a hard-to-assemble area, it might be an error in the assembly, or, alternatively, an error in the mapping. That is supported by the fact that the variants called in intraspecific comparisons are supported by the entirety of mapping reads. Of all individual callsets created in the TN reference & CN query, only the genome-to-genome approach discovered this variant, which is supportive of this call not being a true between-species polymorphism. Variant C is supported by approximately half of the ONT reads from TN in CN reference & TN query interspecific comparison, suggesting that the sequenced TN individual might be heterozygous for this variant. This variant was, however, not called in the TN intraspecific comparison by any caller, which contradicts this possibility. Further inspection revealed that this variant was also called by SVIM-asm and SVIM in the TN reference & CN query analysis, but not in intraspecific CN analysis.

Unpaired variants from the TN reference & CN query high-confidence dataset (i.e. variants called by all three SV callers in the TN reference & CN query analysis, but not in the CN reference & TN query analysis) include variants 3, 4, 7, 8, 13, 15, 16, 18 and 19. Variant 3 is likely an error in the assembly, as it is called by an entirety of reads in the TN reference & CN query as well as in the TN reference & TN query. The same goes for the variants 15, 18 and 19, although the support of the reads as called by SVIM is not completely

unanimous (Supplementary Table 3). Missassembly would not be surprising particularly for variants 18 and 19 located in close proximity to the ends of respective scaffolds. This is also the case for variant 4 that spans the very end of scaffold 1 in the TN reference assembly. Variants 7, 8 and 16 are probably false negative calls from the CN reference & TN query pipeline, as the visual inspection of alignments in the respective areas suggests the existence of such variants (Supplementary Figures 2, 3 and 4). These variants were not convincingly called in any of the intraspecific comparisons and are not visible from the visualisation of intraspecific alignments and thus can be true between the species. Variant 13 is likely a heterozygous site in the sequenced TN individual with support below the threshold needed to be called by Sniffles in the intraspecific comparison. It therefore might be a case of a true inversion that is heterozygous in at least one of the species, although the support in the CN reference & TN query alignments is scarce (Supplementary Figure 5).

In total, we identified 9 inversions polymorphic between species. 5 of them were discovered in both CN reference & TN query and TN reference & CN query analyses and 4 of them were discovered only in the TN reference & CN query analyses, however, it turned out they were false negatives in the analysis in the other direction. Lengths of recognized inversions range between 300 bp and 2 500 bp (Table 7). 6 inversions called in the high-confidence datasets turned out to be artefacts caused probably by errors in the assembly.

6. Discussion

Interspecies comparisons with the common nightingale (CN) as the reference and the thrush nightingale (TN) as query identified 47 047 SVs called by Sniffles, 702 952 SVs called by SVIM and 71 782 SVs called by SVIM-asm. Analyses in the other direction (i.e. with TN as reference and CN as query) yielded 46 430 SVs called by Sniffles, 643 754 SVs called by SVIM and 70 893 SVs called by SVIM-asm. The intersection of the intermediate SV callsets resulted in high-confidence sets of 18 839 variants in the TN against the CN reference and 19 864 variants in the CN against the TN reference. Insertions and deletions shorter than 1 000 bp, common sources of variation in genomes, account for over 94% of both these datasets. There were 7 inversions in the CN reference high-confidence dataset and 19 inversions in the TN reference one. After closer inspection of these, 9 true between-species inversions were identified with lengths ranging between 300 bp and 2 500 bp. 5 inversions from the high-confidence datasets were assessed as false positives and 6 as possible errors in the assembly.

From the comparison of performances of individual SV callers in the CN reference & TN query and TN reference & CN query approaches it is evident that the choice of the reference and query species considerably affects the SV calling results. Probably the most striking difference between the two directions of the analysis is in the numbers of called translocation breakpoints in the interspecific comparisons by Sniffles and SVIM. The almost eight-fold higher number of breakpoint calls in the SVIM CN reference & TN query analysis compared to the number of breakpoint calls in the SVIM TN reference & CN query analysis is likely a consequence of the substantially higher number of reads of shorter lengths sequenced from the TN individual than from the CN individual. The numbers of calls reported by SVIM represent results unfiltered by any threshold score. The higher amount of shorter reads can produce more erroneous inter-scaffold split-reads mappings (i.e. in repetitive elements) which then can result in more translocation breakpoint calls with little support in the data. However, the results of this thesis also provide evidence that different properties of the sequencing data are not the only source of unambiguity in SV calling. The number of calls from the comparisons conducted by SVIM-asm (*de novo* assembly-based approach) from genome-to-genome alignments are also not in complete concordance between the CN reference & TN query and TN reference & CN query analyses, even though the data provided to the SV caller were essentially the same. This highlights the benefits of

conducting the analyses in both directions to discover the full scope of variation, which is however not a widely adopted approach.

In total, the intersection of results from different SV callers yielded 5 inversions that were called in both directions of the analysis (i.e. both in the CN reference & TN query analysis and CN reference & TN query analysis). Notably, it is presumable that the set of inversions described here is not exhaustive. While the outcomes of a high-confidence analysis workflow created by intersecting intermediate results from diverse SV callers are most straightforward and greatly reduce the number of false positive hits, the sensitivity of such an approach is substantially lower than when considering a union of all intermediate callsets (De Coster et al., 2019). An affirmation to that is also the fact that apart from the 5 inversions identified by both merged high-confidence datasets we also discovered 4 that were not called by all SV callers, but still represent true inversions.

Although some evidence suggests that inversions are commonly present on the sex chromosomes of closely related bird species (Hooper et al., 2019; Hooper & Price, 2015), none of the putative inversions in the high-confidence dataset is localized on the Z sex chromosome. Nevertheless, given the workflow aiming at high confidence of called variants, it might be the case that potentially true inversions on sex chromosomes were not called by all three SV callers and therefore were filtered out in the process. The coverage by sequencing reads on the sex chromosomes is half of the coverage of autosomes, therefore it might be the case that potential inversions on Z did not pass the filter for the minimum number of supportive reads in Sniffles and therefore are not included in the final high-confidence callset. The eventuality of inversions on the Z chromosome playing a role in the evolution of the common nightingale and the thrush nightingale species thus cannot be definitively ruled out.

