Chapter 3
Studying Speciation: Genomic Essentials and Approaches

Daronja Trense and Dieter Thomas Tietze

Abstract A genome comprises the entire genetic material of an organism and consists of DNA, which is in turn constructed of hundreds to billions of nucleotides. Nucleotides are organic molecules composed of three subunits: nitrogenous base, sugar (deoxyribose), and phosphate group. The DNA is differentiated into coding (genes) and noncoding regions. A gene is a specific region of DNA that encodes a function. All genes present within an organism represent its genotype. The genotype determines the phenotype, which is, however, additionally affected by the environment and the individual development (ontogenesis). A gene may affect a single or several phenotypic features (pleiotropy). Likewise, a phenotypic feature may be affected by one or several genes, with the latter comprising polygenic traits. In the process of gene expression, the information encoded by a gene is used to generate a product. The expression of genes is regulated by the external (temperature, stress, resource availability) and internal environment (metabolism, cell division cycle), and the gene-specific role in the respective tissue organism. Several processes underlying evolutionary change, e.g., mutation, genetic drift, gene duplication, selection, and migration, may change the genome at the level of single bases through genes to the organism. Such changes may result in population differentiation and eventually speciation. Molecular genetic studies on microevolution and speciation started with single genetic markers, e.g., the COI marker gene. Today, mainly genomic and transcriptomic approaches, making use of a large number of markers such as single nucleotide polymorphisms or microsatellites, are used to compare species, populations, and individuals.

D. Trense
Institute of Pharmacy and Molecular Biotechnology, Heidelberg University, Heidelberg, Germany

D. T. Tietze (✉)
Natural History Museum Basel, Basel, Switzerland
e-mail: thomas.tietze@bs.ch

Understanding the evolution of life is one of the most distinguished tasks of biological research. Recent advances in molecular techniques offer unprecedented opportunities to tackle such issues in a diverse array of taxa including birds, comprising our focal group here. Using genomic data requires detailed knowledge on the composition and function of each component of the genome. In the first section of this chapter, we therefore give an overview from the smallest elements including pentose sugars, double-ringed nitrogenous bases, and phosphate groups, via DNA, genes, and chromosomes to the entire avian genome. Furthermore, we cover functional aspects from DNA replication via transcription and translation through features of cells, tissues, and individuals, in the second section. The third section is about the evolution of the genome. We here highlight the mechanisms bringing about variation in individual genomes as well as the genomes of the next generation. In the last section, we will provide an overview about the components of the genomes that were or are used for understanding speciation, molecular systematics, and other research fields.

3.1 What Is an Avian Genome?

3.1.1 Structure of the Genetic Material

The deoxyribonucleic acid (DNA) is constructed of hundreds to billions of nucleotides, which in turn are constructed of nucleosides. A nucleoside consists of a purine or pyrimidine base linked to a pentose sugar, whereby purine is a double-ringed nitrogenous base such as adenine (A) and guanine (G) and pyrimidine a single-ringed nitrogenous base such as cytosine (C) and thymine (T) (Fig. 3.1) or uracil (U). If a nucleoside is linked to a phosphate group on either the 5′ or 3′ carbon on the deoxyribose, it is called a nucleotide. Two pentose sugar molecules each of two different nucleotide monomers are connected through an individual phosphate molecule, resulting in nucleotides being connected to a long chain. Such a chain creates a single strand of DNA with one end of the chain having a free 5′ and the other a free 3′ end. Two antiparallel and complementary strands can be connected by hydrogen bounds between guanine and cytosine or adenine and thymine, respectively (Alberts et al. 2014). If both DNA strands are wound around each other in an opposite direction, this is called the DNA double helix (Fig. 3.1). In eukaryotes, such as plants, mammals, birds, and many more, which are organisms whose cells have a nucleus and other organelles that are enclosed by membranes, the DNA is organized into chromosomes (Fig. 3.1) within the cell nucleus (plus another DNA molecule in each mitochondrion). Functionally, the genome is divided into genes, i.e., sequences of DNA that encode a single type of ribonucleic acid (RNA).

Due to diploidy of eukaryotic organisms, gene loci occur twice in eukaryotic genomes, as one maternally and one paternally inherited copy. An allele is one of
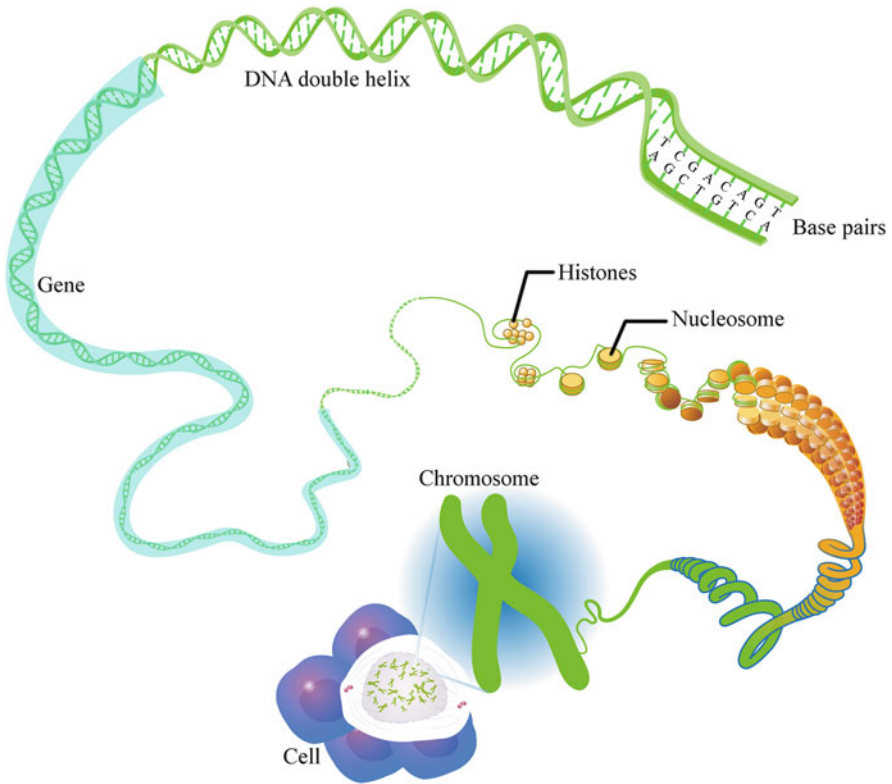
Fig. 3.1 Each cell contains a nucleus with chromosomes. These chromosomes comprise nucleosomes to pack the genetic material in the nucleus. The nucleosome consists of a DNA section, which is wound around the histone. Certain parts of the DNA (genes) carry information for the cell to encode a specific function. The DNA is structured in a double helix and consists of the four bases adenine (A), guanine (G), cytosine (C), and thymine (T) (The graphic was modified for this book chapter and was used from the National Human Genome Research Institute (NHGRI); website www.genome.gov)

several alternative forms of a gene occupying a given locus on a chromosome. All genes of one individual, which were transmitted from its parents, make up the genotype. The genotype produces the phenotype, which is the collection of all observable traits of one organism, e.g., height and eye color (Lesk 2012).

Three bases represent jointly a codon or triplet, and genes include a series of codons that are read sequentially from a starting point on one end to a termination point on the other end. Each triplet codes for a single amino acid in a corresponding protein. There are 64 (4 bases$^{3\text{ nucleotides}}$) possible codons but only 20 naturally occurring amino acids. This means that several codons correspond to the same amino acid (Alberts et al. 2014).

