UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1762

Evolutionary genomics in Corvids

– *From single nucleotides to structural variants*

MATTHIAS H. WEISSENSTEINER

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2019

ISSN 1651-6214
ISBN 978-91-513-0550-9
urn:nbn:se:uu:diva-369878

Dissertation presented at Uppsala University to be publicly examined in Ekmansalen, Norbyvägen 14 A, Uppsala, Sunday, 25 February 2018 at 13:00 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Professor Kateryna M. Makova (Department of Biology, Penn State University, USA).

Abstract
Weissensteiner, M. H. 2019. Evolutionary genomics in Corvids – From single nucleotides to structural variants. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 1762. 42 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-0550-9.

Heritable genetic variation is the raw material of evolution and can occur in many different forms, from altering single nucleotides to rearranging stretches of millions at once. DNA mutations that result in phenotypic differences are the basis upon which natural selection can act, leading to a shift of the frequency of those mutations.

In this thesis I aim to comprehensively characterize and quantify genetic variation in a natural system, the songbird genus *Corvus*.

First, we expand on previous work from a hybrid zone of different populations of Eurasian crows. All black carrion crows and black-and-grey hooded crows meet in a narrow hybrid zone in central Europe, and also in central and Southeast Asia. Comparing population genetic data acquired from these three hybrid zones yielded no single genetic region as a candidate responsible for phenotypic divergence, yet a parallelism in sets of genes and gene networks was evident.

Second, we capitalize on varying evolutionary timescales to investigate the driver of the heterogeneous genetic differentiation landscape observed in multiple avian species. Genetic diversity, and thus differentiation, seems to be correlated both between populations within single species and between species which diverged 50 million years ago. This pattern is best explained by conserved broad-scale recombination rate variation, which is in turn likely associated with chromosomal features such as centromeres and telomeres.

Third, we introduce a *de-novo* assembly of the hooded crow based on long-read sequencing and optical mapping. The use of this technology allowed a glimpse into previously hidden regions of the genome, and uncovered large-scale tandem repeat arrays consisting of a 14-kbp satellite repeat or its 1.2-kpb subunit. Furthermore, these tandem repeat arrays are associated with regions of reduced recombination rate.

Lastly, we extend the population genetic analysis to structural genomic variation, such as insertions and deletions. A large-scale population re-sequencing data set based on short-read and long-read technologies, spread across the entire genus is the foundation of a fine-scale genome-wide map of structural variation. A differentiation outlier approach between all-black carrion and black-and-grey hooded crows identified a 2.25-kilobase LTR retrotransposon inserted 20-kb upstream of the *NDP* gene. The element, which is fixed in the hooded crow population, is associated with decreased expression of *NDP* and may be responsible for differences in plumage color.

Keywords: evolutionary genetics, genomics, population genetics, selection, recombination, chromosomal features, colouration, insertion, deletion, inversion, crow, tandem repeat, transposable element, gene expression

Matthias H. Weissensteiner, Department of Ecology and Genetics, Evolutionary Biology, Norbyvägen 18D, Uppsala University, SE-75236 Uppsala, Sweden.

ISSN 1651-6214
ISBN 978-91-513-0550-9
urn:nbn:se:uu:diva-369878 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-369878)

Für Bernadette, Kilian und Emil

List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I Vijay, N., Bossu, C.M., Poelstra, J.W., Weissensteiner, M.H., Suh, A., Kryukov, A.P. & Wolf J.B.W. (2016) Evolution of heterogeneous genome differentiation across multiple contact zones in a crow species complex. *Nature Communications,* 7: 13195

II Vijay, N., Weissensteiner, M.H., Burri, R., Kawakami, T., Ellegren, H. & Wolf J.B.W. (2017) Genome-wide patterns of variation in genetic diversity are shared among populations, species and higher-order taxa. *Molecular Ecology*, 26, no. 16: 4284-4295

III Weissensteiner, M.H., Pang, A.W.C., Bunikis, I., Höijer, I., Vinnere-Pettersson, O., Suh, A. & Wolf J.B.W. (2017) Combination of short-read, long-read and optical mapping assemblies reveals large-scale tandem repeat arrays with population genetic implications. *Genome Research*, 27: 697-708

IV Weissensteiner, M.H., Bunikis, I., Knief, U., Peona, V., Pophaly, S.D., Suh, A., Sedlazeck, F.J., Warmuth, V.M. & Wolf, J.B.W. Fine-scale analysis of structural genomic variation in natural populations. *Manuscript*

Additional Papers

The following papers were published during the course of my doctoral studies but are not part of this thesis.

Shafer, A. B. A., Wolf, J. B. W., Alves, P. C., Bergström, L., Bruford, M. W., Brännström, I., Colling, G., Dalen, L., Meester, L.D., Ekblom, R., Fawcett, K.D., Fior, S., Hajibabaei, M., Hill, J.A., Hoezel, A.R., Höglund, J., Jensen, E.L., Krause, J., Kristensen, T.N., Krützen, M., McKay, J.K., Norman, A.J., Ogden, R., Österling, E.M., Ouborg, J.N., Piccolo, J., Popovic, D., Primmer C.R., Reed, F.A., Roumet, M., Salmona, J., Schenekar, T., Schwartz, M.K., Segelbacher, G., Senn, H., Thaulow, J., Valtonen, M., Veale, A., Vergeer, P. Vijay, N., Vilá, C., Weissensteiner, M.H., Wennerström, L., Wheat, C.W. & Zieliński, P. (2015). Genomics and the challenging translation into conservation practice. *Trends in Ecology and Evolution*, 30(2), 78–87.

Weissensteiner, M.H., Poelstra, J.W. and Wolf, J.B., 2015. Low-budget ready-to-fly unmanned aerial vehicles: An effective tool for evaluating the nesting status of canopy-breeding bird species. *Journal of Avian Biology*, 46(4), 425-430.

Shafer, A. B. A., Wolf, J. B. W., Alves, P. C., Bergström, L., Brännström, I., Colling, G., Dalen, L., Meester, L.D., Ekblom, R., Fior, S., Hajibabaei, Hoezel, A.R., Höglund, J., Jensen, E.L., Norman, A.J., Ogden, R., Österling, E.M., Ouborg, J.N., Piccolo, J., Primmer C.R., Reed, F.A., Roumet, M., Salmona, J., Schwartz, M.K., Segelbacher, G., Thaulow, J., Valtonen, M., Vergeer, P., Vilá, C., Weissensteiner, M.H., Wheat, C.W. & Zieliński, P. (2016). Reply to Garner *et al.*. *Trends in Ecology and Evolution*, 31(2), 83–84.

Peona, V., Weissensteiner, M.H. & Suh, A. (2018). How complete are "complete" genome assemblies?—An avian perspective. *Molecular Ecology Resources*, 18(6), 1188-1195.

Contents

Abbreviations

bp, kb, Mb	base pairs, Kilo base, Mega base
DNA	deoxyribonucleic acid
Mya	million years ago
SNP	single nucleotide polymorphism
SV	structural variation
SMRT	single-molecule real-time
LTR	long terminal repeat
LINE	long interspersed element
ERV	endogenous retrovirus
LD	linkage disequilibrium
LR	long read
SR	short read
OM	optical map
RAM	repetitive anchored map

1. Introduction

Genetic variation is the raw material for evolution. Individual differences in heritable units – collectively referred to as 'genomes' – can lead to the manifestation of different phenotypic traits, affecting the outer appearance, behavior or physiology of an organism. These differences are in turn the substrate upon which *selection* can act, leading to differentially adapted individuals, and a change in frequency of the underlying genotypes in populations. Selection however is not the only force capable of changing the composition of genetic variation. Genetic drift, essentially the process of random sampling, also drives changes in genetic variation. Migration on the other hand changes genotype frequencies by introducing individuals from other populations and admixing them with the population in focus. Lastly, there is one obvious mechanism influencing genetic variation: mutation, the creation of entirely new genotypes (Hartl and Clark 1997).

Depending on a variety of parameters, these four forces influence the observed genetic variation very differently. Among the most important parameters is the type of the observed variation. Genomes can vary on many different levels; from point mutations which only exchange one nucleotide for another, to genome duplications which double sometimes billions of base pairs. Knowing this, it seems obvious that our image of genetic variation is biased, a fact that is strongly driven by available technologies. In the beginning, our observations were enabled by the invention of light microscopes, limited to karyotypic variation, i.e. differences in chromosome number and size (Schwann 1847). About the same time, genetic variation was observed indirectly through the investigation of phenotypes (Bateson and Mendel 1913), an approach which also led to the discovery of heritability (Morgan 1919). Through improvements in microscope technology, predictions of genetic variation got more and more refined, leading to the discovery of chromosomal rearrangements – that is, the insertion, deletion, translocation or inversion of a large part of the chromosome. Later on, phenotypic traits could eventually be mapped to narrow regions on chromosomes, without knowing the actual sequence or structure of the DNA (Sturtevant 1913).

The field radically changed with the invention Sanger sequencing in the nineteen seventies (Sanger, Nicklen, and Coulson 1977), which then allowed the determination of the actual nucleotide sequence of the DNA. Consequently, single nucleotide polymorphisms (SNPs) of individual genes became the most commonly studied form of genetic variation for more than thirty

years, with the noticeable exception of microsatellites, small tandemly repeated DNA sequences which expand and contract rapidly leading to increased allelic diversity (Litt and Luty 1989; H. Ellegren 2004).

With the advent of high-throughput sequencing technologies in the early 2000s, SNPs became even more popular (and feasible) to study because of the sudden availability of whole genome sequences for multiple individuals per species, yielding thousands or even millions of informative SNP loci (Goodwin, McPherson, and McCombie 2016). Roughly fifteen years later, we are now again at the verge of a major shift in the field of genomic research. Technological advance has made it possible to study structural genomic variation, which was first discovered long before, at a new, base pair-level resolution (Sedlazeck et al. 2018). Catalyzed by the availability of new methods, this type of variation is becoming more and more recognized as an important contributor to phenotypic variation.