None of the 9 identified true inversions surpasses 2 500 bp in length. That diminishes the probability of their influence on meiotic pairing and/or recombination suppression across multiple genes, as rearrangements need to span larger areas (several Mb) to cause serious problems with the assembly of synaptonemal complexes (Kirkpatrick, 2010). Although the detected inversions are not likely to cause the reproductive isolation between the common and the thrush nightingale directly, their effect might still have contributed to the speciation process. Inversion breakpoints, if located strategically, can disrupt a reading frame or cause a change in gene expression, as was shown for example in *Arabidopsis thaliana*

(Tsuchimatsu et al., 2010) or in *Drosophila buzzatii* (Puig et al., 2004), and in that way affect phenotype directly, possibly creating an adaptive mutation, on which selection can later act on (Kirkpatrick, 2010). To fully explore this possibility, it would be beneficial to annotate the genomes and decide whether the identified inversions overlap with protein-coding areas.

The approach to the identification of inversion polymorphisms between two species adapted in this thesis proved to be suitable. Including the intraspecific comparisons in the process have also been found useful, as they made it possible to filter out errors in the assembly and distinguish heterozygous inversions. It would be useful to conduct analysis in a similar way for other SV types, including deletions, duplications and translocations.

7. Conclusion

This thesis investigated structural variation between two nightingale species, the common nightingale (*Luscinia megarhynchos*) and the thrush nightingale (*Luscinia luscinia*). Our results suggest that SVs in genomes are common and therefore their further investigation is needed. Although the high-confidence set of variants constructed here shows a promising start, more thorough analyses of the full range of SVs between these two species across multiple individuals are needed. Among consensus variants identified by the combination of long-read mapping-based and *de novo* assembly-based approaches, we identified 9 inversions. These could potentially play a role in the speciation of nightingales. However, the short length of the inversions excludes them from the possibility of promoting speciation by suppressing recombination or causing problems with meiotic pairing. A possible way, in which they could contribute to the speciation, is by disrupting a reading frame or by enforcing a change in gene expression. These alternatives should be studied in the future in the context of genome annotation.

The conclusions that can be derived from the findings of this thesis are burdened by the low number of compared individuals, making it difficult to distinguish between true fixed variants and within-population polymorphisms. However, the initial findings reported here provide a useful insight into the matter and a solid ground for more robust research of structural variation between the two closely related nightingale species in the future.

8. Literature

Allen, E., Xie, Z., Gustafson, A. M., Sung, G.-H., Spatafora, J. W., & Carrington, J. C. (2004). Evolution of microRNA genes by inverted duplication of target gene sequences in Arabidopsis thaliana. *Nature Genetics*, *36*(12), 1282–1290.

Balachandran, P., & Beck, C. R. (2020). Structural variant identification and characterization. *Chromosome Research : An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology*, *28*(1), 31–47.

Bridges, C. B. (1936). The Bar "Gene" a Duplication. *Science*, *83*(2148), 210–211.

Cameron, D. L., Schröder, J., Penington, J. S., Do, H., Molania, R., Dobrovic, A., Speed, T. P., & Papenfuss, A. T. (2017). GRIDSS: Sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. *Genome Research*, *27*(12), 2050–2060.

Catanach, A., Crowhurst, R., Deng, C., David, C., Bernatchez, L., & Wellenreuther, M. (2019). The genomic pool of standing structural variation outnumbers single nucleotide polymorphism by threefold in the marine teleost Chrysophrys auratus. *Molecular Ecology*, *28*(6), 1210–1223.

Chen, J.-M., Chuzhanova, N., Stenson, P. D., Férec, C., & Cooper, D. N. (2005). Complex gene rearrangements caused by serial replication slippage. *Human Mutation*, *26*(2), 125–134.

Chen, K., Wallis, J. W., McLellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., McGrath, S. D., Wendl, M. C., Zhang, Q., Locke, D. P., Shi, X., Fulton, R. S., Ley, T. J., Wilson, R. K., Ding, L., & Mardis, E. R. (2009). BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. *Nature Methods*, *6*(9), 677–681.

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A. J., Kruglyak, S., & Saunders, C. T. (2016). Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics (Oxford, England)*, *32*(8), 1220–1222.

Damas, J., O'Connor, R., Griffin, D., & Larkin, D. (2019). *Avian Chromosomal Evolution* (pp. 69–92).

Daněček, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008.

De Coster, W., De Rijk, P., De Roeck, A., De Pooter, T., D'Hert, S., Strazisar, M.,

Sleegers, K., & Van Broeckhoven, C. (2019). Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. *Genome Research*, *29*(7), 1178–1187.

De Coster, W., Weissensteiner, M. H., & Sedlazeck, F. J. (2021). Towards population-scale long-read sequencing. *Nature Reviews. Genetics*, 1–16.

Delcher, A. L., Kasif, S., Fleischmann, R. D., Peterson, J., White, O., & Salzberg, S. L. (1999). Alignment of whole genomes. *Nucleic Acids Research*, *27*(11), 2369–2376.

Dobzhansky, Th., & Sturtevant, A. H. (1938). Inversions in the Chromosomes of Drosophila Pseudoobscura. *Genetics*, *23*(1), 28–64.

Dresser, M. E., Ewing, D. J., Harwell, S. N., Coody, D., & Conrad, M. N. (1994). Nonhomologous Synapsis and Reduced Crossing over in a Heterozygous Paracentric Inversion in Saccharomyces Cerevisiae. *Genetics*, *138*(3), 633–647.

Ederveen, A., Lai, Y., van Driel, M. A., Gerats, T., & Peters, J. L. (2015). Modulating crossover positioning by introducing large structural changes in chromosomes. *BMC Genomics*, *16*, 89.

English, A. C., Salerno, W. J., & Reid, J. G. (2014). PBHoney: Identifying genomic variants via long-read discordance and interrupted mapping. *BMC Bioinformatics*, *15*(1), 180.

Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. *Nature Reviews. Genetics*, *7*(2), 85–97.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., & Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, *151*(4), 1531–1545.

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333–351.

Gu, W., Zhang, F., & Lupski, J. R. (2008). Mechanisms for human genomic rearrangements. *PathoGenetics*, *1*(1), 4.