In contrast to the DNA, the RNA is evolutionarily older and has a different sugar (ribose) and a different base (uracil), which is replaced in the DNA by base thymine. The ribose makes the RNA less stable than DNA, and the production of uracil is less complex, because uracil is the unmethylated form of thymine (Alberts et al. 2014).

3.1.1.1 Noncoding and Coding Regions

The coding regions encode RNAs, which result in a protein, (messenger RNA, mRNA) or work directly in the case of other functional RNAs. The mRNA is a single-stranded RNA. The protein coding part(s) of a eukaryotic gene is/are the exon (s). The number of exons can vary. Introns separate the exons from each other such that the introns and exons alternate. After the splicing process, all introns will be removed and only exons remain. All exons are described by open reading frames (ORF), beginning with a start codon (ATG) and ending with a stop codon (Lesk 2012). Additionally, the 5′ and 3′ untranslated regions (UTR), which are the edges of the mRNA, do not code for parts of the protein. Introns tend to have a higher mutation rate than exons due to the fact that they do not encode part of a protein sequence. Thus, the sequence of an exon is more conserved than an intron sequence. Introns play an important role, because a single eukaryotic gene can code for several proteins, which can have different lengths due to alternative splicing.

The noncoding regions are parts of the DNA, which do not encode functional RNAs. Noncoding regions consist of transposable elements (TEs), retroviruses, and long and short interspersed nuclear elements (LINEs and SINEs), among others. TEs are selfish genetic elements, which either copy or paste through an RNA intermediate or directly cut and paste in their DNA form (Kapusta and Suh 2017). An abundant transposable element in birds is the CR1 element. Until now, 14 CR1 families have been described in birds (Kapusta and Suh 2017). TEs can be classified in LINEs and SINEs. LINEs are autonomous retrotransposons and consist usually of two ORFs (Kapusta and Suh 2017). SINEs are non-autonomous non-long terminal repeat retrotransposons, which parasitize LINEs (in birds 6000–17,000 SINEs versus 1,500,000 in humans or less than 0.1% of all avian genome sequences) (Kapusta and Suh 2017). Another example for a noncoding region is the retrovirus, an RNA virus, which can convert its sequence into DNA by reverse transcription (explained in Sect. 3.2.2). Endogenous viral elements (EVEs) are retroviruses that rely on obligate integration into the host genome and can be classified as LTR retrotransposons (Kapusta and Suh 2017).

3.1.1.2 Autosomes Versus Sex Chromosomes

A chromosome contains part of the genetic material of a eukaryotic organism and consists of chromatin, which is a complex of DNA and proteins. Most of these proteins are histones (Fig. 3.1), which wrap up the DNA in the nucleus. The number and appearance of chromosomes is called karyotype. Eukaryotic cells can be present either in a diploid or haploid condition. The term haploid means that chromosomes occur in a single set, while diploid cells have a double set of chromosomes. Most eukaryotic organisms have diploid cells; thus, all chromosomes appear twice. However, eukaryotic organisms differ in the number of chromosomes, for example, humans have 46 chromosomes. Furthermore, chromosomes are divided into

autosomes and sex chromosomes. Autosomes are pairs of chromosomes in a diploid cell, which have the same form, but each chromosome pair has a specific length. Humans have 44 autosomes and 2 sex chromosomes. Sex chromosomes differ from autosomes in length and function of their genes. They include the sex-determining region Y (SRY) gene on the Y chromosome. Furthermore, in humans, men have two different sex chromosomes, the X and Y chromosomes, while women have two X chromosomes. However, this XY sex-determination system is not present in all eukaryotes, in humans, most other mammals, several insects, some snakes, and a few plants. Another system is the ZW sex-determination system, which can be found in birds, several fishes, crustaceans, some insects, and reptiles. In the ZW sex-determination system, males have two Z chromosomes, while females have a Z and a W chromosome (Scanes 2015). Responsible for the sex determination in birds is probably a gene on the W chromosome, which is similar to the SRY gene on the Y chromosome. The Z and X chromosomes are larger and contain more genes than the W and Y chromosomes. Not only the sex chromosomes can be different in eukaryotic organisms, also the autosomes. In lizards, snakes, turtles, and birds, the autosomes can be divided in micro- and macrochromosomes (Matsubara et al. 2006; Ellegren 2013). Microchromosomes are tiny chromosomes with a length under 20,000,000 bp, while macrochromosomes are larger than 40,000,000 bp and resemble the mammalian autosomes in size. Characteristics of microchromosomes are that they include high rates of meiotic recombination, have high guanine-cytosine (GC) contents, short introns, high densities of genes, and **cytosine-phosphate-guanine (CpG) islands**, low densities of transposable elements and other repeats, but many repetitive sequences (Scanes 2015; Kapusta and Suh 2017). Another important aspect concerning chromosomes is the synteny, which describes the location of genetic loci on the same chromosome within an individual or species or even among species.

3.1.1.3 Nuclear Genome and Mitochondrial Genome

A common feature of eukaryotic organisms (except plants and other photosynthetically active eukaryotes) is that they have two genomes, the nuclear and the mitochondrial genome. The nuclear genome is organized in chromosomes as detailed above, while the mitochondrial genome is circularly or linearly organized and located within the mitochondria (that derived from bacteria). Generally, the nuclear genome is larger and contains many more genes than the mitochondrial genome.

3.1.2 The Chicken Model: History and Overview

An important model organism is the chicken *Gallus gallus domesticus*, because it is the first agricultural animal of which the genome was sequenced and has a relatively recent common ancestor with mammals. The ancestor of mammals and birds diverged 310 million years ago according to mitochondrial findings (Griffin et al.

2007). Furthermore, the chicken is the main laboratory model for over 10,828 extant bird species (Gill and Donsker 2017). The chicken genome has a size of 1.2 gigabases (Gb) (Lesk 2012), while avian genomes average around 1.35 Gb from the smallest, the Black-chinned Hummingbird *Archilochus alexandri* with 0.9 Gb, to the largest, the Common Ostrich *Struthio camelus* with 2.1 Gb (Scanes 2015; Kapusta and Suh 2017). The genome of the chicken was sequenced for the first time in 2004 with a 6.6 X **coverage**, using whole-genome shotgun reads. The resulting **assembly** of the chicken was composed of 933,000,000 bp and a genome size of 1.05 Gb (Hillier et al. 2004).

In chicken, it is difficult to identify genes on the W chromosome as well as on the microchromosomes due to the high number of repetitive sequences. However, the Z chromosome is well explored and contains nearly the same genes in all birds and is therefore highly conserved among bird species. On the Z chromosome of the chicken are about 1000 genes located, which are absent from the W chromosome. The W chromosome is degraded to different extents in some bird lineages (Marshall Graves 2015). This is why it is smaller, poorer in genes, but richer in repeats in most birds (Scanes 2015).

The avian karyotypes have been unusually stable during evolution, but there are some exceptions with chromosome numbers from 40 to 126 due to numerous microchromosomes (Griffin et al. 2007; Scanes 2015). A typical avian karyotype has a 2*n* of 76–80 (Ellegren 2013), and the chicken's haploid karyotype is defined by 39 chromosomes: chromosomes 1 through 10 are macrochromosomes, chromosomes 11 through 38 are microchromosomes, and the 39th chromosome is the sex chromosome (Hillier et al. 2004; Ellegren 2005).