In this thesis, I follow this transition by employing a variety of technologies to answer a central question of biology: What is genetic variation, which evolutionary processes shape its distribution across the genome, and how is genetic variation linked to observed phenotypes?

2. Types of Genetic Variation

2.1 Single nucleotide variation

Replication machinery in living cells is imperfect, and errors occur when DNA is replicated. One such form of errors are point mutations which change a single letter (or nucleotide) in the DNA sequence. These mutations have some interesting properties. Besides the rate being variable between species and even between different genomic regions within species (Baer, Miyamoto, and Denver 2007) , certain nucleotide alterations happen more frequently than others. If an exchange happens between purine bases (adenine (A) to guanine (G) or vice versa) or pyrimidine bases (cytosine (C) and with a thymine (T) or vice versa), this is called a transition, whereas the replacement of a purine base (A or G) with a pyrimidine base (C or T) (or vice versa) is a transversion. Due to biochemical features of the DNA molecule, transitions occur more frequently than transversions (Lynch 2007).

There are further features by which single-nucleotide polymorphisms (SNPs) are differentiated. An important one is the position in the codon triplet, which leads to the distinction of synonymous and non-synonymous changes. These terms apply to SNPs occurring in genic regions and refer to the effect on the resulting protein: While synonymous changes do not affect the sequence of amino acids (and are thus silent mutations), non-synonymous changes alter the amino acid sequence and thus likely change the resulting protein structure (Graur and Li 2000). It is somewhat obvious that these features lead to different population genetic trajectories. Since non-synonymous SNPs likely change the phenotype of an organism, they are subject to selection, thus they cannot be considered as evolving neutrally. Synonymous changes on the other hand are usually under much lower selective constraints, and can thus be used for inferences of population history.

In contemporary genomic studies, SNPs are usually the first choice when it comes to investigations of genetic diversity in any given organism. Besides the availability of well-developed tools, also the underlying theoretical expectations are well established (Brookes 1999).

2.2 Structural variation

Genetic variation that affects a whole segment of DNA at once is referred to as structural variation. The precise definitions have been changing over time and are mainly concerning the size range of a variant, which is also somewhat arbitrary. Feuk et al. (2006) classify structural variation into microscopic (> 3 Mb) and sub-microscopic (1 kb to 3 Mb), while Alkan et al. (2011) consider every alteration larger than 50 bp as a structural variant. Although technically everything which is not a SNP could be considered as structural variation, it is advisable to adhere to certain standards for the purpose of comparability. Thus, throughout this thesis, I will follow the definition established by Alkan et al. (2011) and treat alterations larger than 50 bp as structural variants.

Depending on the way the DNA sequence is altered, different types of structural variation can be distinguished. If, compared to a reference, a sequence of nucleotides is inserted at a certain locus, this is classified as an insertion. Conditional on the identity of the inserted sequence, there are two subtypes of insertions, namely novel sequence insertions and transpositions. The latter usually refer to interspersed repeat insertions and will be discussed in more detail below. Essentially, duplications (either in tandem or interspersed) and genes with variable copy number could also be considered as insertions, but due to overlapping definitions they are usually treated separately (Redon et al. 2006; Sharp et al. 2005). When a stretch of DNA is excised, this is considered a deletion. As evident from the description of insertions above, it is rather a matter of perspective, whether a variant is treated as an insertion or a deletion. This is particularly true if no further investigation of the features of the concerned sequence is carried out, because knowledge on the underlying mutational mechanism (e.g. interspersed repeat insertion or tandem repeat contraction or expansion) can be used to elucidate the evolutionary history of a variant. Given the wealth of structural variation commonly occurring in genomes, such a detailed analysis is often not feasible however, and in these cases, insertions and deletions are analyzed together. A third class of structural variation are inversions. Here, a segment of DNA is inverted compared to the reference. Because there is no gain or loss of sequence, inversions are, together with reciprocal deletions and insertions, recognized as balanced polymorphisms (Feuk, Carson, and Scherer 2006).

There are different types of selection that can act upon SVs. First, structural mutations can change the protein encoded by a given gene by altering the amino acid sequence via a frameshift mutation or a change in the nucleotide sequence. Because this usually entails a negative phenotypic effect for the carrier, these kinds of mutations are best studied in (human) genetic diseases (Feuk, Carson, and Scherer 2006). Another, arguably more common way of SVs influencing phenotypes is by altering the expression of a gene or set of genes nearby. The underlying mechanisms involve positional effects (e.g. in-

versions changing the order of genes on a chromosome and thus their chromatin environment affecting expression) (Weiler and Wakimoto 1995), change of the number of copies of a gene expressed (a famous example is the human amylase gene which duplicated presumably in response to a different diet (Perry et al. 2007)) or modifying the promoters and enhancers of genes (Chuong, Elde, and Feschotte 2017). Since structural variations affect a much larger amount of genomic sequence, it is not surprising that they are in general considered to be more deleterious than SNPs. That is not to say that SVs are necessarily always deleterious, but while synonymous SNPs at four-fold degenerate sites and many SNPs in non-coding DNA can evolve effectively neutral, the possibilities for structural mutations to occur where they neither change the phenotype nor impact genome stability of an organism are likely much rarer. It can thus be expected that the allele frequency distribution of SVs segregating in populations will differ from that of SNPs, although this is likely dependent on the class and size of variants.

3. Repetitive DNA

Repetitive DNA has often been ignored or neglected in genomic studies, either because it is difficult to assess or because it is considered (neutrally evolving) junk. It is important to note though that repeats account for a substantial proportion of structural genetic variation and should thus be recognized in evolutionary genetic studies.

3.1 Interspersed Repeats

Repetitive DNA occurs in two flavors: interspersed and tandemly repeated. Interspersed repeats denote any kind of DNA sequence (of a certain length) which is not unique in the genome, with individual copies scattered across the genome (Lee and Langley 2010; Smit 1996). Depending on the type of the repetitive element, they can exist from very few to thousands of copies within a single genome. An important feature of some types of interspersed repeats is that they are capable of moving around in the genome, thus the often-used alternative term 'mobile genetic element'. Two major groups have been characterized so far: Transposable elements (TEs) (Kazazian Jr. 2004) and endogenous viral elements (EVEs) (Katzourakis and Gifford 2010). TEs are separated into two distinct classes, which are based on their means of mobilization. Class I TEs, which include retrotransposons and retroviruses, proliferate via a copy-and-paste mechanism that involves reverse transcription and complementary DNA integration at a novel location (Levin and Moran 2011). Within class I TEs (retrotransposons) there are three main subgroups, namely long interspersed elements (LINEs) (Luan et al. 1993; Martin 2006), short interspersed elements (SINEs) (Wicker et al. 2007; Kapitonov and Jurka 2006) and long terminal repeat (LTR) retrotransposons (including retroviruses) (Kazazian Jr. 2004). While LINEs and LTR retrotransposons both carry the sequence encoding their retrotransposition machinery (reverse transcriptase for target-primed reverse transcription in case of LINEs and several different proteins for replicative retrotransposition in LTR retrotransposons), SINEs lack a transposition machinery of their own and essentially hijack the reverse transcriptase protein of LINEs. Class II TEs (also known as DNA transposons) can move via a cut-and-paste mechanism, essentially translocating their own DNA sequence to another location (Kidwell 2005). A key feature of DNA transposons are terminal inverted repeats (TIRs) and target site duplications

(TSDs) (Wicker et al. 2007). Finally, endogenous viral elements (EVEs) constitute relics of viral infections of the host and subsequent integration in the germline genome (Feschotte and Gilbert 2012). Most commonly, EVEs are retroviruses which are reverse-transcribed and integrated into the host genome via retrotransposition. Thus they can essentially be classified as LTR retrotransposons, although there exist various EVEs of non-retroviral origin (Feschotte and Gilbert 2012).

3.2 Tandem repeats

Tandem repeats occur when the same stretch of DNA (with varying degrees of sequence identity) is repeated in a head-to-tail fashion, forming an array. The different classes of tandem repeats are mainly differentiated by the length of the single repeat unit which ranges from a one base pair to several thousands of base pairs. Among the smallest tandem repeats are homopolymers (a single nucleotide repeated many times) and microsatellites. The latter, which are also called simple sequence repeats, usually exhibit unit sizes of 5 to 25 bp and arrays consisting of up to hundred copies (Litt and Luty 1989). They belong to the most rapidly evolving sequences in the genome, rapidly expanding and contracting via replication slippage, the formation of a hairpin loop during replication that leads to the deletion or insertion of several copies in the array (H. Ellegren 2004; Schlötterer and Tautz 1992). The high mutation rate of these repeats has been utilized by a number of research areas, with DNA fingerprinting among the most important ones. Because microsatellite loci acquire new mutations so rapidly, they can be used to uniquely identify individuals for parental testing. Likewise, microsatellites have been the marker of choice in population genetic studies prior to the rise of high-throughput sequencing technologies and catalyzed the emergence of the field of molecular ecology (Schlötterer 2004). Larger types of tandem repeats comprise minisatellites and satellites. Minisatellites have also been popular in individual genotyping, in fact DNA fingerprinting was first described using minisatellites (Jeffreys, Wilson, and Thein 1985). They exhibit similar features as microsatellites, except a slightly larger unit and array size (López-Flores and Garrido-Ramos 2012). Satellites are the least studied fraction of tandem repeats. Partly this is due to their size: single repeat units of several kilobases and megabase-scale arrays have been prohibitive to their exploration (Plohl 2010). Only recently it has become possible to generate sequence data spanning the length of a single repeat unit, yet the entirety of the array is still often beyond the capabilities of present-day technologies (Miga 2015).