Guerrero, R. F., Rousset, F., & Kirkpatrick, M. (2012). Coalescent patterns for chromosomal inversions in divergent populations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1587), 430–438.

Haldane, J. B. S. (1922). Sex ratio and unisexual sterility in hybrid animals. *Journal of Genetics*, *12*(2), 101–109.

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*(1), e1000327.

Heller, D., & Vingron, M. (2019). SVIM: Structural variant identification using mapped long reads. *Bioinformatics*, *35*(17), 2907–2915.

Heller, D., & Vingron, M. (2020). SVIM-asm: Structural variant detection from haploid and diploid genome assemblies. *Bioinformatics (Oxford, England)*, btaa1034.

Hermetz, K. E., Newman, S., Conneely, K. N., Martin, C. L., Ballif, B. C., Shaffer, L. G., Cody, J. D., & Rudd, M. K. (2014). Large Inverted Duplications in the Human Genome Form via a Fold-Back Mechanism. *PLoS Genetics*, *10*(1), e1004139.

Hooper, D. M., Griffith, S. C., & Price, T. D. (2019). Sex chromosome inversions enforce reproductive isolation across an avian hybrid zone. *Molecular Ecology*, *28*(6), 1246–1262.

Hooper, D. M., & Price, T. D. (2015). Rates of karyotypic evolution in Estrildid finches differ between island and continental clades. *Evolution*, *69*(4), 890–903.

Hooper, D. M., & Price, T. D. (2017). Chromosomal inversion differences correlate with range overlap in passerine birds. *Nature Ecology & Evolution*, *1*(10), 1526–1534.

Huddleston, J., Chaisson, M. J. P., Steinberg, K. M., Warren, W., Hoekzema, K., Gordon, D., Graves-Lindsay, T. A., Munson, K. M., Kronenberg, Z. N., Vives, L., Peluso, P., Boitano, M., Chin, C.-S., Korlach, J., Wilson, R. K., & Eichler, E. E. (2017). Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Research*, *27*(5), 677–685.

Jacobs, P., Brown, W. M. C., Baikie, A. G., & Strong, J. A. (1959). The somatic chromosomes in mongolism. *The Lancet*, *273*(7075), 710.

Jeffares, D. C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., Balloux, F., Dessimoz, C., Bähler, J., & Sedlazeck, F. J. (2017). Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature Communications*, *8*(1), 14061.

Johnson, N. A. (2008). Hybrid incompatibility and speciation. *Nature Education*, *1(1)*(20).

Joron, M., Frezal, L., Jones, R. T., Chamberlain, N. L., Lee, S. F., Haag, C. R., Whibley, A., Becuwe, M., Baxter, S. W., Ferguson, L., Wilkinson, P. A., Salazar, C., Davidson, C., Clark, R., Quail, M. A., Beasley, H., Glithero, R., Lloyd, C., Sims, S., … ffrench-Constant, R. H. (2011). Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*, *477*(7363), 203–206.

Kirkpatrick, M. (2010). How and Why Chromosome Inversions Evolve. *PLoS Biology*, *8*(9), e1000501.

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, *37*(5), 540–546.

Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M., & Kamatani, Y. (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biology*, *20*, 117.

Kulathinal, R. J., Stevison, L. S., & Noor, M. A. F. (2009). The Genomics of Speciation in Drosophila: Diversity, Divergence, and Introgression Estimated Using Low-Coverage Genome Sequencing. *PLoS Genetics*, *5*(7), e1000550.

Küpper, C., Stocks, M., Risse, J. E., Dos Remedios, N., Farrell, L. L., McRae, S. B., Morgan, T. C., Karlionova, N., Pinchuk, P., Verkuil, Y. I., Kitaysky, A. S., Wingfield, J. C., Piersma, T., Zeng, K., Slate, J., Blaxter, M., Lank, D. B., & Burke, T. (2016). A supergene determines highly divergent male reproductive morphs in the ruff. *Nature Genetics*, *48*(1), 79–83.

Kverek, P., Reifova (Storchova), R., Reif, J., & MW, N. (2008). Occurrence of a hybrid between the Common Nightingale (Luscinia megarhynchos) and the Thrush Nightingale (Luscinia luscinia) in the Czech Republic confirmed by genetic analysis. *Sylvia*, *44*, 17–26.

Lai, Z., Nakazato, T., Salmaso, M., Burke, J. M., Tang, S., Knapp, S. J., & Rieseberg, L. H. (2005). Extensive Chromosomal Repatterning and the Evolution of Sterility Barriers in Hybrid Sunflower Species. *Genetics*, *171*(1), 291–303.

Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: A probabilistic framework for structural variant discovery. *Genome Biology*, *15*(6), R84.

Lee, J. A., Carvalho, C. M. B., & Lupski, J. R. (2007). A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders. *Cell*, *131*(7), 1235–1247.

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100.

Liu, Y., Zhang, M., Sun, J., Chang, W., Sun, M., Zhang, S., & Wu, J. (2020). Comparison of multiple algorithms to reliably detect structural variants in pears. *BMC Genomics*, *21*(1), 61.

Lowry, D. B., & Willis, J. H. (2010). A Widespread Chromosomal Inversion Polymorphism Contributes to a Major Life-History Transition, Local Adaptation, and Reproductive Isolation. *PLOS Biology*, *8*(9), e1000500.

Lupski, J. R. (1998). Genomic disorders: Structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends in Genetics: TIG*, *14*(10), 417–422.

Lynch, M., & Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics*, *154*(1), 459–473.

Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., & Sedlazeck, F. J. (2019). Structural variant calling: The long and the short of it. *Genome Biology*, *20*(1), 246.

Manoukis, N. C., Powell, J. R., Touré, M. B., Sacko, A., Edillo, F. E., Coulibaly, M. B., Traoré, S. F., Taylor, C. E., & Besansky, N. J. (2008). A test of the chromosomal theory of ecotypic speciation in Anopheles gambiae. *Proceedings of the National Academy of Sciences*, *105*(8), 2940–2945.

Masly, J. P., Jones, C. D., Noor, M. A. F., Locke, J., & Orr, H. A. (2006). Gene transposition as a cause of hybrid sterility in Drosophila. *Science (New York, N.Y.)*, *313*(5792), 1448–1450.