In comparison to other eukaryotic organisms, the reduction of the avian genome size and transposable elements density, which is about 10% in avian genomes, began after the split of birds and crocodilians 250 million years ago (Griffin et al. 2007; Kapusta and Suh 2017). Thus, avian genomes are compact and were selected due to the evolution of flight (Hughes and Piontkivska 2005; Scanes 2015). In comparison with flightless birds, flying birds have a smaller genome. This might be due to the larger body size and longer generation times of flightless birds (Kapusta and Suh 2017).

Birds expanded their repertoire of keratin genes such as feather and claw keratins, and retained genes for egg production (Scanes 2015). The chicken genome has several genes encoding egg-related proteins, which are not represented in the mammalian genome. These are examples for gene losses in the mammalian lineage. On the other hand, there are some genes in chicken and humans that might have changed their function. In contrast to birds, which excrete uric acid, mammals excrete urea. Concomitantly, it seems that the function of some genes is altered in mammals, because genes encoding the enzymes of the mammalian urea cycle are also found in the chicken genome (Hillier et al. 2004).

The **alignment** of the chicken and human genome shows that at least 70,000,000 bp of sequence are likely to be functional in both species (Hillier et al. 2004). It is estimated that 20,000–23,000 protein-coding genes occur in the chicken genome (Hillier et al. 2004). In the human genome, some 20,000 genes have been detected until now. About 60% of protein-coding genes in chicken have a single

human ortholog. From these conserved genes in human and chicken, 72% are also conserved in the Japanese pufferfish *Takifugu rubripes* genome. Thus, these genes are most likely present in most vertebrates (Hillier et al. 2004).

3.2 How Does the Genome "Work"?

3.2.1 Replication of the DNA

DNA replication is the process of copying DNA within a cell of an organism. The process starts with the opening of the DNA double helix by the enzyme DNA helicase at a specific position by breaking the hydrogen bonds. Both DNA strands of the DNA double helix serve as a template for the replication of a new complementary strand. After opening, the enzyme DNA polymerase adds complementary nucleotides one by one to the growing DNA chain. The unwinding and adding of new nucleotides to the growing chain stops, if it reaches a region, which is either already replicated, or if a protein binds to the DNA sequence to stop the replication. Afterward, the new DNA strands will be checked by proofreading to remove the mismatches. The results of the DNA replication comprise two DNA double helices with one old and one new DNA strand. Apart from a very small number of copying errors, the two daughter molecules are identical in sequence with the original DNA molecule.

3.2.2 Transcription: RNA Synthesis

When a cell needs a specific protein, the transcription of the respective gene (copying DNA into RNA) starts, which is followed by translation of the nucleotide sequence into the amino acid sequence. The transcription begins with the opening of a small portion of the DNA double helix and its unwinding to display the bases. The enzyme RNA polymerase performs the transcription and knows its target position through a promotor, which is a specific nucleotide sequence of the DNA (Alberts et al. 2014). The promotor is located before the coding region and regulates the expression of genes. One strand of the DNA double helix acts as a template for the synthesis of the mRNA (Fig. 3.2). The sequence of the mRNA chain is defined by complementary base-pairing between free nucleotides and the DNA template. This DNA template is exactly complementary to the precursor messenger RNA (pre-mRNA). The transcription stops at a terminator, which represents the end of a gene (Alberts et al. 2014). Thus, the pre-mRNA is released. In eukaryotes, the pre-mRNA goes through several steps of processing such as polyadenylation, capping, and splicing. Polyadenylation adds a poly(A) tail to the pre-mRNA. This means that a specific number of adenine bases are added to the pre-mRNA. Capping of the pre-mRNA places a specific nucleotide and associated proteins to the 5′ end to stabilize the mRNA. Splicing removes the introns—intragenic regions—from the pre-mRNA; therefore, only exonic sequences exist in the mature mRNA.
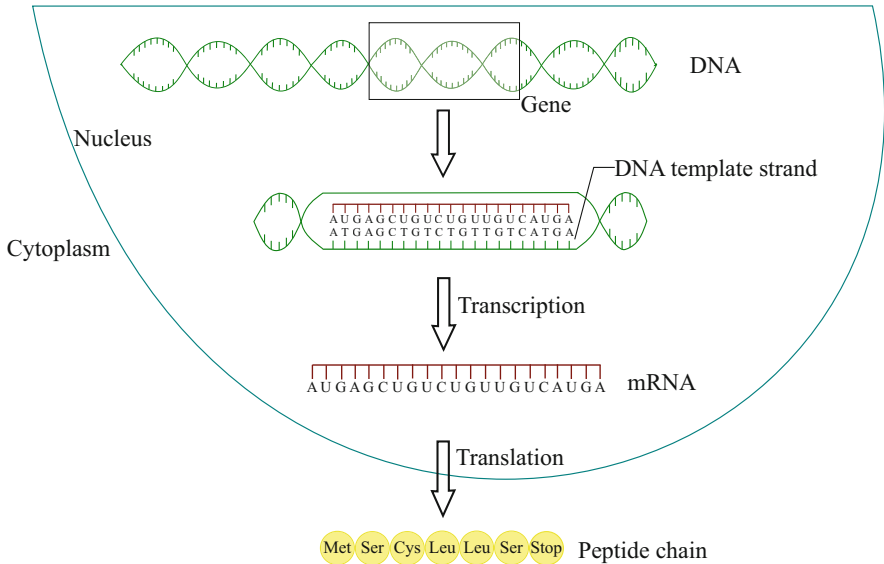
Fig. 3.2 A gene in the DNA provides on the template strand the nucleotide sequence that is transcribed into RNA (change from base thymine to uracil). This synthesis of RNA based on DNA is known as transcription. After the transcription, the mRNA will be released from the nucleus to the cytoplasm. In the cytoplasm, the translation proceeds with synthesizing a peptide chain (protein) based on the nucleotide sequence of the mRNA. In this example, the peptide chain consists of methionine (Met), serine (Ser), cysteine (Cys), leucine (Leu), and a stop codon, which leads to the termination of the translation

3.2.3 Translation

After these steps, the translation begins in the cytoplasm on ribosomes, which are complexes of proteins and ribosomal RNA (rRNA). RNA copies are used directly to synthesize the protein (Fig. 3.2) (Alberts et al. 2014). The information of the DNA (or mRNA sequence) comprises the genetic code, which is read by small RNA molecules, the transfer RNA (tRNAs). The tRNA attaches to one end to a specific amino acid and displays at the other end a specific nucleotide triplet, the anticodon. This anticodon recognizes, due to base pairing, a codon in the mRNA. A stop codon is a nucleotide triplet, which has no corresponding tRNA (Alberts et al. 2014); thus, reaching the stop codon in the mRNA terminates translation. Proteins are important for development and functioning: They form parts and build the structure of an organism; perform metabolic reactions, which are necessary for life; participate in regulation as transcription factors and receptors; are key players in signal transduction pathways; and can act as enzymes to catalyze chemical reactions.

3.2.4 One Gene: One Function?

Historically, it has been assumed that each gene encodes a single function. Today though it is well-known that one gene may have different functions. For instance, some genes encode only a subunit of a protein, because several proteins consist of polypeptides encoded by different genes. In other cases, genes do not encode polypeptides, but functional RNA molecules. Furthermore, genes can encode several proteins due to alternative splicing, which is a process following the actual transcription in eukaryotes. During alternative splicing, some exons can be excluded from the pre-mRNA. Thus, different proteins can be coded by the very same gene. This implies that one gene can influence more than one and even unrelated phenotypic features in one individual (pleiotropy). On the other hand, different genes may influence the same (polygenic) phenotypic feature in one individual.