4. Meiotic Recombination and Linkage Disequilibrium

Meiotic recombination is a genetic mechanism ubiquitous in sexually reproducing organisms (Graur and Li 2000). It describes the process of shuffling alleles during germ cell proliferation and is an important mediator of natural selection (Cutter and Payseur 2013) and source of genetic variation (Hans Ellegren and Galtier 2016). During meiosis, homologous chromosomes align with each other, and breaks in the double-stranded DNA occur. These double-strand breaks occur non-randomly and are, depending on the organism, clustered in certain genomic regions (Massy 2013). Double-strand breaks can result in crossing-overs, the exchange of DNA strands of homologous chromosomes, or be resolved as non-crossing-overs. The former is the actual process that leads to different allelic combinations or haplotypes, increasing the efficacy of selection for individual loci (Cutter and Payseur 2013).

Meiotic recombination can be studied and quantified in different ways. The first is the direct observation of crossing-overs on chromosomes. However, one of the most common ways to infer the recombination rate is by generating a linkage map. Genotypes of genetic markers are determined for parents and offspring over several generations to trace the inherited allelic combinations. Given sufficiently many meiosis (i.e. offspring from the same parents or generations) and a dense marker regime, the entire genome can be mapped and assigned to linkage groups (Graw 2015). Lastly, there is another, indirect, measure of recombination: the population-scaled recombination rate ρ. This approach relies on information from phased genotypes of a number of individuals of a population and capitalizes on deviations of allelic combinations (Wall 2000). If a certain allelic combination is more common than expected from random assortment according to their respective allele frequencies, it is said to be in linkage disequilibrium (LD) (Lewontin and Kojima 1960). The term essentially describes the statistical association between two loci and should not be confused with physical linkage, which is the mere presence of two loci on the same chromosome. The fact that recombination erodes the statistical association between two loci can be used to infer its rate based on observed patterns in LD, with the limitation that other forces such as selection or demographic processes also influence LD (Baird 2015).

An important factor influencing recombination rate is structural variation. If large stretches of DNA are altered between homologous chromosomes, this

can cause difficulties in pairing and subsequent chromosome mis-segregation leading to infertility or inviability of the organism (Lupski and Stankiewicz 2005). Thus, strong selection can be expected to either remove deleterious variants, or to alter the recombination rate regime to avoid such consequences.

5. Corvids in the songird genus *Corvus*

Corvids are an enigmatic group of songbirds. They have been accompanying humans since prehistoric times, and their extraordinary intelligence and sophisticated social behavior have fascinated scientists and laymen alike. The 40 to 44 species in the genus *Corvus*, which comprise crows, ravens and jackdaws, are distributed on all continents except Antarctica and South America and range from medium to large size (Del Hoyo et al. 2015; Jønsson et al. 2016). Phenotypically they are rather uniform with most species similar in shape and an all-black plumage color. However, a few species exhibit a peculiar black-and-white or black-and-grey pattern, which is scattered across the *Corvus* phylogeny, usually together with an all-black sister species. The most prominent example for such a black versus black-and-grey species pair is the case of the hooded *C. (corone*) *cornix* and carrion crow *C. (corone) corone*. They meet in a narrow hybrid zone in central Europe which has served as a textbook example for post-glacial processes (Meise 1928; Mayr 1942). Surprisingly, the genetic differentiation between the two phenotypically very different taxa is extremely low, with only a few genomic regions harboring diagnostic SNP differences (Poelstra et al. 2014). A recent study on the genetic composition of hybrids across the hybrid zone has identified an interaction of two genomic regions on different chromosomes as the main factors responsible for phenotypic divergence (Knief et al., *in revision*). There seem to be a gradient in the degree of genetic differentiation between all-black and black-and-grey species pairs. While in the house crow *C. splendens* different phenotypes merely represent morphs within an admixed population, the all black Eurasian jackdaw *C. monedula* and black-and-white Daurian jackdaw *C. dauricus* are genetically divergent and potentially do not even hybridize (Madge and Burn 1994). This heterogeneity in genetic differentiation between independent phenotypically different species pairs provides an ideal system to investigate the link between phenotype and genotype and unveil the genetic basis of population separation.

6. Methods

6.1 Single-molecule sequencing

With the completion of the first human genome draft assembly in the early 2000s, there was a burst in development of sequencing technologies (Lander et al. 2001; Venter et al. 2001). Owing its name to the massive amount of data which could be suddenly produced, the first generation of high-throughput sequencing had one major limitation: the length of single reads. With only about 30 bp when first introduced, even nowadays the reads of this kind of technology do not exceed a few hundred base pairs (Goodwin, McPherson, and McCombie 2016).

The advent of single-molecule sequencing changed the situation dramatically (Eid et al. 2009; Mikheyev and Tin 2014). Two competing companies with fundamentally different approaches are currently dominating this segment: Pacific Biosystems (now owned by Illumina) and Oxford Nanopore Technologies. Sequencing data from Pacific Biosystems – termed single-molecule real-time sequencing (SMRT) - is the main data type used in this thesis and will thus be described in more detail. At the beginning of essentially all long-read or long-range technologies stands the extraction of high-molecular weight (HMW) DNA. This is a crucial step, since fragmented DNA limits the resulting read length. In an enzymatic ligation, template DNA is circularized with an adaptor and loaded onto the sequencing chip. The actual sequencing begins when the single-stranded template DNA is guided through a DNA polymerase and the nucleotides used in synthesizing the complementary strand are emitting fluorescent signals corresponding to each of the four bases. All of this is happening in a less than a hundred nanometers wide well, called zero-mode wave guide, which detects the nucleotide sequence of the synthesized strand. Sequence reads of over 100 kb can be achieved and average read lengths lie now routinely above 15 kb. Besides read length, the major advantage of the SMRT-sequencing technology is that the circular template molecule can be sequenced up to over 50 times, enabling the creation of a consensus over them. To some extent this alleviates the fact that the sequencing error rate of individual reads is quite high with around 15%. Furthermore, SMRT sequencing does not include any amplification step, which is known to be biased against certain sequence compositions (Dohm et al. 2008). The other main competitor among long-read sequencing technologies is Oxford Nanopore Technologies (ONT). Here, individual DNA molecules pass through

pores a few nanometers wide which sit on a membrane and are able to detect minute fluctuations in the electric current. DNA molecules passing through the pore cause characteristic disruptions in the electric current, which is used by the pore to record the nucleotide sequence. As opposed to SMRT-sequencing, a given template molecule is usually sequenced only once, without the possibility to draw a consensus and a resulting higher error rate of the raw data. In turn, ONT holds the current record in sequencing read length, which has in some cases approached one million bases (Jain et al. 2018). With long-read sequencing, far more contiguous assemblies than with previously available technologies can be generated, and facilitated by decreasing sequencing costs, it is now possible to produce high quality reference genomes for almost any organism (Sedlazeck et al. 2018).

6.2 Optical mapping

Another technology which capitalizes on large single DNA molecules is optical mapping. Having its origin in the early 1990s (Schwartz et al. 1993), the underlying concept has been commercialized by two different companies (OpGen and BioNano Genomics) (Lam et al. 2012; Latreille et al. 2007). Long DNA molecules (> 150 kb) are stretched out uniformly in nano-channels and treated with a restriction enzyme which cuts at a specific recognition motif (e.g. BspQI: 5'-GCTCTTC-3') and either inserts a fluorescently labelled nick strand or disrupts the molecule entirely. The order of these two steps depends on the technology. While in the original description and the OpGen technology the molecule is first elongated and then digested, in preparation for the BioNano technology input DNA is first digested and then stretched in nanochannels. Subsequently an image is taken and the resulting pattern represents a physical map of a single DNA molecule. Performed in a massively parallel fashion, this yields a multitude of single-molecule maps, which can be assembled into a genome-wide consensus map.

Importantly, optical mapping is not a sequencing technology, however the physical information gained from single molecule and consensus maps can be used for comparison with *in-silico* restriction maps from an existing genome assembly. Individual scaffolds and contigs of the assembly can be merged into super-scaffolds and mismatches in the comparison can be used to detect misassemblies. Optical mapping constitutes a complementary approach to sequencing and can therefore be used as an independent tool for quality control of genome assemblies (Shelton et al. 2015). The facilitation of long-range information (single molecules are over 150 kb in size) and independence of sequence context also renders optical maps ideal candidates to detect large-scale structure variation, i.e. variants which are either too large or repetitive to be detected via short- or long-read sequencing. This has been demonstrated in a variety of organisms (e.g. Chakraborty et al. 2018; Kronenberg et al. 2018),

but the reliance on the targeted restriction motif to be present in regions of interest is certainly a limitation.

6.3 Detecting genetic variation

6.3.1 Genome assembly

At the beginning of almost every study looking at genetic variation is the determination of a genomic reference. For organisms for which there is no reference assembly available yet, it can be assembled *de novo*, resulting in an entirely novel representation of the genome. The process of genome assembly can probably be best described with a metaphor: it is nothing more than a jigsaw puzzle, with the input sequencing reads corresponding to the puzzle pieces and an end-to-end genome assembly to the finished jigsaw puzzle. The basic principle is rather simple: the overlap between any two sequence reads can be used to merge them into longer sequences, also called contigs (contiguous sequence) (Yandell and Ence 2012). This process is repeated until chromosomes in chromosome length are generated. Obviously, reality is far from being that straightforward. One main factor constraining genome assembly is the sheer vastness of genomes. To put things in perspective: If one would print out the 1.3 billion base pairs of the crow genome (Venturini, D'Ambrogi, and Capanna 1986) – which is rather small for vertebrates – in 1 millimeter letters, it would cover more than the distance from Rome to Paris. To find an overlap between any two sequence reads (which would correspond to lengths between 15 centimeters and 20 meters) seems very hard to imagine. However, the development of efficient algorithms and an ever-increasing computational power have made this possible. In the early days of assembling large complex genomes (such as mouse and human), mid-sized stretches of the DNA stemming from Sanger-sequenced bacterial artificial clones were assembled with an approach called overlap-layout consensus. With the onset of high-throughput sequencing, the number of available reads to start with increased dramatically, but the length decreased substantially. The result of this new situation was the development of assembly methods based on de Bruijn graphs, which essentially break each read into smaller, overlapping pieces of fixed length, which are then merged to yield assembled contigs. Technological development once again spurred algorithmic development when long-read data became available (Chaisson, Wilson, and Eichler 2015). Contemporary long-read assembly tools, which are commonly based on string graphs, can deal with the constraints inherent in long, error-prone reads and deliver the most contiguous assemblies to date. Most recently, the FALCON-UNZIP assembly tool has been able to capitalize on another advantage of long reads: the information on haplotype structure (Chin et al. 2016). This assembler is able to separate the two chromosomes of a diploid genome and assembles them into

a set of primary and associated contigs (or haplotigs). Previous assembly methods (of all kinds) were – at best – only able to pick on or the other haplotype of a diploid input, but often the resulting haploid assembly included chimeric sequences which represented neither of the chromosomes. A diploid-aware assembly tool circumvents this problem, resulting in a more accurate reference.