McClintock, B. (1931). *Cytological observations of deficiencies involving known genes, translocations and an inversion in Zea mays*. University of Missouri, College of Agriculture, Agricultural Experiment Station.

McVey, M., & Lee, S. E. (2008). MMEJ repair of double-strand breaks (director's cut): Deleted sequences and alternative endings. *Trends in Genetics : TIG*, *24*(11), 529–538.

Mohiyuddin, M., Mu, J. C., Li, J., Bani Asadi, N., Gerstein, M. B., Abyzov, A., Wong, W. H., & Lam, H. Y. K. (2015). MetaSV: An accurate and integrative structural-variant caller for next generation sequencing. *Bioinformatics*, *31*(16), 2741–2744.

Moore, J. K., & Haber, J. E. (1996). Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in Saccharomyces cerevisiae. *Molecular and Cellular Biology*, *16*(5), 2164–2173.

Nattestad, M., & Schatz, M. C. (2016). Assemblytics: A web analytics tool for the detection of variants from an assembly. *Bioinformatics (Oxford, England)*, *32*(19), 3021–3023.

Nowell, P. C., & Hungerford, D. A. (1960). Chromosome Studies on Normal and Leukemic Human Leukocytes. *JNCI: Journal of the National Cancer Institute*, *25*(1), 85–109.

Ottaviani, D., LeCain, M., & Sheer, D. (2014). The role of microhomology in genomic structural variation. *Trends in Genetics: TIG*, *30*(3), 85–94.

Pang, A. W., MacDonald, J. R., Pinto, D., Wei, J., Rafiq, M. A., Conrad, D. F., Park, H., Hurles, M. E., Lee, C., Venter, J. C., Kirkness, E. F., Levy, S., Feuk, L., & Scherer, S. W. (2010). Towards a comprehensive structural variation map of an individual human genome. *Genome Biology*, *11*(5), R52.

Pellestor, F. (2019). Chromoanagenesis: Cataclysms behind complex chromosomal rearrangements. *Molecular Cytogenetics*, *12*, 6.

Puig, M., Cáceres, M., & Ruiz, A. (2004). Silencing of a gene adjacent to the breakpoint of a widespread Drosophila inversion by a transposon-induced antisense RNA. *Proceedings of the National Academy of Sciences*, *101*(24), 9013–9018.

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., & Korbel, J. O. (2012). DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, *28*(18), i333–i339.

Reifová, R., Kverek, P., & Reif, J. (2011). The first record of a female hybrid between the Common Nightingale (Luscinia megarhynchos) and the Thrush Nightingale (Luscinia luscinia) in nature. *Journal of Ornithology*, *152*(4), 1063–1068.

Rieseberg, L. H. (2001). Chromosomal rearrangements and speciation. *Trends in Ecology & Evolution*, *16*(7), 351–358.

Rieseberg, L. H., Whitton, J., & Gardner, K. (1999). Hybrid Zones and the Genetic Architecture of a Barrier to Gene Flow Between Two Sunflower Species. *Genetics*, *152*(2), 713–727.

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R. F., Wilkie, A. O. M., McVean, G., & Lunter, G. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics*, *46*(8), 912–918.

Robberecht, C., Voet, T., Esteki, M. Z., Nowakowska, B. A., & Vermeesch, J. R. (2013). Nonallelic homologous recombination between retrotransposable elements is a driver of de novo unbalanced translocations. *Genome Research*, *23*(3), 411–418.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24–26.

RStudio Team. (2020). *RStudio: Integrated Development for R.* PBC, Boston, MA. http://www.rstudio.com/

Sedlazeck, F. J., Lee, H., Darby, C. A., & Schatz, M. C. (2018). Piercing the dark matter: Bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics*, *19*(6), 329–346.

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., & Schatz, M. C. (2018). Accurate detection of complex structural variations using single molecule sequencing. *Nature Methods*, *15*(6), 461–468.

Stewart, C., Kural, D., Strömberg, M. P., Walker, J. A., Konkel, M. K., Stütz, A. M., Urban, A. E., Grubert, F., Lam, H. Y. K., Lee, W.-P., Busby, M., Indap, A. R.,

Garrison, E., Huff, C., Xing, J., Snyder, M. P., Jorde, L. B., Batzer, M. A., Korbel, J. O., … Project, 1000 Genomes. (2011). A Comprehensive Map of Mobile Element Insertion Polymorphisms in Humans. *PLOS Genetics*, *7*(8), e1002236.

Storchová, R., Reif, J., & Nachman, M. W. (2010). Female heterogamety and speciation: Reduced introgression of the Z chromosome between two species of nightingales. *Evolution*, *64*(2), 456–471.

Sturtevant, A. H. (1917). Genetic Factors Affecting the Strength of Linkage in Drosophila. *Proceedings of the National Academy of Sciences*, *3*(9), 555–558.

Sturtevant, A. H. (1921). A Case of Rearrangement of Genes in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, *7*(8), 235–237.

Sturtevant, A. H., & Mather, K. (1938). The Interrelations of Inversions, Heterosis and Recombination. *The American Naturalist*, *72*(742), 447–452.

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M. H.-Y., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., Casale, F. P., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., … Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*(7571), 75–81.

Todesco, M., Owens, G. L., Bercovich, N., Légaré, J.-S., Soudi, S., Burge, D. O., Huang, K., Ostevik, K. L., Drummond, E. B. M., Imerovski, I., Lande, K., Pascual-Robles, M. A., Nanavati, M., Jahani, M., Cheung, W., Staton, S. E., Muños, S., Nielsen, R., Donovan, L. A., … Rieseberg, L. H. (2020). Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature*, *584*(7822), 602–607.

Tsuchimatsu, T., Suwabe, K., Shimizu-Inatsugi, R., Isokawa, S., Pavlidis, P., Städler, T., Suzuki, G., Takayama, S., Watanabe, M., & Shimizu, K. K. (2010). Evolution of self-compatibility in Arabidopsis by a mutation in the male specificity gene. *Nature*, *464*(7293), 1342–1346.