3.2.5 Categorical vs. Quantitative Traits

A trait is defined as a feature of one individual, and this feature can be characterized by an attribute of the physical appearance (e.g., feather color) or a special behavior of the individual (e.g., alarm calls). These traits may be influenced by one or many genes. A categorical trait can be present or absent (e.g., feather crest), depending on the presence or absence of specific genes or alleles. Or in the case of multiple states, trait values can be categorized, for example, as white or black or yellow plumage. In contrast to categorical traits, quantitative traits show no categories, but continuous variation such as beak length.

3.2.6 Phenotypic Plasticity

The phenotypic variation encountered within and among populations may be caused by genetic or environmental factors. Genetically controlled phenotypic variation is caused by genetic polymorphisms, though not all genetic polymorphisms, e.g., at selectively neutral loci, are imperative for phenotypic variation. However, different phenotypes may alternatively persist within a population due to variation under certain environmental conditions (Pigliucci 2001). If such environmentally induced phenotypes result in different morphs, they are referred to as polyphenisms. Polyphenisms are, thus, the result of phenotypic plasticity, which is defined as the ability of a single genotype to produce different phenotypes in different environments (Pigliucci 2001; West-Eberhard 2003). Such changes include modifications of developmental processes as well as in adult phenotypes in response to environmental stimuli. As phenotypic plasticity may quickly change phenotypic traits, it enables an organism to respond to changing environments (Merilä and Hendry 2014). Environmental variation may induce changes in behavior, morphology, or physiology,

which may be transient or irreversible. More importantly, phenotypic plasticity may be adaptive or reflect nonadaptive interactions between an organism and its environment (Pigliucci 2001). If adaptive, plasticity alters the fitness of an organism under specific environmental conditions. Consequently, phenotypic plasticity may play an important role in adaptive evolution (Fusco and Minelli 2010).

It may, for instance, shield genotypes from selection, thus slowing down evolutionary rates, or, alternatively, facilitate adaptive evolution through genetic assimilation of environmentally induced phenotypes (Ghalambor et al. 2007). Furthermore, note that effects of genes and the environment may be easily confused. Some environmental conditions may produce phenotypes similar to those produced by genetic factors and vice versa. Finally, both environmental conditions and genetic constitution interact with one another to generate the best adapted phenotype (Fusco and Minelli 2010).

3.3 How Does the Genome Evolve?

3.3.1 Modification of the DNA

DNA methylation is a chemical modification of chromatin. In the methylation process, small molecules (i.e., methyl groups consisting of one carbon atom and three hydrogen atoms) attach to the DNA. If a methyl group is attached to a part of a gene, the gene will be turned off. Modifying the wrong gene or other failures can result in abnormal gene activity or false inactivity of a gene. These errors in the epigenetic processes can lead to, for example, cancer and metabolic disorders. Epigenetics examined why the expression of the gene is activated at a specific time in the development of an organism. Furthermore, it describes an inheritable phenotype, which is created from changing chromosomes without alterations of their DNA sequences (Toraño et al. 2016).

Not only DNA methylation but also acetylation and phosphorylation can result in similar changes. Acetylation is the reaction of an acetyl functional group into a chemical compound, while protein phosphorylation is a modification of a protein by which an amino acid is phosphorylated by the addition of a phosphate group. Both modifications are important in biological regulation such as gene and enzyme regulation.

Another type of modification affects the histones by the attachment of chemical compounds. These chemical compounds can be used by other proteins to decide, if a DNA sequence should be active or ignored within in a specific cell. Covalent histone modifications generate or stabilize the location of specific binding partners to chromatin, while non-covalent mechanisms provide the cell with further tools for introductory changes into the chromatin template. Chromatin remodeling and inclusion of specialized histone variation are examples for non-covalent mechanisms. However, covalent and non-covalent mechanisms can also be combined (Goldberg et al. 2007).

3.3.2 Mutation

Biological diversity would not exist without some degree of error in the hereditary process. Such errors occur from the higher level of karyotype down to the DNA base sequence. The rate at which changes occur in DNA sequences is defined as mutation rate (Alberts et al. 2014). A mutation in which one pyrimidine base is replaced by the other or in which one purine base is replaced by the other is called transition. A transversion is a mutation in which a purine base is replaced by a pyrimidine base or the other way around. A point mutation changes only a single base. A synonymous mutation appears, if the substitution does not change the amino acid sequence of the polypeptide product. This is a type of silent mutation, which is more likely to be fixed by drift. A non-synonymous mutation in a coding region does change the sequence of the amino acid and therefore the polypeptide product. This could result in either the production of a different amino acid or a nonsense or termination codon. Selectively neutral mutations occur, when changes in coding regions have no effects on the phenotype. Further modes of mutation include insertions and deletions of a single base or a sequence of bases. An insertion can be reverted by deletion of the inserted sequence, but a deletion of a sequence cannot be reverted in the absence of some mechanism to restore the lost sequence. There have been very precise repair mechanisms for billions of years, but some mutation rate remains. In most cells, this only affects the actual individual, but in germline cells (egg and sperm cells and their precursors), this leads to genotypic changes in the offspring and potentially their phenotype.

3.3.3 Selection

Selection is a process, which acts on the phenotype and can benefit individuals with a certain feature or genotype. This leads to the spread of traits beneficial to survival and reproduction while eliminating detrimental ones. Individuals with advantageous traits have a higher chance to survive and produce more offspring than individuals with unfavorable traits. A negative or purifying selection eliminates new mutations, because the phenotype is negatively affected by the mutation. If an individual with an advantageous mutation survives, it can produce more fertile offspring than individuals without the mutation (positive selection). Sexual selection is an individual's choice of mates of the other sex from the same species by preferring a presumably advantageous feature. This often led to an arms race (e.g., in plumage coloration, song variability).

3.3.4 Genetic Drift

Genetic drift is a random change in the frequency of a heritably gene variant (allele) in a given population. It may occur in all populations, but its strength is strongly

dependent on population size: The smaller the population, the larger the effect of genetic drift. Thus, genetic drift does not depend on specific alleles, either beneficial or harmful. Genetic drift may even lead to the fixation of a harmful allele or the disappearance of beneficial alleles and generally reduces the genetic variation within a population or species. Genetic drift is often associated with founder effects (e.g., settlement on a small island) and population bottlenecks (e.g., glaciation reducing the inhabitable area), owing to the concomitantly reduced population size.

3.3.5   Geographic Variation and Dispersal

The physiological or morphological variation based on genetic features between populations of the same species in its whole range is called geographic variation. Geographic variation may often result from local adaptation, with specific genetic factors of a population being favored by natural selection.

Dispersal means the range expansion of a population by individuals that are adapted to new habitats or places.

3.3.6   Recombination and Migration

The exchange of genetic material either between chromosomes or different regions on the same chromosome is called recombination. Recombination creates new combinations of alleles and genes and gives rise to much of the genetic variability within populations due to different combinations in offspring compared with their parents. In sexually propagating organisms such as birds and humans, it occurs in every generation during the preparation of the germ cells (eggs and sperms). This forms the basis for adaption to changing environmental conditions.

Migration is defined as the change of gene frequency by introducing new allele or more copies of one alleles into a population by a migrant.