Whenever multiple assemblies are available, for example two haplotypes of a single individual as stated above, it is possible to detect genetic variation through comparison between them. After whole-genome alignment, discordant regions ranging from single to millions of base pairs indicate the presence of genetic variants (Chaisson, Wilson, and Eichler 2015; Kronenberg et al. 2018; Chakraborty et al. 2018). In case of large rearrangements, discordant regions are often identifiable via manual inspections of dotplots (a visual summary of a pairwise alignment). For large genomes, however, automated variant calling algorithms are the method of choice. Note that discordant regions in the alignment do not necessarily represent a true genetic variant, but can also mean an error in either of the assemblies. A combination of different approaches (e.g. read mapping, see below), is thus advisable to differentiate false positives from true genetic variation.

6.3.2 Re-sequencing and read alignment

There are situations when *de-novo* assembly of genomes to identify genetic variation is not an option, for example where genetic material is insufficient to yield sufficient sequencing coverage, or when genome size or number of individuals are prohibitive in terms of cost. Then, single or multiple individuals can be re-sequenced and genetic variants are identified via read alignments to a reference assembly. For single-nucleotide variation, this is relatively straightforward. Pileups of unambiguously aligned reads are screened for nucleotides with differences consistent across all reads (or half of them in case of heterozygotes) which are then, usually according to a model which incorporates the likelihood of the mutation and population information, scored as SNPs (Ekblom and Wolf 2014). The discovery of single-nucleotide variation in non-repetitive regions is largely independent of read length, since even very short reads can usually be mapped unambiguously to unique stretches of sequence in a reference genome.

This is not true however for the detection of structural variation, especially since this type of variation is commonly either constituting repetitive DNA directly or associated with it. Most structural variants can be identified via so-called split-read mapping (in case of long reads) or discordant read pairs (in case of paired-end short reads) (Sedlazeck et al. 2018). This approach uses the contradicting information coming from an alignment of a single read or read pair. If at least two discordant alignments are reported for a single read, this indicates that the re-sequenced individual does not conform to the reference.

While deletions are relatively simple to identify via gaps in the alignment, insertions often contain novel sequence and are thus reliant on the read entirely spanning the insertion event to confidently locate it in the reference assembly. The same is true for inversions. In principle it is possible to identify and anchor breakpoints for inversions larger than the aligned read length, however it may become extremely complicated if one (or both) breakpoints contain DNA repeats which lead to ambiguous alignments (Chaisson, Wilson, and Eichler 2015; Sedlazeck et al. 2018). Furthermore, palindromes in the reference can lead to a read alignment erroneously suggesting an inversion. Altogether, the identification and proper characterization of structural variation via read mapping is still far from trivial, but the constant development of long read technologies is promising for the field.

Research aims

General aims

The central aim of my thesis is to characterize the full extent of genetic variation in the songbird genus *Corvus* and understand its relationship with phenotypic divergence between closely related species. My focus lies on structural variation - genetic mutations beyond the single-nucleotide level that alter large stretches of the DNA sequence at once. This form of variation in genomes has been largely neglected in previous studies of natural populations, but has potentially dramatic effects on the phenotype.

Specific aims

Paper I - To investigate patterns of nucleotide diversity and divergence across multiple populations of *Corvus* crows across Europe and Asia. To establish if the same genomic regions show elevated genetic differentiation for parallel phenotypic contrasts in plumage pigmentation and may thus be involved in facilitating speciation.

Paper II -To compare landscapes of genome-wide nucleotide diversity across a wide range of evolutionary timespans and determine whether parallel patterns are associated with chromosomal features.

Paper III - To establish a new high-quality genomic reference for the hooded crow by assembling the individual previously used for the reference assembly with single-molecule real-time sequencing and optical mapping data. To investigate the link between large, potentially centromeric tandem repeat arrays and population genetic parameters.

Paper IV - To improve the hooded crow reference assembly to chromosome-level. To combine population-level short-read, long-read and optical mapping data to investigate structural genomic variation in the genus *Corvus* and its population genetic parameters. To identify candidate mutations responsible for the phenotypic contrast across the European hybrid zone.

Summary of the papers

Paper I - Evolution of heterogenous genome differentiation across multiple contact zones in a crow species complex

As populations diverge, genetic differences accumulate in their genomes. Due to the complex interaction of mutation, selection, recombination, and random genetic drift, genetic differentiation is not randomly distributed along the genome and – depending on how closely related two populations in a comparison are – distinct differentiation peaks can be distinguished (Wolf and Ellegren 2017; Cutter and Payseur 2013). However, the driving forces behind such a heterogeneous differentiation landscape are difficult to tease apart, since processes such as divergent selection against gene flow or linked selection on deleterious mutations result in very similar genomic signatures. A powerful approach to address this problem is to investigate multiple populations with independent contact zones of phenotypically different individuals. While in these contact zones a scenario of selection against gene flow can be assumed, and differentiation peaks likely contain genes responsible for phenotypic differences, allopatric population comparisons within the same phenotype should show a differentiation landscape concordant with linked selection and unrelated to the phenotypic contrast or genetic drift.

The Eurasian *Corvus (corone)* ssp. crow species complex is uniquely suited to investigate this question. Spread from Spain to Eastern China and Japan, populations exhibit two distinct phenotypes with either all-black and black-and-grey plumage color and meet in hybrid zones in Central Europe (*C. (c.) corone* & *C. (c.) cornix*) and Russia (*C. (c.) orientalis* & *C. (c.) cornix*, as well in another contact zone with all-black *C. (c.) orientalis* and black-and-white *C. (c.) pectoralis* in East Asia.

We generated whole-genome short read re-sequencing data for 128 individuals from 16 populations across the entire geographic distribution including three contact zones with phenotypic contrasts. Coalescent-based inference indicated a common ancestor of all populations roughly 300 000 years ago, a result consistent with population divergence during the 'Riss' Pleistocene glacial. We inferred population structure based on principal component analysis and showed that, somewhat surprisingly, populations do not cluster according to phenotype. Instead, deeply divergent lineages are contained in the all-black

populations on both the Western (Spain) and Eastern (Russia) end of the geographic distribution.

We calculated z-transformed F_{ST} (F_{ST}') in non-overlapping 50-kb windows to investigate genome-wide differentiation and used an 99th percentile outlier approach to identify genomic regions which are potentially under selection. Interestingly, differentiation peaks were not shared between population comparisons with the same phenotypic contrasts. The European hybrid zone yielded the same previously identified differentiation peak, spanning 2.8 Mb on chromosome 18 and containing genes involved in melanogenesis (PRKCA, SLC16A6, AXIN2, CACNG1, CACNG4 and CACNG5) and visual perception (RGS9) (Poelstra et al. 2014). Differentiation peaks in the Russian *C. (c.) cornix* and *C. (c.) orientalis* hybrid zones were less pronounced and located on different chromosomes (chromosome 21 and Z chromosome), but also contained genes from the melanogenesis pathway (CLCN6, MFN2 and MTOR).

Only one fixed difference, located in the LRP5 gene that interacts with the WNT pathway, was shared between both of the hybrid zones. The third across-phenotype comparison (*C. (c.) orientalis* and *C. (c.) pectoralis*) yielded one prominent differentiation peak on chromosome 23. While the genomic location of outlier regions were not shared between contact zone comparisons with idiosyncratic processes driving patterns of differentiation, a parallelism in genetic pathways resulting in similar phenotypic contrasts is suggesting a polygenetic architecture of the trait.

Paper II – Genome-wide patterns of variation in genetic diversity are shared among populations, species and higher-order taxa

Genetic diversity and hence differentiation between populations and species is distributed non-randomly along the genome (Wolf and Ellegren 2017). To identify the mechanisms underlying this heterogeneity is nontrivial, especially since they are not mutually exclusive. A scenario often invoked to explain heterogeneous differentiation landscapes with pronounced peaks of elevated differentiation are so-called 'islands of speciation' (Feder, Egan, and Nosil 2012; Nosil and Feder 2013). Hereby selection acting on loci responsible for reproductive isolation is counteracting homogenizing gene flow which reduces background levels of genetic differentiation. Another explanation for the observed heterogeneity in patterns of diversity is background selection. In regions of low recombination, recurrent deleterious mutations lead to a reduction of neutral diversity and thus an increase of genetic differentiation (Cruickshank and Hahn 2014; Charlesworth, Morgan, and Charlesworth 1993; Cutter and Payseur 2013).