Van't Hof, A. E., Campagne, P., Rigden, D. J., Yung, C. J., Lingley, J., Quail, M. A., Hall, N., Darby, A. C., & Saccheri, I. J. (2016). The industrial melanism mutation in British peppered moths is a transposable element. *Nature*, *534*(7605), 102–105.

Viguera, E., Canceill, D., & Ehrlich, S. D. (2001). Replication slippage involves DNA polymerase pausing and dissociation. *The EMBO Journal*, *20*(10), 2587–2595.

Villoutreix, R., Carvalho, C. F. de, Soria-Carrasco, V., Lindtke, D., De-la-Mora, M., Muschick, M., Feder, J. L., Parchman, T. L., Gompert, Z., & Nosil, P. (2020). Large-scale mutation in the evolution of a gene complex for cryptic coloration. *Science*, *369*(6502), 460–466.

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*(7434), 664–668.

Weissensteiner, M. H., Bunikis, I., Catalán, A., Francoijs, K.-J., Knief, U., Heim, W., Peona, V., Pophaly, S. D., Sedlazeck, F. J., Suh, A., Warmuth, V. M., & Wolf, J. B. W. (2020). Discovery and population genomics of structural variation in a songbird genus. *Nature Communications*, *11*(1), 3403.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org

Ye, K., Schulz, M. H., Long, Q., Apweiler, R., & Ning, Z. (2009). Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics (Oxford, England)*, *25*(21), 2865–2871.

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*(7), 849–853.

Zhang, L., Reifová, R., Halenková, Z., & Gompert, Z. (2021). How Important Are Structural Variants for Speciation? *Genes*, *12*(7), 1084.

Zuellig, M. P., & Sweigart, A. L. (2018). A two-locus hybrid incompatibility is widespread, polymorphic, and active in natural populations of Mimulus*. *Evolution*, *72*(11), 2394–2405.

9. Supplementary data

Supplementary Table 1: Number of SV calls by Sniffles with the common nightingale genome as the reference and the common nightingale sequencing data as the query (CN reference & CN query) and with the thrush nightingale as the reference and the thrush nightingale sequencing data as the query (TN reference & TN query) split up by type and size. In the CN reference & CN query, one variant with an unresolved type between a deletion and an inversion is reported here as a deletion. In the TN reference & TN query analysis, one variant of an unresolved type between a deletion and an inversion is reported here as a deletion. DEL stands for deletions, DUP for duplications, INV for inversions, INS for insertions and BND for translocation breakpoints.

	CN reference & CN query					**TN reference & TN query**				
Length	DEL	DUP	INV	INS	BND	DEL	DUP	INV	INS	BND
50-100 bp	1 759	0	0	2 922	4 045	1 964	0	0	3 983	6 669
100-1000 bp	2 771	7	8	3 288		2 816	5	11	3 514	
1-10 kb	339	30	14	217		276	45	19	179	
> 10 kb	63	51	77	2		64	78	90	3	
Total	**4 932**	**88**	**99**	**6 429**	**4 045**	**5 120**	**128**	**120**	**7 679**	**6 669**

Supplementary Table 2: Number of SV calls by SVIM with the common nightingale genome as the reference and the common nightingale sequencing data as the query (CN reference & CN query) and with the thrush nightingale as the reference and the thrush nightingale sequencing data as the query (TN reference & TN query) split up by type and size. Numbers in brackets are the counts after applying a modest filter to filter out the calls with the value of the score computed by SVIM below 5. DEL stands for deletions, DUP for duplications, INV for inversions, INS for insertions and BND for translocation breakpoints and SVs longer than 100 kb.

	CN reference & CN query					**TN reference & TN query**				
Length	DEL	DUP	INV	INS	BND	DEL	DUP	INV	INS	BND
50-100 bp	164 850 (6 068)	54 (0)	4 (0)	52 421 (11 276)	47 316 (2 832)	60 155 (7 064)	64 (1)	6 (0)	75 975 (13 862)	228 894 (5 668)
100-1000 bp	161 209 (8 132)	667 (14)	128 (8)	63 760 (12 057)		63 064 (9 372)	1 050 (23)	184 (8)	71 292 (13 431)	
1-10 kb	3 733 (549)	508 (69)	101 (4)	2 833 (548)		2 439 (477)	797 (98)	183 (6)	2 410 (455)	
> 10 kb	109 (13)	332 (48)	62 (0)	56 (8)		147 (15)	639 (72)	210 (0)	80 (2)	
Total	**329 901 (14 762)**	**1 561 (131)**	**295 (12)**	**119 070 (23 889)**	**47 316 (2 832)**	**125 805 (16 928)**	**2 550 (194)**	**583 (14)**	**149 757 (27 750)**	**228 894 (5 668)**

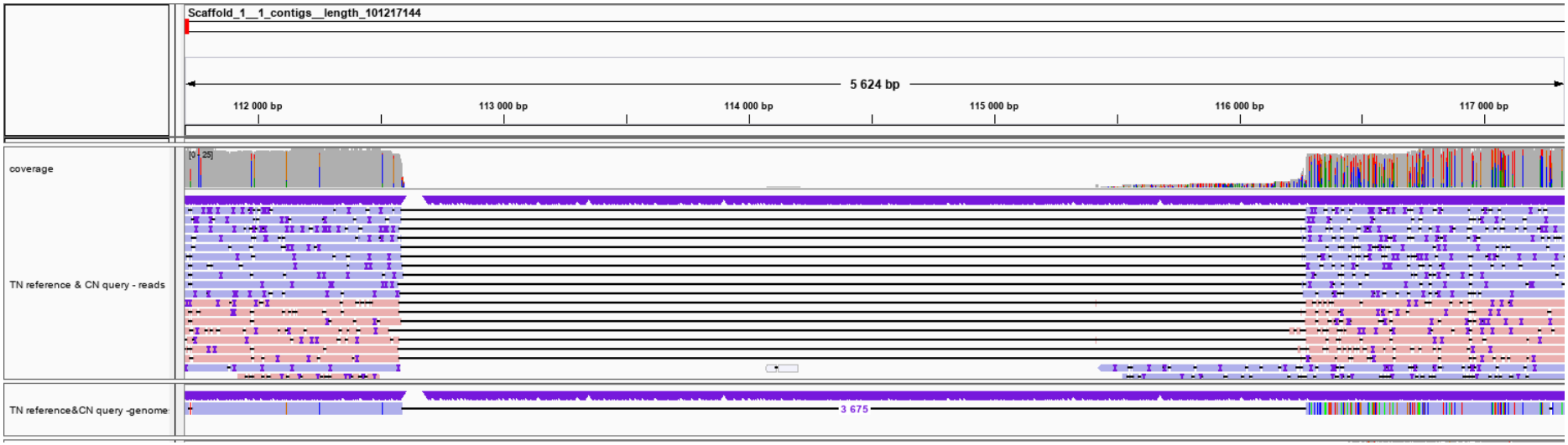
Supplementary Table 3: Inversions from high-confidence datasets and their support by different SV calling pipelines.Variants identified as true inversions in closer examination are highlighted in bold.