3.3.7   Gene Duplication

A duplication of a DNA section, which contains a gene, is defined as gene duplication. Gene duplication can occur during the processes of DNA replication and recombination or when an mRNA is converted back to DNA and new genes integrate into the genome. Gene duplication may allow for the development of a new function. Gene duplication may affect a phenotype, e.g., copies of a gene can lead to a surplus of the gene-specific protein, because the amount of a synthesized protein is regularly proportional to the present number of gene copies (Clancy and Shaw 2008).

Another type of duplication is the duplication of whole chromosomes. This process can occur during cell division when the chromosomes do not separate correctly between the two cells.

3.4 How to Study Speciation Using Genomic Features?

The first molecular markers for species delimitation and taxonomy were isozymes and allozymes. Isozymes describe different molecular forms of an enzyme, which are encoded by different loci. In contrast, allozymes characterize different molecular forms of an enzyme, produced by different alleles at the same locus (Duminil and Di Michele 2009). The term locus refers to a specific position of a gene, while the term gene is related to a DNA section, which contains the information to produce an RNA molecule. The principle approach when using allozymes or isoenzymes is to identify the variation of an enzyme among individuals using electrophoresis. However, nowadays almost exclusively DNA markers instead of protein markers are used for speciation studies because of low resolution due to synonymous mutations.

3.4.1 PCR-Based Molecular Markers

DNA markers can be **codominant** or dominant such as amplified fragment length polymorphisms (AFLP), restriction fragment length polymorphism (RFLP), and random amplified **polymorphic** DNA (RAPD) (Duminil and Di Michele 2009). AFLP studies use restriction enzymes, which digest genomic DNA, followed by the **ligation** of adapters to the **sticky ends** of the restriction fragments. A selection of the restriction fragments will be amplified with polymerase chain reaction (PCR) primers, which have a corresponding adaptor and restriction-site specific sequences. Afterward, the amplicons will be separated through electrophoresis on a gel and visualized. RFLP is a technique, which starts with the cutting process of DNA fragments by restriction enzymes, followed by a gel electrophoresis to order the DNA fragments by their length. RAPD is a special method of the PCR, because it uses short primers and the results are random DNA sequences. The gel electrophoresis shows individual patterns.

There are **codominan**t molecular markers, which can be used for species delimitation and taxonomy. For most of these markers, the PCR method is used to multiply a specific DNA sequence of a sample. The method starts with the denaturation of the double-stranded DNA into single strands, called templates. Short DNA sequences, which are generally 18–20 bp long and are known as primers, bind to the templates. This step is called annealing. The next step is elongation, in which the enzyme DNA polymerase synthesizes a new DNA strand, which is complementary to the template, by adding free nucleotides to the single DNA strand. Afterward, the annealing and elongation are repeated in a definite number of cycles, until enough target DNA

sequences are available (Semagn et al. 2006). This method can be used to sequence and analyze different DNA sequences for a variety of scientific questions.

3.4.1.1 Ribosomal Genes

The nuclear rDNA encodes rRNA, and both contain highly conserved and variable domains, which is a good condition for analyzing phylogenetic relationships (Hwang and Kim 1999; Patwardhan et al. 2014).

The nuclear small subunit (SSU) rDNA is a highly conserved region of the DNA, which has been used for the reconstruction of phylogenetic relationships in kingdoms, phyla, classes, and orders. The nuclear large subunit (LSU) rDNA contains more variation than the SSU rDNA, and the size of its genes varies among phyla. The LSU rDNA is used for studying genetic relationships in orders and families (Hwang and Kim 1999). Further highly conserved regions like the nuclear SSU rDNA are the 12S and 16S rDNA. They encode the ribosomal RNA, which is part of the small ribosomal subunit of a ribosome in a mitochondrion. The 12S rDNA has been used to study the phylogeny of phyla and subphyla, while the 16S rDNA has been used for analyzing the phylogenetic relationships within families and genera, because the 16S rDNA is more variable than the 12S rDNA (Hwang and Kim 1999).

3.4.1.2 Mitochondrial DNA Markers

Due to the fact that the mitochondrial DNA evolves faster than the nuclear genome, mitochondrial protein-coding regions have been used for analyzing the phylogenetic relationships within families, genera, and species (Hwang and Kim 1999). The first mitochondrial marker used was the control region, which is located in the noncoding region and is part of the regulation and initiation of the mitochondrial DNA replication and transcription (Patwardhan et al. 2014). The mitochondrial control region is variable in size and contains many variations also between individuals of the same species. Thus, it is used for studying genetic relationships in species, subspecies, and populations (Hwang and Kim 1999).

The second mitochondrial marker was the cytochrome oxidase I/II (COI/II), which is a well-known protein of an electron transport chain. In the cytochrome c oxidase complex, the COI and COII genes code for two polypeptide subunits. Both have been used for phylogenetic relationships among orders, families, subfamilies, genera, and species. The sequence of the COI gene is one of the sequences that can be used as a barcode for the identification of species (Patwardhan et al. 2014). DNA barcoding is a method to identify species by using short sequences.

Further widely used mitochondrial markers to reconstruct the phylogeny among genera and species are the cytochrome b (cytb) and NADH dehydrogenase 2 (nd2) genes.

3.4.1.3 Microsatellites

A microsatellite is a specific DNA motif with a length of two to six base pairs (Fig. 3.3). Microsatellites are used to detect the number of repeats of a sequence to identify an individual. Similar to microsatellites are minisatellites, but their repeat motifs are longer. Microsatellites can be amplified by PCR, for which labeled primers are needed, followed by analyzing the length of the fragment (microsatellite). A large advantage is the small amount of DNA needed for the PCR. Microsatellites are locus-specific, **codominant**, and highly **polymorphic**. A disadvantage of microsatellites is their taxon-specificity. Thus microsatellite libraries need to be generated for each species or closely related sister species (Delaney 2014). Microsatellites are currently mainly used for paternity tests and population genetics but hold large potential for speciation studies due to their potential to distinguish lineages within a species. It is necessary to work with more than one microsatellite locus to have reliable results.

3.4.2 Expressed Sequence Tags

Genes must be converted into mRNA, but RNA is unstable outside the cell. Hence, mRNA needs to be converted into complementary DNA (cDNA) by the reverse transcriptase enzyme. The production of cDNA is the reverse process of transcription, because mRNA is used as the template instead of the DNA. cDNA is more stable than mRNA and contains generally only exons due to splicing of the pre-mRNA. This means that cDNA represents an expressed gene or a part of it. When the cDNA has been isolated, various nucleotides can be sequenced to create expressed sequence tags (ESTs) with a length of 100–800 bp. They allow the discovery of unknown genes and a comparison between different species due to high conservation in the coding regions (Semagn et al. 2006). From ESTs it is possible to develop primer pairs for sequencing genes in other species and to detect single nucleotide polymorphisms (SNPs) (Schlötterer 2004; Semagn et al. 2006).