In this study, we employ a comparative approach to investigate the processes leading to a heterogenous differentiation landscape and disentangle the contribution of shared linked selection and divergent selection against gene flow. We used population genomic data from three songbird clades (*Ficedula* flycatchers, Geospizinae Darwin's finches and *Corvus* crows) with in total 444 re-sequenced avian genomes (Burri et al. 2015; Lamichhaney et al. 2016; Vijay et al. 2016) to determine patterns of genome-wide nucleotide diversity π, genetic differentiation (F_{ST}, PBS), divergence (d_{xy}) and population-scaled recombination rate (ρ). Surprisingly, we found that broad-scale patterns of nucleotide diversity, differentiation, and divergence as well recombination rate are conserved both within and across the three clades, which are 50 million years divergent. Correlation coefficients for all population genetic parameters were consistently positive and lower between clades (0.082 – 0.32) compared to within clades (0.19 – 0.98). We then investigated the amount of overlap of differentiation outlier windows with putative locations of centromeres and telomeres inferred in zebra finch (Knief and Forstmeier 2016). For all three clades, we found overlap significantly greater than expected by chance (flycatchers: 58.53 % and 60.98 %, crows: 21.95 % and 31.7 %, Darwin's finches: 14.63 % and 29.27 %).

In conclusion, we found that genomic landscapes of genetic diversity, differentiation and recombination are conserved across broad evolutionary timescales in songbirds. This speaks for a role of linked selection rather than divergent selection against gene flow as a key driver for the observed pattern heterogeneity. Overall, it becomes clear that comparative approaches using multiple independent study systems are a valuable tool to study the processes generating heterogeneous landscapes of genetic diversity.

Paper III – Combination of short-read, long-read and optical mapping assemblies reveals large-scale tandem repeat arrays with population genetic implications.

The generation of accurate and contiguous genome assemblies is central to molecular and population genetic studies. While this necessity is omnipresent, its achievement is often hampered by limitations caused by repetitive DNA. The reason for that is that DNA repeats introduce ambiguity in the task of finding overlaps between single sequencing reads and sequence contigs, leading to either fragmented or incorrect assemblies. Furthermore, the occurrence of large tandem repeat arrays – high copy number repeats arranged in a head-to-tail orientation – is often associated with chromosomal features such as centromeres and telomeres, which in turn influence sequence composition and recombination rate. The ability to locate and anchor such tandem repeat arrays is thus a vital prerequisite in population genetic studies.

In this study we used a combination of different technologies and methodological approaches to characterize and locate large tandem repeat arrays in an avian model for speciation, the hooded crow *Corvus (corone) cornix*. We sequenced the genome of a male individual to a 52-fold coverage depth using Pacific Biosystems single-molecule real-time sequencing and generated a highly contiguous assembly (8.58 Mb contig N50). We then compared this assembly with a previously published short-read assembly and an independently assembled optical map and identified 36 large-scale tandem repeat arrays at the ends of scaffolds. In the vicinity of these repetitive anchored maps (RAMs) we identified a novel satellite repeat which we termed "crowSat1". In the majority of scaffold ends which contained a RAM, we also found this 14-kb DNA satellite repeat or its 1.2-kb subunit, supporting the notion that the observed RAMs consist of large tandem repeat arrays of this repeat.

Large tandem repeat arrays are commonly associated with chromosomal features such as centromeres and telomeres, which influence meiotic recombination and the distribution of nucleotide diversity. Thus, we then looked at the relationship of RAMs and crowSat1 with population genetic parameters of adjacent genomic regions. Using a large-scale population re-sequencing data set (Vijay et al. 2016), we estimated the population-scaled recombination rate ρ and calculated the weighted mean in 50-kb windows for two crow populations. Windows adjacent to RAMs or containing crowSat1 show a significantly reduced recombination rate both compared to the genome-wide average and other windows at scaffold ends. Our findings thus concur with an effect of large tandem repeat arrays on recombination and point towards a centromeric or telomeric association. Knowledge on these large tandem repeat arrays can be incorporated in downstream population genetic analysis and prevent potential biases in the inference of selection or demographic history. In conclusion, this study demonstrates how the combination of different, complementary technologies can lead to the discovery of previously hidden, yet important genomic features.

Paper IV – Fine-scale analysis of Structural Genomic Variation in Natural Populations

In most organisms, structural genomic variation accounts for a large proportion of the variation between two genomes (Chaisson, Wilson, and Eichler 2015; Huddleston and Eichler 2016). Despite the potential to cause diseases and drastically alter phenotypes and thus be of evolutionary importance (Küpper et al. 2016; Feuk, Carson, and Scherer 2006; Feulner and De-Kayne 2017), structural variation has received much less attention than for example single nucleotide variation (Huddleston and Eichler 2016). In part this is due to the fact that technological limitations have thus far constrained the discovery of

this type of variation. More specifically, the presence of repetitive DNA is hampering the reliable discovery and genotyping of structural variants. Both sequence assembly and read mapping require sequence reads containing unambiguous information to yield confident results. However, if a single read is not able to span an interspersed repeat or a tandem repeat array entirely this is not possible. Thus, the ability to reliably detect structural variants greatly depends on the length of single reads (Sedlazeck et al. 2018).

In this study we aimed to characterize and quantify the full size spectrum of structural genomic variation in a natural population setting. Using a model system for speciation, the Eurasian crow *Corvus (c.)* ssp. species complex and further members of the same songbird genus, we portray genetic variation beyond the single-nucleotide level and investigate its role in the divergence of phenotypically different populations. To generate a phased, diploid genome reference, we re-assembled SMRT-sequencing long-read data with the FALCON-UNZIP assembly tool (Chin et al. 2016). We then further improved assembly contigs with a previously published optical map (Weissensteiner et al. 2017) and Hi-C chromatin interaction mapping data resulting in chromosome-level scaffolds. We also generated long-read and optical mapping data from a jackdaw *Corvus monedula* individual, to produce another reference assembly from the same genus and included a previously published assembly of the Hawaiian crow *Corvus hawaiiensis* (Sutton et al. 2018). Contigs for all three assemblies have been generated with FALCON-UNZIP and similar input data.

We then investigated genetic variation on the single individual level by aligning the two haplotypes of each assembly and identifying single-nucleotide and structural variation. Median densities of SVs and SNPs per 1-Mb windows were similar for the hooded crow and jackdaw with 7 SVs (2,228 SNPs) and 5 SVs (1,558 SNPs), respectively. In contrast, median SV and SNP densities were only 0 and 75 per Mb window for the Hawaiian crow, a result expected given the heavily inbred population of this species (Sutton et al. 2018). Next, we looked at SV segregating within and between natural populations. We generated long-read (LR) data for 31 individuals across the *Corvus* phylogeny, comprising the European jackdaw (*C. monedula*), the Daurian jackdaw (*C. dauuricus*), the American crow (*C. brachyrhynchos*) and three populations of the Eurasian crow complex (*C. (corone)* ssp.). Using a read-mapping based approach, we identified in total 51,405 variants across all individuals, of which 1,767 were private to single individuals. To account for errors in genotyping, we exploited the phylogenetic information contained in the sampling setup and retained a set of 35,065 variants. Given the large divergence time between clades (13 million years), we assumed that polymorphisms shared between clades should be rare and thus likely constitute genotyping errors. Among retained variants, we identified 1,165 inversions, 13,134 insertions and 16,794 deletions. Of all insertions and deletions, 6,998 could be associated with tandem repeats and 8,558 with interspersed repeats.

We then quantified the amount of segregating polymorphism in the crow and jackdaw clade and found 24,036 and 12,147 variants to be polymorphic, respectively. Folded site frequency spectra across all populations in each clade exhibited a strongly right skewed pattern indicating an excess of rare alleles. This was not the case however for insertions and deletions associated with simple and low complexity repeats, which showed elevated intermediate frequencies. Apart from a potential technical bias, it is possible that a high amount of forward and backward mutation typical for these types of tandem repeat is responsible for the observed pattern. To increase the number of sampled individuals, populations and species, we employed a short-read (SR) based SV detection approach using a previously published data set (Vijay et al. 2016; Poelstra et al. 2014). We identified 132,025 SVs, of which only 3,206 overlapped with the LR variant set. This difference cannot solely be explained by the difference in sample size, indicating a high amount of false-positives and an in general lower suitability of short read data to detect SV. Next, we compared the two phenotypically divergent European crow populations to detect variants which are potentially under selection. We estimated genetic differentiation (F_{ST}) for each variant in the LR variant set and performed an outlier scan, yielding 163 SVs. Among these we identified a 2.25-kb LTR retrotransposon insertion 20 kb upstream of the *NDP* gene on chromosome 1. This gene has been shown to play a role in plumage patterning in pigeons (Vickrey et al. 2018) and is involved in maintaining the hybrid zone of black-and-grey and all-black crows in Europe (Poelstra et al. 2014; Knief et al., n.d.). Interestingly, the insertion is polymorphic or absent in European all-black carrion crows, but fixed in black-and-grey hooded crows. Furthermore, we used normalized gene expression data of *NDP* in skin (Poelstra et al. 2015) and find a significant association with the insertion genotype, suggesting a causal role of the insertion in the phenotypic divergence.

In summary, we show that long-read sequencing has the potential to greatly expand our view on genetic variation and facilitates the detailed characterization and population genetic analysis of previously unknown variants. We demonstrate this approach in an evolutionary model system and pinpoint a candidate structural variant potentially responsible for population divergence.