	common nightingale					thrush nightingale					Intraspecific CN ref. & CN q.		Intraspecific TN ref. & TN q.		Interspecific CN ref. & TN q.			Interspecific TN ref. & CN q.		
	Scaf[1]	Len (Mb)[2]	Start (bp)[1]	End (bp)[1]	Avg. cov.[3]	Scaf [1]	Len (Mb)[2]	Start (bp)[1]	End (bp)[1]	Avg. cov.[3]	SVIM[4]	Sniffles [5]	SVIM[4]	Sniffles[5]	SVIM[4]	Sniffles [5]	SVIM -asm[6]	SVIM[4]	Sniffles [5]	SVIM -asm[6]
A	**1**	**111.7**	**73 416 550**	**73 417 895**	**~26x**	**2**	**111.4**	**72 702 376**	**72 703 701**	**~17x**	**0**	**0**	**18:9**	**10:14**	**15:11**	**9:13**	**1**	**13**	**0:20**	**1**
B	1	111.7	97 990 307	97 991 192	~28x	2	111.4	97 243 491	97 244 377	~28x	0:25	0:29	0	0	3:20	0:23	1	0	0	1
C	8	71.3	19 077 482	19 082 345	~25x	5	71.4	52 073 273	52 078 136	~26x	0	0	0	0	9:11	13:11	1	22:6	0	1
D	**9**	**37.9**	**22 146 910**	**22 148 042**	**~26x**	**9**	**37.7**	**22 293 947**	**22 295 085**	**~24x**	**0**	**0**	**0**	**0**	**0:27**	**0:30**	**1**	**22**	**0:25**	**1**
E	**9**	**37.9**	**24 244 427**	**24 246 871**	**~20x**	**9**	**37.7**	**24 390 642**	**24 393 090**	**~22x**	**13:16**	**0:15**	**0**	**0**	**0:23**	**0:26**	**1**	**11:13**	**2:14**	**1**
F	**10**	**35.7**	**5 859 081**	**5 859 404**	**~18x**	**10**	**35.1**	**29 606 041**	**29 606 364**	**~23x**	**0**	**0**	**0**	**0**	**19**	**0:20**	**1**	**0:23**	**0:25**	**1**
G	**17**	**20.2**	**13 690 667**	**13 692 687**	**~19x**	**17**	**20.0**	**6 422 552**	**6 424 572**	**~21x**	**0**	**0**	**0**	**0**	**14**	**0:20**	**1**	**10**	**0:19**	**1**
3	2	101.3	47 650 593	47 651 149	~22x	1	101.2	53 554 409	53 554 965	~19x	0	0	0:18	0:24	0	0:00	1	0:16	0:21	1
4	2	101.3	1	4 578	~14x	1	101.2	101 212 565	101 217 145	~16x	0	0	0	0	0	0	0	8	0:16	1
5	see A																			
7	**3**	**89.3**	**7 306 429**	**7 307 645**	**~22x**	**3**	**89.0**	**81 944 149**	**81 945 365**	**~19x**	**0**	**0**	**0**	**0**	**4**	**5:25**	**0**	**1:13**	**0:23**	**1**
8	**8**	**71.3**	**25 275 486**	**25 277 039**	**~12x**	**5**	**71.4**	**45 861 817**	**45 863 401**	**~18x**	**0**	**0**	**0**	**0**	**7**	**0**	**0**	**4:13**	**0:15**	**1**
10	see D																			
11	see E																			
12	see F																			
13	**14**	**20.8**	**7 984 347**	**7 984 817**	**~15x**	**14**	**20.6**	**7 960 174**	**7 960 644**	**~16x**	**0**	**0**	**3**	**0**	**2**	**0**	**1**	**1:14**	**0:16**	**1**
14	see G																			
15	24	11.1	8 565 213	8 565 498	~23x	24	11.0	2 505 550	2 505 835	~15x	0	0	1:20	0:21	0	0	1	0:14	0:15	1
16	**28**	**7.2**	**1 847 927**	**1 849 250**	**~10x**	**27**	**7.0**	**3 917 611**	**3 918 935**	**~14x**	**0**	**0**	**1**	**0**	**2**	**0**	**1**	**6**	**0:14**	**1**
18	25	8.0	110 527	110 985	~19x	30	6.5	138 272	138 730	~17x	0	0	6:11	0:18	0	0	1	4:14	0:16	1
19	39	0.8	488 764	489 497	~23x	42	0.3	45 582	46 315	~32x	0	0	5:14	0:16	0	0	0	3:20	0:22	1