3.4.3 Single Nucleotide Polymorphisms

A single nucleotide polymorphism (SNP) is the change of a single base in the DNA sequence (Fig. 3.3) (Semagn et al. 2006). Generally, two different nucleotides can be found per position, and SNPs mostly occur in noncoding regions (Grover and Sharma 2016). The simplest method to identify SNPs is to screen a high-quality DNA sequence or an EST. The most common methods like restriction-site-associated DNA sequencing (RAD-seq) and genotyping by sequencing (GBS) will be explained in the following two sections. A comprehensive strategy for detecting SNPs in a genome is the generation of shotgun genome sequences. For this method,

	SNP 1	**SNP 2**	**Microsatellite 1**

Individual 1 AGTCTTGATCGAATCATGACTGACTCAGGTCCTAGTAACTCACACACACACAGTCAACTGA
Individual 2 AGTCTTGATCGAATCATGACTGACTCTGGTCCTAGTAACTCACACACACACAGTCAACTGA
Individual 3 AGTCTTGATCGAATTATGACTGACTCAGGTCCTAGTAACTCACACACACACACAGTCAACTGA
Individual 4 AGTCTTGATCGAATCATGACTGACTCTGGTCCTAGTAACTCACACACACACACAGTCAACTGA
Individual 5 AGTCTTGATCGAATTATGACTGACTCTGGTCCTAGTAACTCACACACACACAGTCAACTGA
Individual 6 AGTCTTGATCGAATTATGACTGACTCAGGTCCTAGTAACTCACACACACAGTCAACTGA
Individual 7 AGTCTTGATCGAATTATGACTGACTCTGGTCCTAGTAACTCACACACACAGTCAACTGA
Individual 8 AGTCTTGATCGAATCATGACTGACTCTGGTCCTAGTAACTCACACACACACAGTCAACTGA
Individual 9 AGTCTTGATCGAATCATGACTGACTCAGGTCCTAGTAACTCACACACACACAGTCAACTGA
Individual 10 AGTCTTGATCGAATTATGACTGACTCAGGTCCTAGTAACTCACACACACACAGTCAACTGA

Fig. 3.3 Ten individuals from one population are represented with fictional sequences. In these sequences, two single nucleotide polymorphisms (SNPs) and one microsatellite occur. The first SNP is a variation of the bases cytosine (C) and thymine (T). The individuals 1, 2, 4, 8, and 9 carry base C, while the other individuals have a T at the same position. In the individuals 1, 3, 6, 9, and 10, an adenine (A) appears as the second base, whereas the individuals 2, 4, 5, 7, and 8 have the base T on this position. The microsatellite in this example is a repetition of two bases, C and A. In the individuals 1, 2, 5, 8, 9, and 10, it is 12 bases long $(CA)_6$. In the individuals 3 and 4, the microsatellite is 14 bp long $(CA)_7$, while it is shorter in the individuals 6 and 7 $(CA)_5$

a pool of DNA from different individuals should be sequenced. A more efficient approach is the shotgun sequencing with a reduced section of the genome, in which the DNA of many different individuals can be sequenced (Schlötterer 2004). Most of these methods are cost- and time-intensive and the information content of one SNP is very low, but they have a low mutation rate (high stability) and high frequency in the genome, and new analytical methods are being developed and open up new opportunities.

SNPs can be used in different research questions, e.g., investigate about natural selection across species (Künstner et al. 2010), examine recent divergence (McCormack et al. 2012), explore the genetic structure of different morphological features in different species (Silva et al. 2017), and investigate hybridization (Manthey et al. 2016).

3.4.4   Restriction-site-associated DNA sequencing

Restriction-site-associated DNA sequencing (RAD-seq) is the genotyping of short DNA fragments, which are adjacent to the cut site of a restriction enzyme (RE). The first step of RAD-seq is the digestion of the genomic DNA with a chosen RE, followed by the ligation of an adapter (P1) to the overhang of the RE (Baird et al. 2008; Davey and Blaxter 2011). This adapter contains a binding site for the forward primer and a barcode for the sample identification. After ligation, the fragments are pooled and size selected (Baird et al. 2008). The DNA fragments are then ligated to a second adapter (P2), which has a reverse primer site and is a Y adapter with divergent ends (Coyne et al. 2004; Baird et al. 2008). The reason for choosing a Y adapter is that all fragments contain the P1 adapter, because the P2 adapter cannot bind to the reverse primer, before the amplification of the P1 adapter has been

finished (Baird et al. 2008; Davey and Blaxter 2011). After ligation of the second adapter, a PCR reaction is performed. The PCR-products are used for next-generation sequencing (3.4.7) (Baird et al. 2008). The resulting reads are trimmed, grouped by barcodes, and mapped to a reference genome or, if no reference genome is available, the same reads are aligned for identifying SNPs (Baird et al. 2008; Davey and Blaxter 2011). The challenges of RAD-seq are the high costs of sequencing and the diversity of RAD-seq protocols with different technical details. Nevertheless, one can choose the protocol most suitable for the own study system or research question (Andrews et al. 2016). RAD-seq can identify and generate thousands of genetic markers, reduces the complexity of the genome, and can be used for species with no or limited existing sequence data (Davey and Blaxter 2011). Furthermore, RAD-seq was extended to use two REs instead of one RE to exclude the step of size selection. This method is called double digest RAD-seq (Peterson et al. 2012).

3.4.5   Genotyping by sequencing

Genotyping by sequencing (GBS) is a highly multiplexed approach for constructing reduced representative libraries for the Illumina next-generation sequencing platform to discover a large number of SNPs. This approach can be used for any species at a low per-sample cost and also incorporates restriction enzymes (RE) to reduce genome complexity (Elshire et al. 2011; Chung et al. 2017). The procedure of GBS like RAD-seq starts with the digestion of DNA by an RE. The selected REs should be suitable for the investigated species by containing an overhang of two to three base pairs, and REs do not cut frequently in the major repetitive fraction of the genome. After the digestion, two adapters are ligated to the ends of the digested DNA. The adapters should be complementary to the overhang of the chosen RE, and one adapter contains a barcode for multiplex sequencing. These adapters contain binding sites for appropriate primers, which are added to perform a PCR reaction to increase the amount of DNA fragments. The PCR products are cleaned up and DNA fragments with a specific size result in a library. Libraries are used for sequencing, followed by filtering reads, which match one of the barcodes and the corresponding cut site of the RE, and are not adapter dimers. These sorted reads are separated by their barcode and after separation the barcode is removed. The filtered reads are mapped to the reference genome, consequently reads, which mapped on the same position are aligned and used to identify SNPs (Elshire et al. 2011). GBS is a cost-effective method to discover SNPs, genotype individuals within a population, and detect molecular markers. The disadvantages are the management of big datasets and the fact that the data do not represent the whole genome, which could have a negative effect on constructing genetic maps (Chung et al. 2017).

3.4.6 Transcriptomics

This is a technique to study an organism's transcriptome, which is the total of all its RNA transcripts. The transcriptome is a snapshot at a specific time of all transcripts in one cell or tissue, for a specific developmental stage. These expressed genes of one organism in different cells, tissues, conditions, or time points give details about the function of uncharacterized genes and the biology of organisms. Furthermore, the comparison of transcriptomes allows the identification of genes, which are expressed in different cells; hence, it gives information about gene regulation. There are two techniques to create a transcriptome: microarrays and RNA-Seq. The microarray approach quantifies a set of predefined sequences, while the RNA-Seq technique uses next-generation sequencing to target "all" expressed genes (Wang et al. 2009).

3.4.7 "Whole" Genome Sequencing

Next-generation sequencing (NGS) is a method to produce a large number of reads of short DNA sequences, between 50 and 150 bp long. The read length of NGS is often short with a high error rate, but this is compensated due to a higher **coverage** of the consensus sequence (Scanes 2015). These reads can be combined to continuous sequences (contigs), and contigs can be in turn linked to scaffolds. Indications about the quality of contigs and scaffolds (genome assemblies) can be provided by the N50 value, which represents the minimum length of long sequences that make up half of the **assembly** of contigs or scaffolds (Kapusta and Suh 2017). Contigs and scaffolds can be used to identify genes, but there are sequences which have no genetic information, which are clustered in chromosome Unknown (chrUn). Annotation is the process of linking DNA reads to information available from previous work (on other taxa) (Scanes 2015).