Sammanfattning på Svenska

Ärftlig genetisk variation är en av grundstenarna i evolutionen. Det beror på att slumpmässiga förändringar (så kallade mutationer) i arvsmassan kan generera skillnader i observerbara egenskaper (fenotyper) hos en organism. Evolutionen verkar på fenotypen, men genom dess koppling till arvsmassan kan evolutionära processer förändra frekvensen av de underliggande mutationerna i en population. Genetisk variation är ett brett begrepp som inkluderar både enstaka basutbyten (s.k. enbaspolymorfier) och storskaliga rearrangemang av miljontals sammanhängande nukleotider samtidigt (strukturella variation). Det innebär att effekten på den resulterande fenotypen kan variera beroende på vilken typ av mutation som sker och på vilken position i arvsmassan den inträffar. Målet med denna avhandling är att ge en överblick över genetisk variation och deras förekomst och frekvens i naturen genom att studera arter inom släktet *Corvus*. För att nå målet använder vi oss av flera olika sekvenseringstekniker och populationsgenetiska metoder för dataanalys. Dessa verktyg hjälper att belysa genetiska variationer i olika tidsperspektiv, från individuell variation inom en population till jämförelser mellan avlägsen besläktade arter. Vi börjar med att studera ett geografiskt område som sedan tidigare är känt för förekomsten av hybrider av olika arter av kråkfåglar. Utbredningsområdena hos svartkråka (*C. (corone) corone*) och gråkråka (*C. (corone) cornix*) överlappar i smala hybridzoner i både Centraleuropa, Centralasien och Sydostasien. I den första studien jämför vi den genetiska variationen från de tre olika hybridzonerna. Den visar att det inte finns några tydliga regioner i genomet som eventuellt bestämmer fenotypiska skillnader. Samtidigt fanns det tydliga överensstämmelser i grupper av gener och gennätverk bland de olika hybridzoner. I den andra studien undersöker vi olika evolutionära tidsramar för att förstå vilka faktorer som driver heterogeniteten i genetisk differentiering som tidigare har observerats i jämförelser mellan många olika fågelarter. Genetisk diversitet, och därmed differentiering, verkar korrelera såväl mellan populationer av samma art, som mellan olika arter som separerade för upp till 50 miljoner år sedan. Detta mönster skapas genom en konserverad variation i rekombinationsfrekvens, som i sin tur sannolikt beror av kromosomstrukturer som centromerer och telomerer. Vi karaktäriserar också hela gråkråkans arvsmassa, genererad med hjälp av sekvensering av långa DNA-molekyler samt optical mapping. Användning av dessa teknologier gav oss inblick i tidigare icke-tillgängliga delar av arvsmassan och vi upptäckte grupper av storskaliga repetitiva delar som förekom i tandem (14 kilobaser långa ”satelliter” med en

underenhet som utgjorde 1.2 kilobaser). Eftersom dessa grupper av tandemrepetitioner är associerade med genetiska regioner som kännetecknas av en minskad rekombinationshastighet inom hela populationen, drar vi slutsatsen att ursprunget skulle kunna vara centromerer eller telomerer. Vi avslutar avhandlingen med att utöka analysen av arvsmassan från enbaspolymorfier till undersökning av större strukturella rearrangemang, såsom insertioner, deletioner och inversioner inom släktet Corvus. Användning av Hi-C-teknik (kartläggning av arvsmassan med hjälp av kromatininteraktion) gör det möjligt att sammanfoga bitar av DNA-sekvens till fullständiga kromosomer. Sekvensering av långa DNA molekyler gör det också möjligt att få information om diploida individers enskilda haplotyper. Vi gjorde en detaljerad kartläggning av strukturella varianter i åtta olika fågelarter med hjälp av olika sekvenseringsteknologier. Våra analyser visade att det fanns en variabel, cirka 2.25-kilobaser lång, repeterad sekvens 20 kb uppströms genen NDP. Sekvensen tillhör ERVK-familjen och är fixerad (icke-variabel) i gråkråka. Sekvensen har tidigare kopplats till sänkt genuttryck av NDP och är därmed en kandidatregion som skulle kunna ligga bakom skillnader i fjäderdräkt mellan olika Corvusarter.

Acknowledgements

There is a large number of I would like to thank and who all contributed to the fact that last five years have been the most exciting in my life so far.

Jochen, thanks for being my main supervisor and accepting me as a grad student, also for the seemingly endless generosity in respect to trying out new sequencing technologies. Manfred, even though we haven't had so much interaction (especially since I moved to Munich), I nevertheless want to thank you for many great discussions. Alex, thanks for the invaluable support in so many occasions and for sparking my interest in the dark side of the genome. Hans, thank you for creating such an amazing research environment, I am more than grateful that I could obtain my PhD here. From the people in the Wolf lab I would like to thank: Jelmer, for introducing me to what it means to be a PhD student and for memorable field moments, Christen, for all the fantastic hours in the crow core team, Andy, for nice discussions in Uppsala and elsewhere, Nagarjun, for being my bioinformatics lifeline more than once, Aaron, for the many great hours discussing science, politics and science politics, Melanie, for being a great office mate for a confused first year student, Chi-Chih, for great help in the lab, Julia, for providing insights into the Swedish culture, Kristaps, for being a fantastic field biologist (the jackdaw goes on your cap and that trip to Fjärnebofjärden I will never forget), Lars Dronasson, for facilitating my first paper, Glib, for many educating discussions, Verena, for spreading happiness, Gaby for saving me in the lab, Saurabh for being the computer wizard, and finally Steffi, Benedikt, Vera, Ana, Fidel and Josh for all the fun and exciting hours in the Wolf lab.

From the Suh lab I'd like to thank Valentina in particular, being a fellow gappie was fun and exciting. Jesper, Anne-Marie and Octavio, or someone else from the Suh lab was always there (on Slack) when I needed support.

In the hope that I don't forget anyone, I would like to express my gratitude towards the following people at the EBC, who have made it an even better place: Nina, Alex C., Mario, Lena, Luciana, TJ, Agnes, Marcin, Torsten, Ayca, Karl, Ruxi, Berrit, Roy, Anna, Linnea S., Linnea B., Jesper, Ghazal, Lore, Paulina, Venkat, Robert and Taki!

Karin, thanks for always sharing some carefully crafted puns and the last minute help with the Svensk samfattning. Niclas, thanks for following my invitation to Turku and for helping to fine-tune the Swedish summary. Willian, thank you for nicely welcoming me in the EvoBio enclave back in 2014. You are a very good colleague and I always enjoyed the discussions with you. Roy,

thanks for enduring me (among other noisy people) in the office and helping out with R questions. Ludo, thanks for being a superb role model, all the bio-info support and table tennis matches. Sergio, thank you for being a fantastic colleague and office mate. Not only were you incomprehensively tolerant about diverse acoustic disturbances from my side, but you were also extremely patient when I asked you for the 300th time how to properly plot something in ggplot. Claire, thank you for sharing an office with me. I was grateful that you came along to Munich and spread the Uppsala vibe. Uli, thank you for helping me to adjust to Munich and providing valuable input on many aspects of my thesis. Fritz, thanks for all your support cutting through the SV jungle. Ricardo, thank you for your company and providing a new perspective. Homa, thank you for just being such a great person, a good friend and full of invaluable knowledge and wisdom. I'm glad that I had the chance to share an office with you.

I also want to express my gratitude towards Smålands nation, the German Ornithologist's society DO-G and the Swiss Ornithological Society Ala, who have supported conference and workshop visits. Dani, thanks for welcoming and hosting us at one of the most beautiful places imaginable. Stephan and Kristina, thank you for introducing me to academia.

Ein ganz herzliches Dankeschön an dieser Stelle an unsere austro-schwedischen Freunde Familie Lindh, Familie Grusell und Familie Primetzhofer! Ohne euch wäre Schweden nicht halb so schön gewesen!

Pipo und Mimi, ihr habts wahrscheinlich mit den größten Dank verdient. Eure ausdauernde Unterstützung und Begeisterung für das was ich so mach sucht wirklich ihresgleichen – Danke!

Bernadette, mit dir ist alles möglich. Ich bin unendlich froh und dankbar dich zu haben. Kilian und Emil, ihr seid meine größten Schätze und erfüllt mein Leben.

References

Alkan, Can, Bradley P. Coe, and Evan E. Eichler. 2011. "Genome Structural Variation Discovery and Genotyping." *Nature Reviews Genetics* 12 (5): 363–76. https://doi.org/10.1038/nrg2958.

Baer, Charles F., Michael M. Miyamoto, and Dee R. Denver. 2007. "Mutation Rate Variation in Multicellular Eukaryotes: Causes and Consequences." *Nature Reviews Genetics* 8 (8): 619–31. https://doi.org/10.1038/nrg2158.

Baird, Stuart J. E. 2015. "Exploring Linkage Disequilibrium." *Molecular Ecology Resources* 15 (5): 1017–19. https://doi.org/10.1111/1755-0998.12424.

Bateson, William, and Gregor Mendel. 1913. *Mendel's Principles of Heredity*. University press.

Brookes, Anthony J. 1999. "The Essence of SNPs." *Gene* 234 (2): 177–86. https://doi.org/10.1016/S0378-1119(99)00219-X.

Burri, Reto, Alexander Nater, Takeshi Kawakami, Carina F. Mugal, Pall I. Olason, Linnea Smeds, Alexander Suh, et al. 2015. "Linked Selection and Recombination Rate Variation Drive the Evolution of the Genomic Landscape of Differentiation across the Speciation Continuum of Ficedula Flycatchers." *Genome Research* 25 (11): 1656–65. https://doi.org/10.1101/gr.196485.115.

Chaisson, Mark J. P., Richard K. Wilson, and Evan E. Eichler. 2015. "Genetic Variation and the de Novo Assembly of Human Genomes." *Nature Reviews Genetics* 16 (11): 627–40. https://doi.org/10.1038/nrg3933.

Chakraborty, Mahul, Nicholas W. VanKuren, Roy Zhao, Xinwen Zhang, Shannon Kalsow, and J. J. Emerson. 2018. "Hidden Genetic Variation Shapes the Structure of Functional Elements in Drosophila." *Nature Genetics* 50 (1): 20–25. https://doi.org/10.1038/s41588-017-0010-y.

Charlesworth, Brian, M. T. Morgan, and Deborah Charlesworth. 1993. "The Effect of Deleterious Mutations on Neutral Molecular Variation." *Genetics* 134 (4): 1289–1303.

Chin, Chen-Shan, Paul Peluso, Fritz J Sedlazeck, Maria Nattestad, Gregory T Concepcion, Alicia Clum, Christopher Dunn, et al. 2016. "Phased Diploid Genome Assembly with Single-Molecule Real-Time Sequencing." *Nature Methods* 13 (October): 1050.

Chuong, Edward B., Nels C. Elde, and Cedric Feschotte. 2017. "Regulatory Activities of Transposable Elements: From Conflicts to Benefits." *Nature Reviews Genetics* 18: 71–86. https://doi.org/10.1038/nrg.2016.139.

Cruickshank, Tami E., and Matthew W. Hahn. 2014. "Reanalysis Suggests That Genomic Islands of Speciation Are Due to Reduced Diversity, Not Reduced Gene Flow." *Molecular Ecology* 23 (13): 3133–57. https://doi.org/10.1111/mec.12796.

Cutter, Asher D., and Bret A. Payseur. 2013. "Genomic Signatures of Selection at Linked Sites: Unifying the Disparity among Species." *Nature Reviews Genetics* 14 (4): 262–74. https://doi.org/10.1038/nrg3425.

Del Hoyo, J., A. Elliott, J. Sargatal, D. A. Christie, and E. de Juana. 2015. *Handbook of the Birds of the World Alive. Lynx Edicions, Barcelona*.

Dohm, J. C., C. Lottaz, T. Borodina, and H. Himmelbauer. 2008. “Substantial Biases in Ultra-Short Read Data Sets from High-Throughput DNA Sequencing.” *Nucleic Acids Research* 36 (16): e105–e105. https://doi.org/10.1093/nar/gkn425.

Eid, J., A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, et al. 2009. “Real-Time DNA Sequencing from Single Polymerase Molecules.” *Science* 323 (5910): 133–38. https://doi.org/10.1126/science.1162986.

Ekblom, Robert, and Jochen B. W. Wolf. 2014. “A Field Guide to Whole-Genome Sequencing, Assembly and Annotation.” *Evolutionary Applications* 7 (9): 1026–42.

Ellegren, H. 2004. “Microsatellites: Simple Sequences with Complex Evolution.” *Nature Reviews* 5 (June): 435–45. https://doi.org/10.1038/nrg1348.

Ellegren, Hans, and Nicolas Galtier. 2016. “Determinants of Genetic Diversity.” *Nature Reviews Genetics* 17 (7): 422–33. https://doi.org/10.1038/nrg.2016.58.

Feder, Jeffrey L., Scott P. Egan, and Patrik Nosil. 2012. “The Genomics of Speciation-with-Gene-Flow.” *Trends in Genetics* 28 (7): 342–50. https://doi.org/10.1016/j.tig.2012.03.009.

Feschotte, Cédric, and Clément Gilbert. 2012. “Endogenous Viruses: Insights into Viral Evolution and Impact on Host Biology.” *Nature Reviews Genetics* 13: 283–296. https://doi.org/10.1038/nrg3199.

Feuk, Lars, Andrew R. Carson, and Stephen W. Scherer. 2006. “Structural Variation in the Human Genome.” *Nature Reviews Genetics* 7 (2): 85–97. https://doi.org/10.1038/nrg1767.

Feulner, P. G. D., and R. De-Kayne. 2017. “Genome Evolution, Structural Rearrangements and Speciation.” *Journal of Evolutionary Biology* 30 (8): 1488–90. https://doi.org/10.1111/jeb.13101.

Goodwin, Sara, John D. McPherson, and W. Richard McCombie. 2016. “Coming of Age: Ten Years of next-Generation Sequencing Technologies.” *Nature Reviews Genetics* 17: 333–51. https://doi.org/10.1038/nrg.2016.49.

Graur, Dan, and Wen-Hsiung Li. 2000. *Fundamentals of Molecular Evolution*. 2nd ed. Sunderland, Mass: Sinauer Associates.

Graw, Jochen. 2015. *Genetik*. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-44817-5.

Hartl, Daniel L., and Andrew G. Clark. 1997. *Principles of Population Genetics*. Vol. 116. Sinauer associates Sunderland.

Huddleston, J., and E. E. Eichler. 2016. “An Incomplete Understanding of Human Genetic Variation.” *Genetics* 202 (4): 1251–54. https://doi.org/10.1534/genetics.115.180539.

Jain, Miten, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, et al. 2018. “Nanopore Sequencing and Assembly of a Human Genome with Ultra-Long Reads.” *Nature Biotechnology* 36 (4): 338–45. https://doi.org/10.1038/nbt.4060.

Jeffreys, Alec J., Victoria Wilson, and Swee Lay Thein. 1985. “Individual-Specific ‘Fingerprints’ of Human DNA.” *Nature* 316 (6023): 76–79.

Jønsson, Knud Andreas, Pierre-Henri Fabre, Jonathan D. Kennedy, Ben G. Holt, Michael K. Borregaard, Carsten Rahbek, and Jon Fjeldså. 2016. “A Supermatrix Phylogeny of Corvoid Passerine Birds (Aves: Corvides).” *Molecular Phylogenetics and Evolution* 94 (January): 87–94. https://doi.org/10.1016/j.ympev.2015.08.020.

Kapitonov, Vladimir V., and Jerzy Jurka. 2006. “Self-Synthesizing DNA Transposons in Eukaryotes.” *Proceedings of the National Academy of Sciences of the United States of America* 103 (March): 4540–45. https://doi.org/10.1073/pnas.0600833103.

Katzourakis, A., and R. J. Gifford. 2010. “Endogenous Viral Elements in Animal Genomes.” *PLoS Genetics* 6: e1001191.
Kazazian Jr., Haig H. 2004. “Mobile Elements: Drivers of Genome Evolution.” *Science* 303 (March): 1626–32. https://doi.org/10.1126/science.1089670.
Kidwell, M. G. 2005. “Transposable Elements.” In *The Evolution of the Genome*, edited by T. Ryan Gregory, 165–221. Elsevier Academic Press.
Knief, Ulrich, Christen M. Bossu, Nicola Saino, Bengt Hansson, Jelmer Poelstra, Nagarjun Vijay, Matthias H. Weissensteiner, and Jochen B. W. Wolf. n.d. “Epistatic Mutations under Divergent Selection Govern Phenotypic Variation in the Crow Hybrid Zone.” *Manuscript*.
Knief, Ulrich, and Wolfgang Forstmeier. 2016. “Mapping Centromeres of Microchromosomes in the Zebra Finch (Taeniopygia Guttata) Using Half-Tetrad Analysis.” *Chromosoma* 125 (4): 757–68. https://doi.org/10.1007/s00412-015-0560-7.
Kronenberg, Zev N., Ian T. Fiddes, David Gordon, Shwetha Murali, Stuart Cantsilieris, Olivia S. Meyerson, Jason G. Underwood, et al. 2018. “High-Resolution Comparative Analysis of Great Ape Genomes.” *Science* 360 (6393): eaar6343. https://doi.org/10.1126/science.aar6343.
Küpper, Clemens, Michael Stocks, Judith E. Risse, Natalie dos Remedios, Lindsay L. Farrell, Susan B. McRae, Tawna C. Morgan, et al. 2016. “A Supergene Determines Highly Divergent Male Reproductive Morphs in the Ruff.” *Nature Genetics* 48: 79–83. https://doi.org/10.1038/ng.3443 http://www.nature.com/ng/journal/v48/n1/abs/ng.3443.html#supplementary-information.
Lam, Ernest T, Alex Hastie, Chin Lin, Dean Ehrlich, Somes K Das, Michael D Austin, Paru Deshpande, et al. 2012. “Genome Mapping on Nanochannel Arrays for Structural Variation Analysis and Sequence Assembly.” *Nature Biotechnology* 30 (8): 771–76. https://doi.org/10.1038/nbt.2303.
Lamichhaney, Sangeet, Guangyi Fan, Fredrik Widemo, Ulrika Gunnarsson, Doreen Schwochow Thalmann, Marc P. Hoeppner, Susanne Kerje, et al. 2016. “Structural Genomic Changes Underlie Alternative Reproductive Strategies in the Ruff (Philomachus Pugnax).” *Nature Genetics* 48: 84–88. https://doi.org/10.1038/ng.3430 http://www.nature.com/ng/journal/v48/n1/abs/ng.3430.html#supplementary-information.
Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, et al. 2001. “Initial Sequencing and Analysis of the Human Genome.” *Nature* 409: 860–921. https://doi.org/citeulike-article-id:3297864.
Latreille, Phil, Stacie Norton, Barry S Goldman, John Henkhaus, Nancy Miller, Brad Barbazuk, Helge B Bode, et al. 2007. “Optical Mapping as a Routine Tool for Bacterial Genome Sequence Finishing.” *BMC Genomics* 8 (1): 321. https://doi.org/10.1186/1471-2164-8-321.
Lee, Yuh Chwen G., and Charles H. Langley. 2010. “Transposable Elements in Natural Populations of *Drosophila Melanogaster*.” *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 365 (April): 1219–28. https://doi.org/10.1098/rstb.2009.0318.
Levin, Henry L., and John V. Moran. 2011. “Dynamic Interactions between Transposable Elements and Their Hosts.” *Nature Reviews Genetics* 12: 615–27.
Lewontin, R. C., and Ken-ichi Kojima. 1960. “The Evolutionary Dynamics of Complex Polymorphisms.” *Evolution* 14 (4): 458–72.
Litt, Michael, and Jeffrey A. Luty. 1989. “A Hypervariable Microsatellite Revealed by in Vitro Amplification of a Dinucleotide Repeat within the Cardiac Muscle Actin Gene.” *American Journal of Human Genetics* 44 (3): 397.
López-Flores, I, and MA Garrido-Ramos. 2012. “The Repetitive DNA Content of Eukaryotic Genomes.” *Repetitive DNA* 7: 1–28.

Luan, Dongmei D., Malka H. Korman, John L. Jakubczak, and Thomas H. Eickbush. 1993. “Reverse Transcription of R2Bm RNA Is Primed by a Nick at the Chromosomal Target Site: A Mechanism for Non-LTR Retrotransposition.” *Cell* 72: 595–605. https://doi.org/10.1016/0092-8674(93)90078-5.

Lupski, James R., and Pawel Stankiewicz. 2005. “Genomic Disorders: Molecular Mechanisms for Rearrangements and Conveyed Phenotypes.” *PLoS Genetics* 1 (6): 0627–33. https://doi.org/10.1371/journal.pgen.0010049.

Lynch, Michael. 2007. *The Origins of Genome Architecture*. Vol. 98. Sinauer Associates Sunderland (MA).

Madge, Steve, and Hilary Burn. 1994. *Crows and Jays: A Guide to the Crows, Jays and Magpies of the World*. A&C Black.

Martin, Sandra L. 2006. “The ORF1 Protein Encoded by LINE-1: Structure and Function during L1 Retrotransposition.” *Journal of Biomedicine and Biotechnology* 2006: 6. https://doi.org/10.1155/jbb/2006/45621.

Massy, Bernard de. 2013. “Initiation of Meiotic Recombination: How and Where? Conservation and Specificities Among Eukaryotes.” *Annual Review of Genetics* 47 (1): 563–99. https://doi.org/10.1146/annurev-genet-110711-155423.

Mayr, Ernst. 1942. *Systematics and the Origin of Species, from the Viewpoint of a Zoologist*. Harvard University Press.

Meise, Wilhelm. 1928. “Die Verbreitung Der Aaskrähe (Formenkreis Corvus Corone L.)...”

Miga, Karen H. 2015. “Completing the Human Genome: The Progress and Challenge of Satellite DNA Assembly.” *Chromosome Research* 23 (3): 421–426. https://doi.org/10.1007/s10577-015-9488-2.

Mikheyev, Alexander S., and Mandy M. Y. Tin. 2014. “A First Look at the Oxford Nanopore MinION Sequencer.” *Molecular Ecology Resources* 14 (6): 1097–1102. https://doi.org/10.1111/1755-0998.12324.

Morgan, Thomas Hunt. 1919. *The Physical Basis of Heredity*. JB Lippincott.

Nosil, Patrik, and Jeffrey L. Feder. 2013. “GENOME EVOLUTION AND SPECIATION: TOWARD QUANTITATIVE DESCRIPTIONS OF PATTERN AND PROCESS: SPECIAL SECTION.” *Evolution* 67 (9): 2461–67. https://doi.org/10.1111/evo.12191.

Perry, George H, Nathaniel J Dominy, Katrina G Claw, Arthur S Lee, Heike Fiegler, Richard Redon, John Werner, et al. 2007. “Diet and the Evolution of Human Amylase Gene Copy Number Variation.” *Nature Genetics* 39 (September): 1256.

Plohl, Miroslav. 2010. “Those Mysterious Sequences of Satellite DNAs.” *Periodicum Biologorum* 112 (4): 403–10.

Poelstra, J. W., N. Vijay, C. M. Bossu, H. Lantz, B. Ryll, I. Müller, V. Baglione, et al. 2014. “The Genomic Landscape Underlying Phenotypic Integrity in the Face of Gene Flow in Crows.” *Science* 344 (6190): 1410–15.

Poelstra, J. W., N. Vijay, M. P. Hoeppner, and J. B. W. Wolf. 2015. “Transcriptomics of Colour Patterning and Coloration Shifts in Crows.” *Molecular Ecology* 24 (18): 4617–28. https://doi.org/10.1111/mec.13353.

Redon, Richard, Shumpei Ishikawa, Karen R. Fitch, Lars Feuk, George H. Perry, T. Daniel Andrews, Heike Fiegler, et al. 2006. “Global Variation in Copy Number in the Human Genome.” *Nature* 444 (7118): 444–54. https://doi.org/10.1038/nature05329.

Sanger, F., S. Nicklen, and A. R. Coulson. 1977. “DNA Sequencing with Chain-Terminating Inhibitors.” *Proceedings of the National Academy of Sciences* 74 (12): 5463–67. https://doi.org/10.1073/pnas.74.12.5463.

Schlötterer, Christian. 2004. “Opinion: The Evolution of Molecular Markers — Just a Matter of Fashion?” *Nature Reviews Genetics* 5 (1): 63–69. https://doi.org/10.1038/nrg1249.

Schlötterer, Christian, and Diethard Tautz. 1992. “Slippage Synthesis of Simple Sequence DNA.” *Nucleic Acids Research* 20 (2): 211–15. https://doi.org/10.1093/nar/20.2.211.

Schwann, T. H. 1847. *Microscopical Researches into the Accordance in the Structure and Growth of Animals and Plants*. Рипол Классик.

Schwartz, D., X Li, L. Hernandez, S. Ramnarain, E. Huff, and Y. Wang. 1993. “Ordered Restriction Maps of Saccharomyces Cerevisiae Chromosomes Constructed by Optical Mapping.” *Science* 262 (5130): 110–14. https://doi.org/10.1126/science.8211116.

Sedlazeck, Fritz J., Hayan Lee, Charlotte A. Darby, and Michael C. Schatz. 2018. “Piercing the Dark Matter: Bioinformatics of Long-Range Sequencing and Mapping.” *Nature Reviews Genetics* 19 (6): 329–46. https://doi.org/10.1038/s41576-018-0003-4.

Sharp, Andrew J., Devin P. Locke, Sean D. McGrath, Ze Cheng, Jeffrey A. Bailey, Rhea U. Vallente, Lisa M. Pertz, et al. 2005. “Segmental Duplications and Copy-Number Variation in the Human Genome.” *The American Journal of Human Genetics* 77 (1): 78–88. https://doi.org/10.1086/431652.

Shelton, Jennifer M., Michelle C. Coleman, Nic Herndon, Nanyan Lu, Ernest T. Lam, Thomas Anantharaman, Palak Sheth, and Susan J. Brown. 2015. “Tools and Pipelines for BioNano Data: Molecule Assembly Pipeline and FASTA Super Scaffolding Tool.” *BMC Genomics* 16 (1). https://doi.org/10.1186/s12864-015-1911-8.

Smit, Arian FA. 1996. “The Origin of Interspersed Repeats in the Human Genome.” *Current Opinion in Genetics & Development* 6 (6): 743–48. https://doi.org/10.1016/S0959-437X(96)80030-X.

Sturtevant, A. H. 1913. “The Linear Arrangement of Six Sex-Linked Factors in Drosophila, as Shown by Their Mode of Association.” *Journal of Experimental Zoology* 14 (1): 43–59. https://doi.org/10.1002/jez.1400140104.

Sutton, Jolene T., Martin Helmkampf, Cynthia C. Steiner, M. Renee Bellinger, Jonas Korlach, Richard Hall, Primo Baybayan, et al. 2018. “A High-Quality, Long-Read De Novo Genome Assembly to Aid Conservation of Hawaii’s Last Remaining Crow Species.” *Genes* 9 (8): 393. https://doi.org/10.3390/genes9080393.

Venter, J. Craig, Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, Granger G. Sutton, Hamilton O. Smith, et al. 2001. “The Sequence of the Human Genome.” *Science* 291 (5507): 1304–51. https://doi.org/10.1126/science.1058040.

Venturini, Giorgio, Raffaella D’Ambrogi, and Ernesto Capanna. 1986. “Size and Structure of the Bird Genome—I DNA Content of 48 Species of Neognathae.” *Comparative Biochemistry and Physiology Part B: Comparative Biochemistry* 85 (1): 61–65. https://doi.org/10.1016/0305-0491(86)90221-X.

Vickrey, Anna I, Rebecca Bruders, Zev Kronenberg, Emma Mackey, Ryan J Bohlender, Emily T Maclary, Raquel Maynez, et al. 2018. “Introgression of Regulatory Alleles and a Missense Coding Mutation Drive Plumage Pattern Diversity in the Rock Pigeon.” *ELife* 7 (July). https://doi.org/10.7554/eLife.34803.

Vijay, Nagarjun, Christen M. Bossu, Jelmer W. Poelstra, Matthias H. Weissensteiner, Alexander Suh, Alexey P. Kryukov, and Jochen B. W. Wolf. 2016. “Evolution of Heterogeneous Genome Differentiation across Multiple Contact Zones in a Crow Species Complex.” *Nature Communications* 7 (October): 13195. https://doi.org/10.1038/ncomms13195.

Wall, Jeffrey D. 2000. “A Comparison of Estimators of the Population Recombination Rate.” *Molecular Biology and Evolution* 17 (1): 156–63. https://doi.org/10.1093/oxfordjournals.molbev.a026228.

Weiler, Karen S., and Barbara T. Wakimoto. 1995. “Heterochromatin and Gene Expression in Drosophila.” *Annual Review of Genetics* 29 (1): 577–605. https://doi.org/10.1146/annurev.ge.29.120195.003045.

Weissensteiner, Matthias H., Andy W.C. Pang, Ignas Bunikis, Ida H?ijer, Olga Vinnere-Petterson, Alexander Suh, and Jochen B.W. Wolf. 2017. “Combination of Short-Read, Long-Read, and Optical Mapping Assemblies Reveals Large-Scale Tandem Repeat Arrays with Population Genetic Implications.” *Genome Research* 27 (5): 697–708. https://doi.org/10.1101/gr.215095.116.

Wicker, Thomas, Francois Sabot, Aurelie Hua-Van, Jeffrey L. Bennetzen, Pierre Capy, Boulos Chalhoub, Andrew Flavell, et al. 2007. “A Unified Classification System for Eukaryotic Transposable Elements.” *Nature Reviews Genetics* 8: 973–82.

Wolf, Jochen B. W., and Hans Ellegren. 2017. “Making Sense of Genomic Islands of Differentiation in Light of Speciation.” *Nature Reviews Genetics* 18 (2): 87–100. https://doi.org/10.1038/nrg.2016.133.

Yandell, Mark, and Daniel Ence. 2012. “A Beginner’s Guide to Eukaryotic Genome Annotation.” *Nature Reviews Genetics* 13: 329–42.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1762

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-369878

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2019