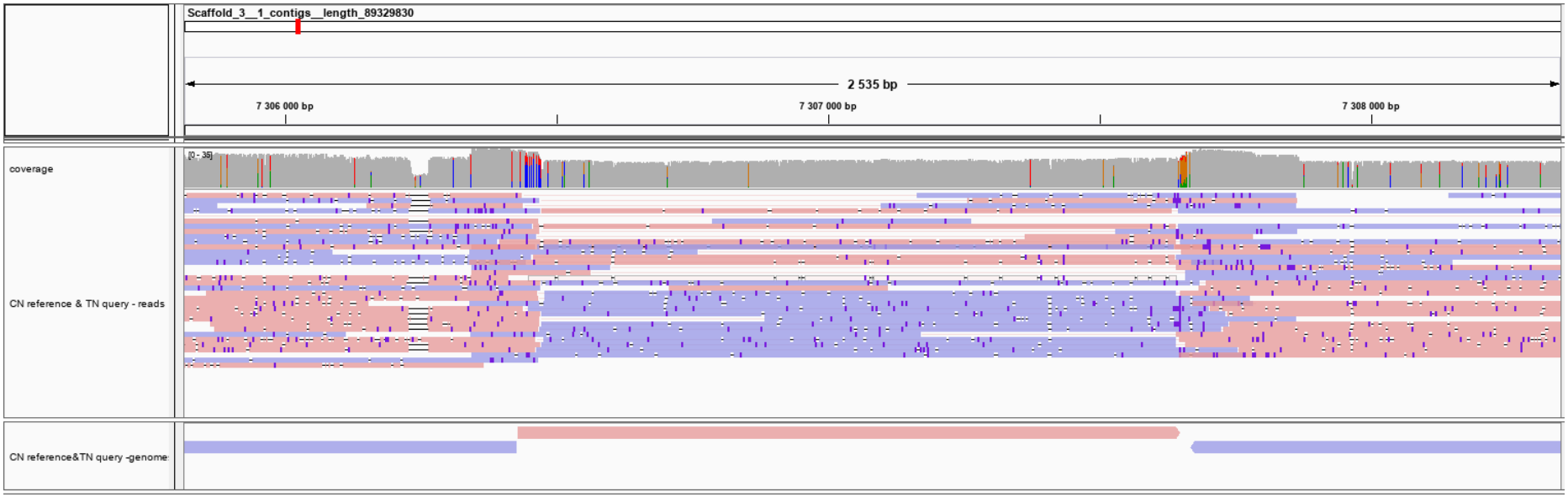
[1] Range of the alignment line spanning the inversion specified by the scaffold, start position and end position (0-based) and range of its mapped counterpart from the other species.

[2] Length of the respective scaffold.
[3] Average coverage of the respective area.
[4] Numbers of supportive reads as called by SVIM. 0 stands for a variant not called in the respective dataset at this position. Records containing colon (e.g. 0:25) represent the number of reads supporting the reference vs. the number of reads supporting the variant in properly called inversions with reads supporting both breakpoints. Non-zero records without colon represent the number of supporting reads for not properly called inversions (i.e. without sufficient support for one of its breakpoints).
[5] Numbers of supportive reads as called by Sniffles. 0 stands for a variant not called in the respective dataset at this position. Non-zero records represent the number of supporting reads for reference and for the variant, respectively.
[6] 1 if there is a variant call in the respective dataset at this position, 0 otherwise.

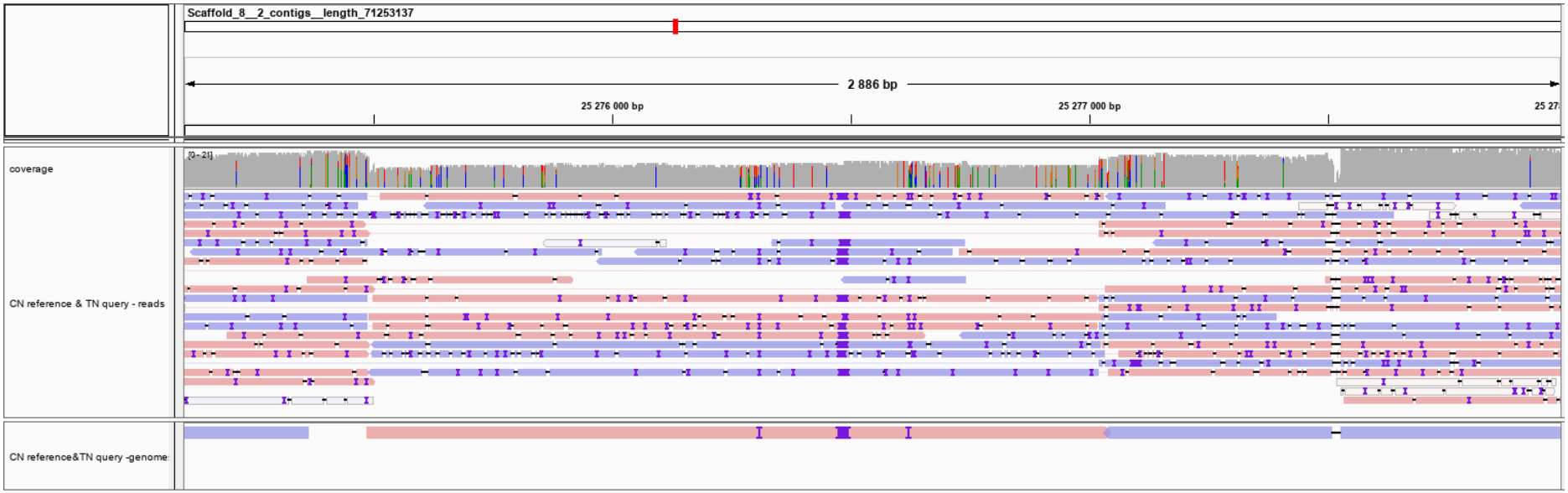
Supplementary Figure 1: Variant 1 (scaffold 1: 111 922 - 115 525 in the thrush nightingale assembly) overlaps a low coverage region in both the alignment of common nightingale sequencing data to the thrush nightingale genome (top track) and in the alignment of common nightingale assembly to the thrush nightingale assembly (bottom track). The image was generated using IGV. Sequences are colored by the orientation in which they map.