3.4.7.1 Different Strategies for Sequencing Genomes

The traditional Sanger sequencing with 1-kb-long sequence reads and the Roche 454 sequencing with up to 800 bp sequence reads have been largely replaced by short-read technologies such as Illumina HiSeq with 150 bp sequence reads. There are also even newer technologies available such as Pacific Biosciences with up to 5 kb sequence reads or Ion Torrent with about 500 bp sequence reads (Ekblom and Wolf 2014). The technology of $10\times$ genomics uses short reads from Illumina sequencing to link the short reads to long molecules. In the long molecules, variation can be detected to identify which reads belong to the father or mother of the examined individual. Another method uses single molecules by detecting them and sequencing their DNA. This is called single-molecule genomics.

One of the most common strategies for genome sequencing is the shotgun sequencing. First, DNA is cut into small random fragments, whereby the size of the fragments depends on the technology used. These fragments will be assembled to a longer contig. This process is known as de novo **assembly**. It is important that there is enough overlap between the sequence reads for a correct **assembly**, and this implies also a high **coverage**. If there are longer fragments like several hundred base pairs, both ends of the sequence will be sequenced called paired-end sequencing. Afterward, the resulted contigs are connected to longer sequences (scaffolds) (Ekblom and Wolf 2014).

The genome annotation uses the whole genome sequences in combination with relevant information from gene models, functional information, microRNA, or epigenetic modifications. Consequently, a lack of genomic information will result in low annotation rates. Annotation describes the process of using data of other genomes or transcriptomes to detect genes or transcripts on the newly assembled genome (Ekblom and Wolf 2014).

3.4.7.2 Limitations of Analyzing Genomes

Usually, a genome draft represents the complete nucleotide base sequence for all chromosomes in one species. Nevertheless, there is not just one sequence for a species, due to individual genomic variation, differences among cells within individuals due to diploidy. Thus, the assembled reference genome sequence of one individual will only comprise a subset of the total variation present within a species. Typically, one individual is sequenced, but sometimes a genome is based on a consensus of a few individuals (Ekblom and Wolf 2014). Furthermore, it is not possible to sequence and assemble all nucleotides in the genome due to sequencing errors (Scanes 2015), and most genome **assembly** methods fail on repetitive elements, which are typically not included in reference genomes (Hoban et al. 2016). However, repetitive regions may be characterized through the annotation of a comprehensive dataset compounded of a high-**coverage** single molecule real-time sequencing **assembly**, an assembled optical map, and a generated high-**coverage** short-read sequence **assembly** to a repeat library (Weissensteiner et al. 2017).

3.4.8 Epigenome

In almost all cells of an individual, the same DNA sequence can be found, but nevertheless cells may differ as the information content encoded within the DNA may be used differently. Such differences may arise from chemical modifications of the DNA or histone proteins without changing the DNA sequence. The resulting epigenome includes chemical compounds, which have been added to the DNA to regulate gene activity. These chemical compounds are not part of, but fixed to, the DNA. Epigenomic changes occur in individual development and tissue

differentiation and may result in cell division, and, in some circumstances, they can be transferred to the next generation. However, the epigenome can also be influenced by environmental conditions, such that the epigenome may vary between individuals. Through epigenetic changes, genes can be turned off or on (expression), thus determining the production of proteins in specific cells. For example, the eye is specialized for light-sensitive proteins and red blood cells for carrying oxygen. Furthermore, epigenetic changes in DNA and histones play a role in regulatory pathways of eukaryotes (Marshall Graves 2015).

3.5 Closing Words

Speciation is one of the main focusses in evolutionary biology and also the starting point to clarify the relationship between species. Morphological traits and reproduction are important for characterizing one species, but over the last decades, genetic tools got more and more influence in the delimitation of species. Therefore, it is necessary to understand the structure and function of the used genetic material. The genetic investigation of speciation began with short sequences and few genes for small sample sizes. Nowadays, more individuals of one species and additionally more species can be examined. Furthermore, SNPs, transcriptomes, and whole genomes are the newest traits to analyze and understand speciation—also in functional respect. However, methods will be further developed to become more cost-effective, faster, and more informative.

References

Alberts B, Johnson A, Lewis J, Morgan D, Raff M, Roberts K, Walter P (2014) Molecular biology of the cell. Garland Science, Taylor & Francis Group

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet 17(2):81–92

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA, Fay JC (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3(10):e3376

Chung YS, Choi SC, Jun T-H, Kim C (2017) Genotyping-by-sequencing: a promising tool for plant genetics research and breeding. Hortic Environ Biotechnol 58(5):425–431

Clancy S, Shaw KM (2008) DNA deletion and duplication and the associated genetic disorders. Nat Educ 1:23

Coyne KJ, Burkholder JM, Feldman RA, Hutchins DA, Cary SC (2004) Modified serial analysis of gene expression method for construction of gene expression profiles of microbial eukaryotic species. Appl Environ Microbiol 70(9):5298–5304

Davey JW, Blaxter ML (2011) RADSeq: next-generation population genetics. Brief Funct Genomics 9(5–6):416–423

Delaney KS (2014) Landscape genetics of urban bird populations. In: Gil D, Brumm H (eds) Avian urban ecology. Oxford University Press

Duminil J, Di Michele M (2009) Plant species delimitation: a comparison of morphological and molecular markers. Plant Biosyst 143:528–542. https://doi.org/10.1080/11263500902722964

Ekblom R, Wolf JBW (2014) A field guide to whole-genome sequencing, assembly and annotation. Evol Appl 7:1026–1042. https://doi.org/10.1111/eva.12178

Ellegren H (2005) The avian genome uncovered. Trends Ecol Evol 20:180–186. https://doi.org/10.1016/j.tree.2005.01.015

Ellegren H (2013) The evolutionary genomics of birds. Annu Rev Ecol Evol Syst 44:239–259. https://doi.org/10.1146/annurev-ecolsys-110411-160327

Elshire RJ, Glaubitz JC, Qi S, Poland JA, Kawamoto K, Buckler ES, Mitchell SE, Orban L (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6(5):e19379

Fusco G, Minelli A (2010) Phenotypic plasticity in development and evolution: facts and concepts. Introduction. Philos Trans R Soc Lond Ser B Biol Sci 365:547–556. https://doi.org/10.1098/rstb.2009.0267

Ghalambor CK, McKay JK, Carroll SP, Reznick DN (2007) Adaptive versus non-adaptive phenotypic plasticity and the potential for contemporary adaptation in new environments. Funct Ecol 21(3):394–407

Gill F, Donsker D (eds) (2017) IOC world bird list (v7.1). https://doi.org/10.14344/IOC.ML.7.1

Goldberg AD, Allis CD, Bernstein E (2007) Epigenetics: a landscape takes shape. Cell 128:635–638. https://doi.org/10.1016/j.cell.2007.02.006

Griffin DK, Robertson LBW, Tempest HG, Skinner BM (2007) The evolution of the avian genome as revealed by comparative molecular cytogenetics. Cytogenet Genome Res 117:64–77. https://doi.org/10.1159/000103166

Grover A, Sharma PC (2016) Development and use of molecular markers: past and present. Crit Rev Biotechnol 36:290–302. https://doi.org/10.3109/07388551.2014.959891

Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, Bork P, Burt DW, Groenen MAM, Delany ME, Dodgson JB, Chinwalla AT, Cliften PF, Clifton SW, Delehaunty KD, Fronick C, Fulton RS, Graves TA, Kremitzki C, Layman D, Magrini V, McPherson JD, Miner TL, Minx P, Nash WE, Nhan MN, Nelson JO, Oddy LG, Pohl CS, Randall-Maher J, Smith SM, Wallis JW, Yang S-P, Romanov MN, Rondelli CM, Paton B, Smith J, Morrice D, Daniels L, Tempest HG, Robertson L, Masabanda JS, Griffin DK, Vignal A, Fillon V, Jacobbson L, Kerje S, Andersson L, Crooijmans RPM, Aerts J, van der Poel JJ, Ellegren H, Caldwell RB, Hubbard SJ, Grafham D V., Kierzek AM, McLaren SR, Overton IM, Arakawa H, Beattie KJ, Bezzubov Y, Boardman PE, Bonfield JK, Croning MDR, Davies RM, Francis MD, Humphray SJ, Scott CE, Taylor RG, Tickle C, Brown WRA, Rogers J, Buerstedde J-M, Wilson SA, Stubbs L, Ovcharenko I, Gordon L, Lucas S, Miller MM, Inoko H, Shiina T, Kaufman J, Salomonsen J, Skjoedt K, Wong GK-S, Wang J, Liu B, Wang J, Yu J, Yang H, Nefedov M, Koriabine M, deJong PJ, Goodstadt L, Webber C, Dickens NJ, Letunic I, Suyama M, Torrents D, von Mering C, Zdobnov EM, Makova K, Nekrutenko A, Elnitski L, Eswara P, King DC, Yang S, Tyekucheva S, Radakrishnan A, Harris RS, Chiaromonte F, Taylor J, He J, Rijnkels M, Griffiths-Jones S, Ureta-Vidal A, Hoffman MM, Severin J, Searle SMJ, Law AS, Speed D, Waddington D, Cheng Z, Tuzun E, Eichler E, Bao Z, Flicek P, Shteynberg DD, Brent MR, Bye JM, Huckle EJ, Chatterji S, Dewey C, Pachter L, Kouranov A, Mourelatos Z, Hatzigeorgiou AG, Paterson AH, Ivarie R, Brandstrom M, Axelsson E, Backstrom N, Berlin S, Webster MT, Pourquie O, Reymond A, Ucla C, Antonarakis SE, Long M, Emerson JJ, Betrán E, Dupanloup I, Kaessmann H, Hinrichs AS, Bejerano G, Furey TS, Harte RA, Raney B, Siepel A, Kent WJ, Haussler D, Eyras E, Castelo R, Abril JF, Castellano S, Camara F, Parra G, Guigo R, Bourque G, Tesler G, Pevzner PA, Smit A, Fulton LA, Mardis ER, Wilson RK. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432:695–716. DOI: https://doi.org/10.1038/nature03154.

Hoban S, Kelley JL, Lotterhos KE, Antolin MF, Bradburd G, Lowry DB, Poss ML, Reed LK, Storfer A, Whitlock MC (2016) Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. Am Nat 188:379–397. https://doi.org/10.1086/688018

Hughes AL, Piontkivska H (2005) DNA repeat arrays in chicken and human genomes and the adaptive evolution of avian genome size. BMC Evol Biol 5:12. https://doi.org/10.1186/1471-2148-5-12

Hwang U-W, Kim W (1999) General properties and phylogenetic utilities of nuclear ribosomal DNA and mitochondrial DNA commonly used in molecular systematics. Korean J Parasitol 37:215–228. https://doi.org/10.3347/kjp.1999.37.4.215

Kapusta A, Suh A. 2017. Evolution of bird genomes – a transposon's-eye view. 1389:164–185. DOI: https://doi.org/10.1111/nyas.13295.

Künstner A, Wolf JBW, Backström N, Whitney O, Balakrishnan CN, Day L, Edwards SV, Janes DE, Schlinger BA, Wilson RK, Jarvis ED, Warren WC, Ellegren H (2010) Comparative genomics based on massive parallel transcriptome sequencing reveals patterns of substitution and selection across 10 bird species. Mol Ecol 19(Suppl 1):266–276. https://doi.org/10.1111/j.1365-294X.2009.04487.x

Lesk AM (2012) Introduction to genomics. Oxford University Press

Manthey JD, Robbins MB, Moyle RG (2016) A genomic investigation of the putative contact zone between divergent Brown Creeper (*Certhia americana*) lineages: chromosomal patterns of genetic differentiation. Genome 59:115–125. https://doi.org/10.1139/gen-2015-0093

Marshall Graves JA (2015) Evolution of vertebrate sex chromosomes and dosage compensation. Nat Rev Genet 17:33–46. https://doi.org/10.1038/nrg.2015.2

Matsubara K, Tarui H, Toriba M, Yamada K, Nishida-Umehara C, Agata K, Matsuda Y (2006) Evidence for different origin of sex chromosomes in snakes, birds, and mammals and step-wise differentiation of snake sex chromosomes. Proc Natl Acad Sci U S A 103:18190–18195. https://doi.org/10.1073/pnas.0605274103

McCormack JE, Maley JM, Hird SM, Derryberry EP, Graves GR, Brumfield RT (2012) Next-generation sequencing reveals phylogeographic structure and a species tree for recent bird divergences. Mol Phylogenet Evol 62:397–406. https://doi.org/10.1016/j.ympev.2011.10.012

Merilä J, Hendry AP (2014) Climate change, adaptation, and phenotypic plasticity: the problem and the evidence. Evol Appl 7(1):1–14

Patwardhan A, Roy A, Ray S (2014) Molecular markers in phylogenetic studies – a review. J Phylogenet Evol Biol 2. https://doi.org/10.4172/2329-9002.1000131

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE, Orlando L (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One 7(5):e37135

Pigliucci M (2001) Phenotypic plasticity: beyond nature and nurture. JHU Press, Baltimore

Scanes CG (2015) Sturkie's avian physiology. Elsevier

Schlötterer C (2004) The evolution of molecular markers – just a matter of fashion? Nat Rev Genet 5:63–69

Semagn K, Bjornstad A, Ndjiondjop MN (2006) An overview of molecular marker methods for plants. Afr J Biotechnol 5:2540–2568. https://doi.org/10.5897/AJB2006.000-5110

Silva CNS, McFarlane SE, Hagen IJ, Rönnegård L, Billing AM, Kvalnes T, Kemppainen P, Rønning B, Ringsby TH, Sæther B-E, Qvarnström A, Ellegren H, Jensen H, Husby A (2017) Insights into the genetic architecture of morphological traits in two passerine bird species. Heredity 119:197–205. https://doi.org/10.1038/hdy.2017.29

Toraño EG, Garciá MG, Fernández-Morera JL, Ninõ-Garciá P, Fernández AF (2016) The impact of external factors on the epigenome: in utero and over lifetime. Biomed Res Int. https://doi.org/10.1155/2016/2568635

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63. https://doi.org/10.1038/nrg2484

Weissensteiner MH, Pang AWC, Bunikis I, Höijer I, Vinnere-Petterson O, Suh A, Wolf JBW (2017) Combination of short-read, long-read, and optical mapping assemblies reveals large-scale tandem repeat arrays with population genetic implications. Genome Res 27:697–708. https://doi.org/10.1101/gr.215095.116

West-Eberhard MJ (2003) Developmental plasticity and evolution. Oxford University Press, Oxford