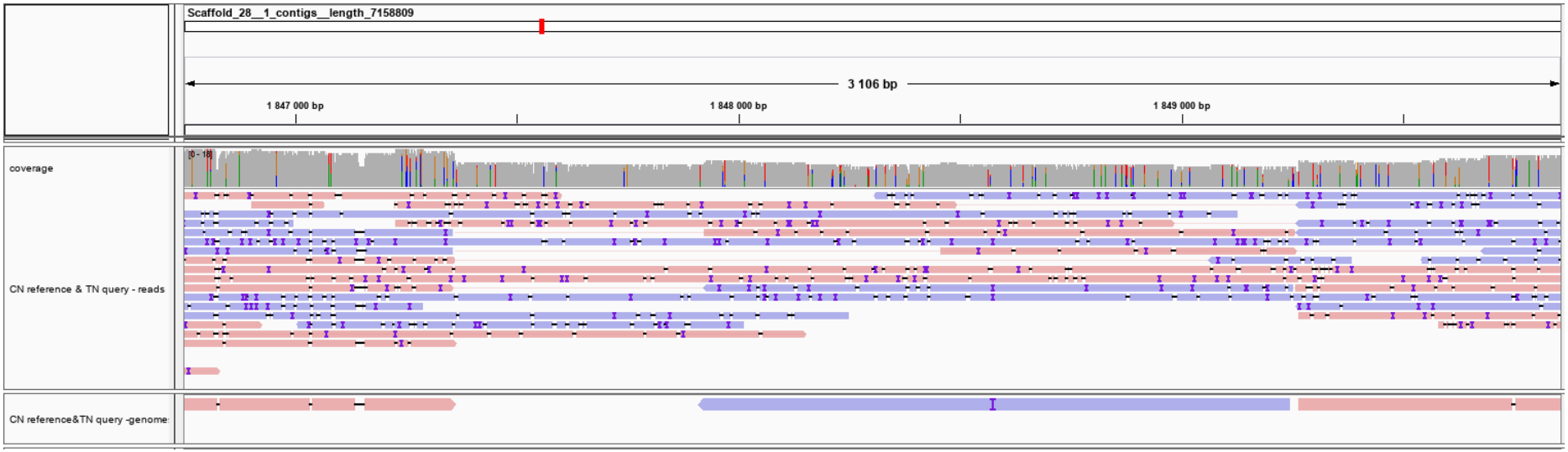
Supplementary Figure 2: Variant 7 (scaffold 3: ~7 306 429 - ~7 307 645 in the common nightingale assembly). Even though variant 7 was not called by all three SV callers in the CN reference & TN query interspecific comparison, the alignment of the TN reads (track at the top) and the TN genome (track at the bottom) to the CN reference suggests its presence. The image was generated using IGV. Sequences are colored by the orientation in which they map.



Supplementary Figure 3: Variant 8 (scaffold 8: ~25 275 486 - ~25 277 039 in the common nightingale assembly). Although variant 8 was not called by all three SV callers in the CN reference & TN query interspecific comparison, the alignment of the TN reads (track at the top) and the TN genome (track at the bottom) to the CN reference suggests its presence. The image was generated using IGV. Sequences are colored by the orientation in which they map.

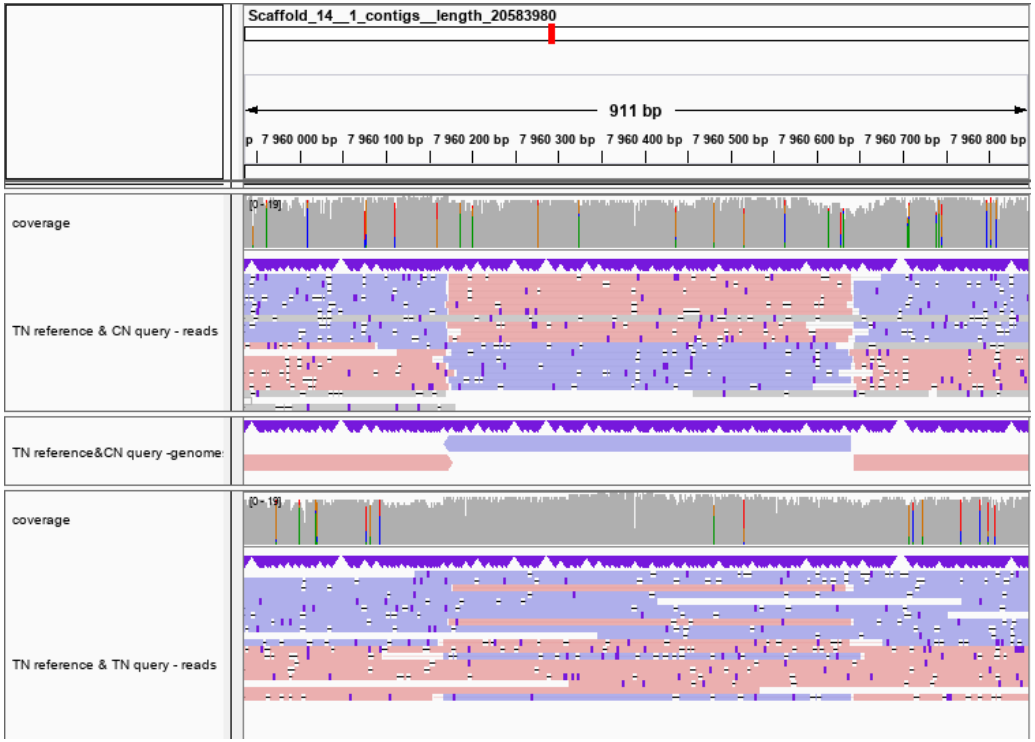


Supplementary Figure 4: Variant 16 (scaffold 28: ~1 847 927 - ~1 849 250 in the common nightingale assembly). Although variant 16 was not called by all three SV callers in the CN reference & TN query interspecific comparison, the alignment of the TN reads (track at the top) and the TN genome (track at the bottom) to the CN reference suggests its presence. The image was generated using IGV. Sequences are colored by the orientation in which they map.



Supplementary Figure 5: Variant 13 (scaffold 14: ~7 984 347 - ~7 984 817 in the common nightingale assembly and scaffold 14: ~7 960 174 - ~7 960 644 in the thrush nightingale assembly). (A) Alignment of CN reads (top track), TN genome (track in the middle) and TN reads (bottom track) to the TN reference genome. (B) Alignment of TN reads (top track), CN genome (track in the middle) and CN reads (bottom track) to the CN reference genome. Variant 13 was confidently called from the TN reference & CN query comparisons (top and middle tracks in figure A), however, from the intraspecific comparison (bottom track in figure A) it is clear, that it is a heterozygous site in TN, which is also partially noticeable in the alignment of reads in CN reference & TN query comparison (top track in figure B). The image was generated using IGV. Sequences are colored by the orientation in which they map.

A:



B: