UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1823

Gene regulatory evolution in flycatchers: statistical approaches for the analysis of allele-specific expression

MI WANG

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2019

ISSN 1651-6214
ISBN 978-91-513-0686-5
urn:nbn:se:uu:diva-384493

Dissertation presented at Uppsala University to be publicly examined in Lindahlsalen, Norbyvägen 14, Uppsala, Wednesday, 4 September 2019 at 13:00 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Associate professor Christopher Wheat (Department of Zoology, Stockholm University).

Abstract
Wang, M. 2019. Gene regulatory evolution in flycatchers: statistical approaches for the analysis of allele-specific expression. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 1823. 46 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-0686-5.

Understanding the molecular mechanisms underlying evolutionary changes in gene expression is a major research topic in biology. While a powerful approach to study this is the analysis of allele-specific expression (ASE), most of previously published methods can only be applied to lab organisms. In this thesis, to enable the analysis of ASE in natural organisms, I developed two methods for ASE detection. The first one was Bayesian negative binomial approach, and the second one was Read-backed Phasing-based ASE approach. Both methods performed well in simulations and comparisons. By applying those methods, I found that ASE was prevalent in natural flycatcher species. Combining the analyses of differential gene expression and ASE, I found a widespread cis-trans compensation and a critical role of tissue-specific regulatory mechanism during gene expression evolution. Moreover, for cis-regulatory sequences, there was a larger proportion of slightly deleterious mutations and weaker signatures of positive selection for genes with ASE than genes without ASE. For coding sequence, no such difference was observed. These results indicated that the evolution of gene expression and coding sequences could be uncoupled and occurred independently.

Keywords: ASE, gene expression evolution, bayesian, RPASE, flycatcher.

Mi Wang, Department of Ecology and Genetics, Evolutionary Biology, Norbyvägen 18D, Uppsala University, SE-75236 Uppsala, Sweden.

ISSN 1651-6214
ISBN 978-91-513-0686-5
urn:nbn:se:uu:diva-384493 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-384493)

愿我们
一生努力
一生被爱
想要的都拥有
得不到的都释怀

List of Papers

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

I **Wang, M.,** Uebbing, S., Ellegren, H. (2017) Bayesian inference of allele-specific gene expression indicates abundant *cis*-regulatory variation in natural flycatcher populations. *Genome Biology and Evolution*, 9(5):1266–1279

II **Wang, M.,** Uebbing, S., Pawitan, Y., Scofield, D. G. (2018) RPASE: Individual-based allele-specific expression detection without *prior* knowledge of haplotype phase. *Molecular Ecology Resources*, 18(6):1247–1262

III Mugal, C. F., **Wang, M.,** Backström, N., Wheatcroft, D., Ålund, M., McFarlane, E. S., Dutoit, L., Qvarnström, A., Ellegren, H. Gene regulatory evolution in natural flycatcher populations is highly tissue-specific and shows distinctive patterns in the testis. *Manuscript*

IV **Wang, M,** Mugal, C. F., Craig, R. J., Dutoit, L., Bolivar, P., Ellegren, H. *cis*-regulatory variation and allele-specific expression in the collared flycatcher (*Ficedula albicollis*) genome. *Manuscript*

Additional papers

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

Craig, R. J., Suh, A., **Wang, M.,** Ellegren, H. (2018). Natural selection beyond genes: Identification and analyses of evolutionarily conserved elements in the genome of the collared flycatcher (*Ficedula albicollis*). *Molecular Ecology*, 27(2): 476–492.

Bolívar, P., Mugal, C. F., Rossi, M., Nater, A., **Wang, M.,** Dutoit, L., Ellegren, H. (2018). Biased inference of selection due to GC-biased gene conversion and the rate of protein evolution in flycatchers when accounting for it. *Molecular Biology and Evolution*, 35(10): 2475–2486.

Dutoit, L., Mugal, C. F., Bolivar, P., **Wang, M.,** Nadachowska-Brzyska K., Smeds, L., Yazdi, H. P., Gustafsson, L., Ellegren, H. (2018) Sex-biased gene expression, sexual antagonism and levels of genetic diversity in the collared flycatcher (*Ficedula albicolli*s) genome. *Molecular Ecology*, 27(18): 3572-3581

Contents

Introduction

Gene expression evolution

Around 160 years ago, Darwin revealed that natural selection can underlie the evolutionary change of species over time. Since then, what is actually being selected at the molecular level is of major importance for many areas of biology. Around 40 years ago, King and Wilson (King & Wilson 1975) proposed that evolution can occur at two levels: gene sequence and gene expression. Translating these two levels into the DNA sequence, selection can act upon coding sequences that encode the protein, and regulatory sequences that regulate expression level of the protein.

While selection on protein coding sequence has been a subject of study for decades, the evolutionary significance of gene regulatory evolution was not widely recognized until recently (Prud'homme *et al.* 2007). Coding sequences were thought to be the main source for adaptation. Technically, it was also more straightforward to define a neutral null hypothesis when inferring selection on coding sequence. In contrast, it has been difficult to identify and annotate regulatory sequences. To circumvent this difficulty, the role of regulatory sequence has mostly been studied through the analysis of gene expression. However, gene expression is temporally and spatially dynamic, and changes in gene expression involve variations in DNA, RNA, proteins, and the environment. Identifying expression and regulatory divergence that are responsible for adaptation is therefore much more complicated than identifying that of coding sequence divergence.

Nowadays, it has become increasingly clear that expression divergence often plays a key part in adaptation and speciation (Mack *et al.* 2019; Romero *et al.* 2012; Wittkopp & Kalay 2012; Wray 2007). This has been supported by rapidly accumulating empirical studies which showed that changes in gene expression are responsible for changes relating to morphology, behavior, and physiology. For example, in *Drosophila*, the pigmentation, which plays an important role in crypsis, thermoregulation and mate choice, is associated with gene expression regulation (Gompel & Carroll 2003; Gompel *et al.* 2005; Hollocher *et al.* 2000; Kopp *et al.* 2000; Kopp & True 2002). In stickleback

fish, the reduction of pelvic is suggested to be caused by changes in gene expression (Cresko *et al.* 2004; Marques *et al.* 2016; Morris *et al.* 2014; Shapiro *et al.* 2004). In humans, lactose persistence, breast cancer, myocardial infarction, as well as many other disease risks are also correlated with gene expression variations (Bersaglieri *et al.* 2004; Glassberg *et al.* 2019; Olds & Sibley 2003; Swallow 2003; Tishkoff *et al.* 2007).

These studies have also promoted the view that some phenotypes are easier to achieve via expression changes than via coding sequence changes. Changing the encoded protein generally affects all cells at all times and at all places where the protein is active, and as a consequence, has a great pleiotropic effect. Changes in protein coding sequence are suggested to evolve under a strong selective constraint. As opposite, changing the level of gene expression has a spatially and temporally circumscribed effect, depending on the cell type, the developmental stage, the tissue, and the environment, which largely minimizes the functional trade-offs and reduces the pleiotropic effect (Fraser 2011; Prud'homme *et al.* 2007; Romero *et al.* 2012; Wittkopp & Kalay 2012; Wray 2007).

Gene regulation appears to evolve under tissue-specific pressures, depending on tissue-specific biological process, tissue development, and tissue-specific function (Romero *et al.* 2012). The rates of gene expression evolution among tissues show substantial differences: brain is usually under the most selective constraint, whereas testis usually shows an excess of divergence between species (Blekhman *et al.* 2008; Brawand *et al.* 2011; Consortium 2017). In addition, there is an enrichment of expression divergence on sex chromosomes compared to autosomes (Blekhman *et al.* 2008; Brawand *et al.* 2011; Li *et al.* 2017).

The evolutionary significance of *cis*- and *trans*-regulatory changes

Regulatory changes can be broken up into *cis*- and *trans*-regulatory changes. *Cis*-regulatory changes refer to sequence variation in *cis*-regulatory elements. *Cis*-regulatory elements, such as insulator, promoter and enhancer, are stretches of non-coding DNA sequences that located linearly or spatially close to the gene involved. *Trans*-regulatory changes refer to changes in *trans*-regulatory factors. *Trans*-regulatory factors, such as transcription factors and non-coding RNAs, are proteins or RNAs that are encoded by coding DNA sequences. *Trans*-regulatory factors bind to *cis*-regulatory elements, and together, they form the genetic basis of gene expression regulation. Changes in

either of them can influence the binding affinity, and as a consequence, change the level of gene expression.

While *cis*- and *trans*-regulations work together, they have different roles on gene regulation. A single *trans*-regulatory factor binds to multiple *cis*-regulatory elements from multiple genes, and thus *trans*-regulatory variation affects expression level of multiple genes. In contrast, a single gene may contain many *cis*-regulatory elements with each having independent influence. *Cis*-regulatory variation has thus only limited influence on one particular gene. As a result, *trans*-regulatory variation is suggested to have a greater mutational target and undergo stronger selective constraint, whereas *cis*-regulatory variation has less deleterious pleiotropic side effect, and consequently, plays an important role in adaptation by bypassing the selective constraint (Mack & Nachman 2017; Prud'homme *et al.* 2007; Wray 2007).

Cis-trans compensation and Dobzhansky-Muller hybrid incompatibility

Previous studies show that *cis*- and *trans*-regulatory variations often influence the same gene (Bell *et al.* 2013; Chen *et al.* 2015; Coolon *et al.* 2014; Davidson & Balakrishnan 2016; Emerson *et al.* 2010; Goncalves *et al.* 2012; Graze *et al.* 2009; Guerrero *et al.* 2016; Landry *et al.* 2005; Mack *et al.* 2016; McManus *et al.* 2010; Schaefke *et al.* 2013; Tirosh *et al.* 2009; Wittkopp *et al.* 2004; Zhuang & Adams 2007). When such influences have the same direction, *cis*- and *trans*-regulatory variations reinforce each other, leading to a change in expression level more extreme than individual regulatory variation. When such influences have the opposite direction, *cis*- and *trans*-regulatory variations compensate each other, leading to a less extreme change than individual regulatory variation. Interestingly, the opposite direction has been found to play a predominant role (Brawand *et al.* 2011; Goncalves *et al.* 2012; Metzger *et al.* 2016; Romero *et al.* 2012; Takahasi *et al.* 2011; Verta *et al.* 2016). This reveals that the interactions of *cis*- and *trans*- regulatory variations often compensate for one another and result in no or limited change on overall expression level (Gilad *et al.* 2006; Lemos *et al.* 2008). Although individual *cis*- or *trans*-regulatory variation destabilizes the gene expression, the compensated pair of *cis*- and *trans*-regulatory variations re-stabilizes it back. Therefore, the conservation of gene expression can be maintained at the same time as variations in regulatory sequences accumulate (Signor & Nuzhdin 2018).

The compensated pairs of *cis*- and *trans*-regulatory variations co-adapt within species during evolutionary process (Mack & Nachman 2017), but such co-

adaption gets disrupted during hybridizing, which consequently induces gene mis-regulation in the inter-species hybrid. As being explained by the Dobzhansky-Muller model (Haerty & Singh 2006; Signor & Nuzhdin 2018), the novel incompatible interaction between divergent *cis*- and *trans*-regulatory sequences can cause hybrid incompatibility and provide molecular basis for the evolution of intrinsic post-zygotic reproductive isolation (Landry *et al.* 2005; Mack *et al.* 2016; Tulchinsky *et al.* 2014). In addition, gene mis-regulation is commonly observed in male-based genes of sterile hybrid, implying that it might underlie hybrid male sterility (Gomes & Civetta 2015; Good *et al.* 2010; Mack *et al.* 2016; Michalak & Noor 2003; Ranz *et al.* 2004; Sundararajan & Civetta 2011; Turner *et al.* 2014).

Identification of genes influenced by *cis*- and *trans*-regulatory variations

Since *cis*-regulation only influences the gene that locates on the same chromosome, it has an allelic effect for individual that is heterozygous for the *cis*-regulatory variation. Such allelic effect gives rise to allele-specific gene expression (ASE), which is defined as maternal and paternal alleles of a gene being expressed at a different level (in diploid organisms). Since both alleles are exposed to the same *trans*-regulatory factors and experience the same environmental conditions, ASE can in turn serve as a powerful tool to identify genes with *cis*-regulatory effect. It must be noted that, ASE analyses are not able to directly identify *cis*-regulatory variation itself. In addition, ASE analyses require heterozygous mutation in protein coding sequence to be able to distinguish maternal and paternal alleles.

While *cis*-regulatory effect can be studied within species, the investigation of *trans*- regulatory effect requires homozygous parents and their heterozygous F1 hybrid. For genes that show differential gene expression between parental species, the presence of ASE in hybrid indicates a *cis*-regulatory divergence whereas the absence of ASE indicates a *trans*-regulatory divergence.

Studying natural populations

Crosses of inbred strains/species with substantially divergent genomes from model organisms in lab settings are mostly used to study the role of *cis*- and *trans*-regulation (Coolon *et al.* 2014; Emerson *et al.* 2010; Glaser-Schmitt & Parsch 2018; Goncalves *et al.* 2012; Mack *et al.* 2016; McManus *et al.* 2010;

Schaefke *et al.* 2013; Tirosh *et al.* 2009; Wittkopp *et al.* 2008). Such crosses have small within- and large between-species divergence, which largely ensures the homozygosity of parental species and heterozygosity of their F1 hybrid.

However, non-model natural populations have not been fully investigated (Tung *et al.* 2015; Wang *et al.* 2017). Apart from lacking convenient features in lab organisms described above, studying natural populations is further challenged by having no complete knowledge of the phase, no replicates that have precisely matched developmental stages and environmental conditions, and frequent polymorphisms. Those limitations have led to a bias in favor of studying lab organisms, and it is still unclear if what is known can directly be transferred to the wild. A broad view of regulatory evolution remains to be formulated, which makes the studying of natural non-model organisms of critical importance.

The flycatcher system

The collared flycatcher (*Ficedula albicollis*) is a trans-Saharan migrant songbird. It breeds in deciduous and mixed coniferous forests in Europe and has an estimated effective population size (N_e) of approximately 200,000 (Nadachowska-Brzyska *et al.* 2013). About less than one million years ago, the collared flycatcher started diverging from its sister species, the pied flycatcher (*Ficedula hypoleuca*), which is likely due to occupying different habitat during the Pleistocene glaciations (SÆTRE *et al.* 2001).

These two species came into contact in a natural hybrid zone on the Swedish islands of Öland in the Baltic Sea since roughly 60 years ago (Qvarnström *et al.* 2010). Their naturally-occurring F1 hybrids show physiological dysfunction in several ways. First, they have low paring success. Hybrid males produce intermediate songs compared to their parental species, and such transgressive signals are found to be sexually unattractive by females of both parental species (Price & Wedell 2008; Svedin *et al.* 2008; Verzijden *et al.* 2012). Second, they have an elevated metabolic rate, which results in an expensive self-maintenance (McFarlane *et al.* 2016). Third, hybrid males have severely reduced fertility due to fewer and malformed sperm compared to their parental species.

The flycatcher species have been used extensively as a natural non-model organism to study the processes of speciation (Burri *et al.* 2015; Ellegren *et al.* 2012), molecular evolution (Bolívar *et al.* 2015; Bolívar *et al.* 2018), recombination rate evolution (Kawakami *et al.* 2017; Kawakami *et al.* 2014), gene

expression evolution (Uebbing *et al.* 2016) and recently ASE (Wang *et al.* 2017; Wang *et al.* 2018). With such great amount of resources and data, the flycatcher species are very good systems to be used to investigate gene regulatory evolution in the wild.

Conserved non-coding elements

With advances in sequencing technology, although the analysis of ASE has become a key tool when studying regulatory evolution, it remains rather difficult to identify the regulatory sequence and to annotate the causal variation that is responsible for ASE. Previous studies showed that, while a phenotypic influence is found to be *cis*-regulatory, very often, the underlying genetic bases are still unclear (Clark *et al.* 2006; Drapeau *et al.* 2006; Marcellini & Simpson 2006). Unlike coding sequence, which is classified as synonymous and nonsynonymous substitution to reflect sequence variation with and without a phenotypic influence at the protein level, there is a lack of method to distinguish *cis*-regulatory variation from noncoding DNA sequences alone, for the vast majority of organisms (Rockman & Wray 2002; Wittkopp & Kalay 2012).

To circumvent this difficulty, identifying *cis*-regulatory elements can be a starting point. Analogous to the high conservation of coding regions relative to surrounding sequences, functional non-coding sequence should also be conserved across evolution. The search for *cis*-regulatory elements can thus be guided by patterns of sequence conservation (Harmston *et al.* 2013).

A majority of tested conserved non-coding DNA sequences, referred as conserved non-coding elements (CNEs), is found to be associated with changes in expression level of neighbouring genes (De La Calle-Mustienes *et al.* 2005; Harmston *et al.* 2013; Sanges *et al.* 2013; Shen *et al.* 2012; Visel *et al.* 2008; Visel *et al.* 2009), and contribute to expression divergence relating to lineage-specific traits (Polychronopoulos *et al.* 2017). CNEs have therefore been suggested to act typically as *cis*-regulatory elements (Pennacchio *et al.* 2006; Sanges *et al.* 2013; Visel *et al.* 2008). Following this, sequence variation at CNEs can be used as a proxy for *cis*-regulatory variation.

Analyses of ASE

The ASE analyses, especially in natural organisms, are challenged by, e.g., mapping bias, overdispersion estimation, haplotype construction and read

counts dependency. Mapping bias is introduced when reads from both maternal and paternal alleles are aligned with a single genome reference since reads from the allele that is least diverged from the reference genome had a larger mapping opportunity, relative to the other allele (Stevenson *et al.* 2013). While such inhomogeneous mapping can be remedied by making two reference genomes with each containing only maternal or paternal alleles (Quinn *et al.* 2014; Rozowsky *et al.* 2011; Van de Geijn *et al.* 2015), by removing biased SNPs from simulation (Degner *et al.* 2009), or by introducing polymorphism-aware aligners (Wu & Nacu 2010), mapping bias cannot be eliminated completely (Castel *et al.* 2015).

Overdispersion refers to the presence of greater variability in the data comparing to the theoretical expectation in a given statistical distribution. It is estimated by using multiple replicates. In RNA-seq analyses in general, overdispersion is introduced by both biological variation, such as genetically/environmentally non-identical replicates, and technical variations, such as non-identical runs of libraries, sequencing error, and mapping bias. Ignoring overdispersion leads to a high false discovery rate, and thus, it is of particular importance to account for overdispersion when analysing RNA-seq. Extra-binomial and extra-Poisson distributions have been used extensively (Edsgärd *et al.* 2016; Leon-Novelo *et al.* 2014; Mayba *et al.* 2014; Wang *et al.* 2017) to study differential gene expression in many species, including plants (Steige *et al.* 2015), *Drosophila* (Graze *et al.* 2012), birds (Davidson & Balakrishnan 2016) and humans (Glassberg *et al.* 2019; McCoy *et al.* 2017).

Estimating overdispersion in ASE detection when studying natural organisms is not straightforward, due to the difficulty of obtaining natural replicates. To circumvent this limitation, multiple heterozygous SNPs within a haplotype can be used to estimate overdispersion (Rozowsky *et al.* 2011; Wang *et al.* 2017). This approach, however, relies on a known phasing for heterozygous SNPs. With the exception of data from humans, most study systems do not have sufficient phased data available.

Haplotype phase can be constructed by different methods. First, it can be inferred from linkage disequilibrium using preferably hundreds of individual samples (Browning & Browning 2007; Menelaou & Marchini 2013). This method however cannot reliably infer rare haplotypes as well as individual haplotypes that may arise from recent recombination event (He *et al.* 2010; Snyder *et al.* 2015). While pseudo-phasing (Mayba *et al.* 2014), which artificially builds a major allele by grouping larger read counts and a minor allele by grouping smaller read counts, improves phasing at an individual level, it can produce spurious ASE calls when genes have discordant read counts (Wang *et al.* 2017). Read-backed phasing deduces physical haplotype conti-

guity. It uses heterozygous SNPs to join reads that share alleles into haplotypes, resulting in a high accuracy at individual phasing (McKenna *et al.* 2010).

Even with perfectly phased individual haplotype, a further difficulty can arise from the fact that read counts from multiple neighbouring SNPs may be dependent since a single RNA-seq read can contribute to read counts at multiple SNP positions. Although this is a natural result of linear transcription, such dependency violates a common assumption of many statistical test that treats read counts as independent observations. Taking together, overcoming these difficulties is of critical importance, which enables the analysis of ASE and allows an investigation of regulatory evolution in the wild.

Methods

All studies in this thesis used RNA-seq data from natural bird populations. The field and lab work had been performed by collaborators. Paper I and Paper II addressed challenges in ASE detection for natural organisms. Paper III and Paper IV used the developed method in Paper II and studied molecular mechanisms and genetic determinants of regulatory evolution.

Sampling

In Paper I, collared and pied flycatchers were sampled from Öland and Uppsala, respectively (Uebbing *et al.* 2016). Five unrelated male individuals for each species were dissected immediately in the field. Brain, kidney, liver, lung, muscle, skin, and gonad were then isolated and stored in RNA-later at -80ºC. Paper II used the above samples but included 4 more unrelated female collared flycatchers and 5 more unrelated female pied flycatchers. All female individuals were processed with the same method.

In Paper IV, 5 male collared flycatchers were sampled on Öland during the breeding season 2014. However, this time, individuals were held in aviaries at the field nearby the sampling sites for at least two weeks prior to tissue dissection. Tissue samples were then placed in RNA-later at -80ºC until RNA extraction. Five tissues of brain, heart, kidney, liver and left testis were studied. In Paper III, beside above samples, we also included 5 pied flycatchers and 3 F1 hybrids between collared and pied flycatchers in the same place. All individuals were processed with an identical method. Species and hybrid identity were identified by plumage score first (Qvarnström *et al.* 2010), and then confirmed by single nucleotide fixed differences from the sequencing data.

RNA extraction, library preparation and sequencing

Tissues were homogenized, and total RNA was extracted according to the manufacturer's instructions. Illumina paired-end libraries for RNA-seq were prepared. For Paper I and II, RNA-seq was generated by using an Illumina

Genome Analyzer IIx for 100 sequencing cycles. For Paper III and IV, sequencing was performed on an Illumina HiSeq instrument, and a sequencing library for the phage PhiX was included as 1% spike-in in the sequencing run.

Read mapping and SNP calling

We used the collared flycatcher assembly FicAlb1.5 (GenBank Accession: GCA_000247815.2) as the reference genome, and the gene annotation was obtained from *Ensembl* (http://www.ensembl.org) release 73. To reduce the mapping bias towards the reference genome, all positions in the collared flycatcher genome that showed fixed differences between collared and pied flycatchers were masked, using SAMtools v.1.3 (Li *et al.* 2009).

For Paper I, TopHat v. 2.0.5 (Kim *et al.* 2013) with default settings was used to map RNA-seq reads to the SNP-masked genome sequence, and for the rest of studies, STAR v.2.5.1b (Dobin *et al.* 2013) was used. Only uniquely mapped reads were used for further analyses.

For Paper I, genome re-sequencing data from the same individuals were obtained (Ellegren *et al.* 2012) and used in SNP calling by GATK v. 2.2 with standard options (DePristo *et al.* 2011). SNPs that passed GATK standard quality criteria were extracted and used in the phasing, by Beagle v. 3.0.4 (Browning & Browning 2007).

For Paper II, III, and IV, SNPs were called individually using GATK 3.5.0, following the best practice with recommended parameter settings (Van der Auwera *et al.* 2013). Differently, for Paper II, we combined reads from genome re-sequencing and RNA-seq in the SNP calling, and for Paper III and IV, we used RNA-seq data only.

The Bayesian negative binomial (NB) approach

The Bayesian NB approach was among one of the first methods that allowed the estimation of overdispersion when testing for ASE at an individual level. It calculated overdispersion using a Bayesian solution and tested for ASE using generalized linear mixed mode framework.

Let Y_{ijp} be the read counts from allele p ($p = 1, 2$) at SNP j ($j = 1, \ldots, J$) in *trans*cript i. Assuming a NB distribution for Y_{ijp}, there is

$$Y_{ijp} \sim NB(\mu_{ijp}, \varphi_i),$$

where φ_i is the overdispersion parameter. In line with extra Poisson distribution, it can also be denoted as

$$Y_{ijp} \sim NB(\mu_{ijp}, \sigma^2_{ijp}),$$

where $\sigma^2_{ijp} = \mu_{ijp} + \varphi_i\mu^2_{ijp}$. The negative binomial distribution has the probability mass function

$$f(y_{ijp}|\mu_{ijp}, \varphi_i) = P\left(Y_{ijp} = (y_{ijp}|\mu_{ijp}, \varphi_i)\right)$$
$$= \frac{\Gamma(y_{ijp} + \varphi_i^{-1})}{\Gamma(\varphi_i^{-1})\Gamma(y_{ijp} + 1)}\left(\frac{1}{1 + \mu_{ijp}\varphi_i}\right)^{\varphi_i^{-1}}\left(\frac{\mu_{ijp}}{\varphi_i^{-1} + \mu_{ijp}}\right)^{y_{ijp}}.$$

(Robinson & Smyth 2007), φ_i is calculated by maximizing a weighted likelihood $WL(\varphi_i)$, where

$$WL(\varphi_i) = APL(\varphi_i) + \gamma_i\, APL_C(\varphi_i),$$

$APL(\varphi_i)$ is the individual likelihood, $APL_C(\varphi_i)$ is the averaged likelihood, and γ_i is the weight given to $APL_C(\varphi_i)$.

This weighted likelihood can be interpreted as an approximate empirical Bayesian solution with the $APL_C(\varphi_i)$ as the prior distribution and $WL(\varphi_i)$ as the posterior distribution for φ_i. While γ_i is fixed for all genes in Robinson and Smyth 2007, it was flexible in the Bayesian NB approach for each transcript, depending on *J*. It was an important and necessary extension for the ASE detection since different transcripts contained different number of SNPs. We defined $\gamma_i = \frac{3}{df_i}$ and $df_i = 2\times J\text{-}2$, where df_i was residual degrees of freedom when estimating φ_i. Such data-driven γ_i allowed the weight to vary flexibly among transcripts based on the potential statistical reliability of individual estimation. For example, when *J* is comparatively large (e.g., $J = 4$), the individual likelihood $APL(\varphi_i)$ is prioritized over the common likelihood $APL_C(\varphi_i)$ because γ_i is less than one ($\gamma_i = 0.5$). On the other hand, if $J = 2$, then $APL_C(\varphi_i)$ is prioritized over $APL(\varphi_i)$ due to $\gamma_i = 1.5$. After fitting the NB model, null hypothesis that two alleles have equal expression can then be tested.

RPASE approach

RPASE consisted of two steps: (i) GATK read-backed phasing and (ii) extra-binomial exact test for ASE. Read-backed phasing deduces a physical phase of SNPs. While it enables the extension of ASE detection from individual SNPs to multiple SNPs within a haplotype, it introduces dependency between read counts since haplotype phase is produced by joining reads that shared

SNPs. The outcome of phased regions was referred as phased blocks, forming a basic unit of the ASE test.

Comparing to binomial test, extra-binomial test takes a greater variability in the data, i.e. overdispersion, into consideration, where overdispersion is calculated by having a random probability of success. Let a random variable y_i has

$$y_i \sim \text{Binomial}\ (n_i, p_i),$$

where $n_i = y_{i1sum} + y_{i2sum}$, y_{i1sum} and y_{i2sum} are two allelic summed counts from the *i*th phased block, and p_i is the probability of success. Allowing a random effect in p_i, we have

$$logitp_i = b_i, \tag{1}$$

where b_i's were iid $N(0, \sigma^2)$. Conditional on b_i, y_{i*} has a marginal probability

$$\begin{aligned} f(k_i; n_i, p_i, \sigma^2) &= \Pr(y_{i*} = k_i) \\ &= E\{\Pr(y_{i*} = k_i | b_i)\} \\ &= \int \binom{n_i}{k_i} {p_i}^{k_i} (1 - p_i)^{n_i - k_i} \phi\left(\frac{b_i}{\sigma}\right) d\left(\frac{b_i}{\sigma}\right). \end{aligned} \tag{2}$$

It describes the probability of observing k_i out of n_i, where $\phi(*)$ is the standard normal density. Overdispersion is calculated by 32-point Gaussian quadrature technique (Pawitan 2001).

As a next step, we addressed the dependency of read counts. Let $y_{i[s]*}$ represented the *s*th minimum counts in a phased block that has *r* SNPs $(1 \le s \le r)$, we then have

$$\begin{cases} t_{i[s]1} = y_{i[s]1}, t_{i[s]2} = y_{i[s]2}, & s = 1 \\ t_{i[s]1} = y_{i[s]1} - y_{i[s-1]1}, t_{i[s]2} = y_{i[s]2} - y_{i[s-1]2}, & s > 1 \end{cases}, \tag{3}$$

with $t_{i[s]*}$ referring to the stratum counts for the *s*th stratum. Under the null hypothesis of no ASE, i.e., allelic frequency (k_i/n_i) of 0.5, the probability mass function can be constructed as

$$f\left(\frac{k_i}{n_i}\right) = \sum_{\Omega} \prod_s f\left(k_{i[s]}; n_{i[s]}, p_{i[s]}, \sigma^2\right), \tag{4}$$

where

$$\Omega = \{\Omega_1, \Omega_2, \ldots \Omega_* \ldots\}, \tag{5}$$

$$\Omega_* = \{k_{i[1]}, k_{i[2]}, \dots, k_{i[s]}, \dots, k_{i[r]}\}, \quad (6)$$

$$\{k_{i[1]}, k_{i[2]}, \dots, k_{i[s]}, \dots, k_{i[r]}\} \in \left\{\frac{\sum_s (r-s+1)k_{i[s]}}{\sum_s (r-s+1)n_{i[s]}} = \frac{k_i}{n_i}\right\}, \quad (7)$$

$p_{i[s]} = 0.5$, $n_{i[s]} = t_{i[s]1} + t_{i2[s]2}$. $f(k_{i[s]};\ n_{i[s]}, p_{i[s]}, \sigma^2)$ denotes the probability of observing $k_{i[s]}$ in the *s*th stratum from the normal binomial distribution as defined in Equation 2.

Finally, we calculated the exact *p*-value. Using the probability mass function defined in Equation 4, the exact *p*-value for the *i*th phased block is

$$p_i = \sum_{\frac{k_i}{n_i} \geq \frac{k_i'}{n_i'}} f(k_i/n_i) \quad (8)$$

or

$$p_i = 1 - \sum_{\frac{k_i}{n_i} \geq \frac{k_i'}{n_i'}} f(k_i/n_i), \quad (9)$$

depending on whether the observed allelic frequency $\frac{k_i'}{n_i'}$ is larger or smaller than 0.5.

Above algorithm is implemented in the R package RPASE, which can be download and installed from Github (https://github.com/wangmi811/RPASE).

Simulations and comparison with other method

To evaluate the performance of the developed ASE methods, we simulated multiple datasets under a variety of scenarios with 8% of genes to be true positives. For genes without ASE, we simulated read counts with a mean allelic frequency of 0.5, and for genes with ASE, we simulated different allelic frequencies. We simulated 1000 genes in each scenario and repeated each simulation 100 times. Average true positive rate and false discovery rate were calculated accordingly.

RNASeqReadSimulator (http://alumni.cs.ucr.edu/~liw/rnaseqreadsimulator.html) was used to simulate paired-end RNA-seq reads for Paper II. We varied

SNP density, allelic frequency, overdispersion, coverage and sequencing error to investigate the performance of RPASE.

As independent validations, the developed ASE methods were compared to MBASED (Mayba *et al.* 2014), which is to our knowledge the only published method that allows for individual phasing before detecting ASE. MBASED has two mode: phased mode and unphased mode. With phased mode, a known phase is required. With unphased mode, it performs pseudo-phasing. As suggested by Mayba et al. (2014), major allelic frequency > 0.7 and Benjamini Hochberg-adjusted p-value < 0.05 are–used as significant thresholds when identifying ASE. In Paper I, MBASED was run in phased mode and compared with Bayesian NB approach, and in Paper II, both phased mode and unphased mode were compared with RPASE.

Differential gene expression analyses, tissue-specificity estimation and Gene Ontology analyses

We investigated differential gene expression (DE) by using the DEseq2 package (Love *et al.* 2014). Following the recommended pipeline, the output from HTseq v0.6.1 with default settings (Anders *et al.* 2015) were fed to DEseq2. For each tissue separately, DE analyses were based on pair-wise contrast, i.e. collared vs. pied flycatchers, collared flycatchers vs. F1 hybrids, and pied flycatchers vs. F1 hybrids.

Tissue-specificity was estimated by using *trans*cripts per million (TPM), where TMP was calculated using RSEM 1.2.29 (Li & Dewey 2011) after STAR mapping. For each tissue in each species separately, we calculated tissue-specificity index τ_i for each gene i following (Yanai *et al.* 2005):

$$\tau_i = \frac{\sum_{t=1}^{k}\left(1 - \frac{log_2(E_{it})}{log_2(E_{imax})}\right)}{k-1},$$

where E_{it} is the TPM value for gene i in tissue t, and E_{imax} is the maximum TPM value across the total number of k tissues for the ith gene.

Flycatcher Gene Ontology (GO) annotations were retrieved from Ensembl 73. We tested the enrichment of GO terms by using topGO package in R. Benjamini Hochberg-adjusted p-value < 0.05 was used to determinant whether test results were significant or not.

Identification of CNEs

We identified conserved element (CE) using phastCons (Siepel *et al.* 2005) with whole-genome alignments, phylogenetic model and parameters from (Craig *et al.* 2018). Conserved non-coding elements (CNEs) were later identified as CEs with no overlap with protein-coding sequence, RNA genes, or pseudogenes. We then assigned CNEs to the nearest gene based on the physical distance between the CNE and the transcription start site (TSS) of a gene.

Estimation of distribution of fitness effects (DFE), ω_a, α, and ω

We grouped genes with ASE into an ASE group and genes without ASE into a control group. Then, we tested if the strength of selection in *cis*-regulatory elements and coding sequences differed between the two groups. DFE-alpha v2.16 (Eyre-Walker & Keightley 2009) was used. It fits a gamma distribution by comparing the folded site frequency spectrum (SFS) for sites under selection to those that are putatively neutral, and the strength of purifying selection is estimated based on the shape and mean of the gamma distribution. For *cis*-regulatory elements, we considered CNEs located 2 kb upstream of TSS (CNE_2kb) as selected sites and non-conserved sequence in 2 kb upstream of the TSS (nonCNE_2kb) as the neutral reference. For coding sequences, we considered 0-fold degenerate sites as selected sites and 4-fold degenerate sites as the neutral reference.

Baseml program in PAML 4.9e (Yang 2007) and sequence alignments between collared flycatcher, zebra finch and chicken (outgroup) were used to estimate the divergence for the collared flycatcher lineage. We calculated ω_a as the rate of adaptive substitution relative to neutral divergence, α as the proportion of adaptive substitution, and ω as the ratio of divergence for *cis*-regulatory elements and coding sequence separately.

95% confidence intervals for the DFE, ω_a and α were generated by bootstrapping genes for 200 times with replacement. After calculating the standard error of the bootstrapped distribution, confidence intervals were defined using the 2.5th and 97.5th percentiles of the Student's t-distribution. We then conducted randomization test following (Eyre-Walker & Keightley 2009) to test differences in the DFE, ω_a and α between the ASE and control groups.

Regression analysis

To investigate what features determinate whether a gene showed ASE or not, we gathered data that have previously been suggested to affect *cis*-regulatory variation and performed logistic regression analyses. Presence or absence of ASE (1 or 0) was used as the response variable. Recombination rate, density of functional sites, local mutation rate, the number of protein-protein interactions (PPI), tissue specificity (τ), gene expression level, and gene length were used as explanatory variables. Among them, recombination rate, the density of functional sites, and mutation rate were selected to represent variables relating to genomic background. The others were used to represent gene-specific functions and features.

We obtained recombination rate in cM/Mb for non-overlapping 200 kb from (Kawakami *et al.* 2014) and assigned gene-wise recombination rates by mapping gene to its corresponding 200 kb window. The density of functional sites was calculated as the density of exons and CNEs. Synonymous substitution rate (d_S) from (Bolívar *et al.* 2015) was used as a proxy for local mutation rate. PPI was obtained from (Uebbing *et al.* 2015). Tissue specificity (τ) was calculated as mentioned above. Gene expression level was computed as a mean of TPM. Gene length was obtained from *Ensembl* gene annotation. All variables were centered and scaled in logistic regression analyses.

We used stepwise regression based on the Akaike Information Criterion to find the best-fit model. To calculate model accuracy, we randomly split the data into training set (80%) and predict set (20%). Estimated coefficients from the best-fit model by the training set were used to predict the presence or absence of ASE in predict set. Model accuracy was calculated as the proportion of genes in the predict set having the correct prediction.

Research Aims

The main objective of this thesis was to develop methods that enabled the analysis of allele-specific gene expression in natural non-model organisms and provided insights into molecular mechanisms and genetic determinants of gene expression evolution. The specific aims of each of these papers were:

Paper I − Develop a Bayesian model that allowed for ASE detection with individual-based gene-level resolution and investigate the prevalence of ASE in wild populations.

Paper II − Develop a computational pipeline that incorporate individual phasing when testing for ASE.

Paper III − Study the determinants of the conservation of gene expression and explore the relationship between *cis*- *trans* compensation and hybrid incompatibilities.

Paper IV − Investigate the evolutionary processes and molecular mechanisms that underline *cis*-regulatory evolution

Summary of papers

Paper I − Bayesian inference of allele-specific gene expression indicates abundant *cis*-regulatory variation in natural flycatcher populations

A powerful approach to study molecular mechanism of gene expression evolution is the analysis of ASE, which is used widely to investigate *cis*-regulatory effects (Arnoult *et al.* 2013; Chen *et al.* 2015; He *et al.* 2010; Pastinen 2010; Steige *et al.* 2015; Yan *et al.* 2002). Crosses of inbred organisms from lab settings largely benefit ASE analysis with good statistical power since inbred lines have small within- and large between-species genetic variation, a sufficient number of replicates, and complete knowledge of the phase. Such methodological advantages have led to a bias in favor of studies on lab organisms (Emerson *et al.* 2010; Glaser-Schmitt *et al.* 2018; Goncalves *et al.* 2012; Mack & Nachman 2017; McManus *et al.* 2010; Schaefke *et al.* 2013; Tirosh *et al.* 2009; Wittkopp *et al.* 2004; Wittkopp *et al.* 2008), and natural populations of non-model organisms are rarely investigated (Tung *et al.* 2015; Wang *et al.* 2017). Therefore, it remains to be explored whether findings in lab settings can be directly transferred to the wild.

To overcome these limitations and study natural organisms, we developed a novel Bayesian negative binomial (NB) approach to detect ASE. Relying on known individual haplotype phase, this approach used multiple heterozygous SNPs within a gene, allowed for overdispersion, calculated a weighted likelihood that captured both gene-wise variation and overall variation, and tested for ASE by an approximate empirical Bayesian solution.

We evaluated the performance of the Bayesian NB approach by (i) simulating multiple data sets under a variety of scenarios, (ii) applying it to a public data set, and (iii) comparing it with another ASE method. Simulation revealed that average true positive rate increased with the number of SNPs, read coverage, and in particular, allelic imbalance. On the other hand, the average false discovery rate was below 0.015 across all tested scenarios. As an independent validation, Bayesian NB approach was applied to an RNA-seq data set from human (Rozowsky *et al.* 2011) and compared to another ASE detection method called MBASED (Mayba *et al.* 2014). We found a considerable

amount of overlap of detected ASE genes in the tested dataset. Taken together, those results indicated that the Bayesian NB approach was a valid method and had powerful control over noise and variation when testing for ASE.

We then used the Bayesian NB approach to study ASE in collared and pied flycatchers. We found that ASE was prevalent, and there were around 7% of genes showing ASE in at least one individual and tissue. Different tissues had different percentage of ASE genes. While many ASE genes were detected in only one tissue or individual, there were 64 ASE genes that were detected in more than one individual-tissue combination. Among those multiple occurred ASE genes, it was more common to have the same individual showing ASE in multiple tissues than multiple individuals showing ASE on the same tissue, highlighting the nature of sequence variation in *cis*-regulation. No particular functionality was found for ASE genes since no significant GO term was enriched.

Note, most of the detected ASE genes showed a major allelic frequency larger than 0.9, reflecting that the power was heavily biased towards an extreme allelic imbalance. This suggested that the estimated ASE percentages in flycatcher populations were likely to be underestimated since subtler allelic imbalance was remained undetected. One important explanation was that sequencing coverage was too low, which prevented ASE detection in lowly expressed genes. In addition, with only a few individuals per species under study, it was a challenge to infer phase from linkage disequilibrium, where preferably, hundreds of individual samples are usually required (Browning & Browning 2007; Menelaou & Marchini 2013). Greater sequencing depth with more sampled individuals could be a possible approach to properly address this issue in the future.

Paper II − RPASE: individual based allele-specific expression detection without prior knowledge of haplotype phase

While recent developments in sequencing technologies allowed the ASE investigation for whole genome at increasingly fine resolution (Edsgärd *et al.* 2016; Leon-Novelo *et al.* 2014; Mayba *et al.* 2014), several technical issues remained to be addressed in the ASE analysis, especially when studying natural organisms. The first one was mapping bias. Mapping bias occurred when reads from two alleles were aligned with a single reference genome. Although many published methods have attempted to solve the issue (Degner *et al.* 2009; Quinn *et al.* 2014; Rozowsky *et al.* 2011; Van de Geijn *et al.* 2015; Wu

& Nacu 2010), mapping bias cannot completely be eliminated (Castel *et al.* 2015). Thus, downstream statistical models that can account for extra variations, manifesting as overdispersion, are of particular importance when testing for ASE. Since overdispersion was estimated from multiple replicates, the second challenge was the difficulty of obtaining wild replicates. Using multiple heterozygous SNPs within a gene was an alternative way to investigate individual ASE (Rozowsky *et al.* 2011; Wang *et al.* 2017). This approach, however, introduced a third challenge since it required a known individual haplotype phase.

To overcome these difficulties and to enable ASE detection within individual organisms, we developed a computational pipeline called RPASE (Read-backed Phasing-based ASE detection). RPASE first used Genome Analysis Toolkit's read-backed phasing to construct individual haplotype, and then, it performed a tailored extra-binomial exact test that allowed for overdispersion and dependency when aggregating multiple SNPs within individual haplotype. RPASE took (i) the mapped RNA-seq reads from a single individual, and (ii) a list of SNPs from the same individual as the input data and produced gene-wise *p*-value for ASE detection.

We evaluated the performance of RPASE by simulating RNA-seq reads data over a range of key parameters. Simulation showed that haplotypes could be constructed correctly for all genes at all settings, reflecting a high accuracy of read-backed phasing. Detection power increased with increasing coverage, increasing SNP density, and decreasing overdispersion, and RPASE had robust control over false discoveries under all simulated scenarios.

We further assessed the performance of RPASE by comparing it with MBASED (Mayba *et al.* 2014). We applied both methods to a public ENCODE dataset from human. When phasing information was available, two methods produced largely concordant ASE results, which served as further support that RPASE had a valid way to circumvent reads dependency. When phasing information was not available, RPASE had a better performance, and MBASED appeared to have spurious ASE calls.

Finally, RPASE was applied to four bird species: collared flycatcher, pied flycatcher, zebra finch and chicken. Our analyses revealed a potentially rich landscape of ASE in those wild birds, demonstrating how RPASE enabled the in-depth ASE analysis with individual-based gene-level resolution.

Paper III − Gene regulatory evolution in natural flycatcher populations is highly tissue-specific and shows distinctive patterns in the testis

Gene expression evolution at the transcription level can be broken up into evolutionary changes in *cis*- and *trans*-regulatory sequence (Fraser 2011; Prud'homme *et al.* 2007; Romero *et al.* 2012; Wittkopp & Kalay 2012; Wray 2007). Recent evidence suggests that *cis*- and *trans*-regulatory changes often modify genes expression in opposite directions with reciprocal effect, i.e., they compensate each other and result in no or little change on overall expression level (Johnson & Porter 2000). Such *cis*-*trans* compensation stabilizes gene expression and maintains a conservation of gene expression (Mack & Nachman 2017).

However, accumulation of such compensatory changes can cause incompatible interactions between regulatory changes in divergent species when hybridizing, and thereby introducing mis-expression in hybrids and gives rise to hybrid dysfunction. This is known as Dobzhansky-Muller hybrid incompatibilities (Haerty & Singh 2006; Mack & Nachman 2017; Signor & Nuzhdin 2018).

To gain insight into the molecular mechanisms underlying gene expression evolution, we analysed gene expression from two flycatcher species as well as their naturally-occurring F1 hybrids. Phylogenetic analysis of mtDNA sequences revealed that all F1 hybrids were from a cross between a female pied flycatcher and a male collared flycatcher.

Differential gene expression analysis between pure species indicated that gene expression evolution was tissue-specific. The tissues that showed the highest number of differentially expressed (DE) gene were testis (1031), followed by heart (160), liver (311), kidney (537), and brain (55). In addition, different tissues showed a different set of DE genes, suggesting either tissue-specific genes were differentially expressed, or broadly expressed genes were regulated in a tissue-specific way. Moreover, a broad Fast-Z effect was observed since the proportion of DE genes in Z-link genes was higher than that of autosomal genes in all tissues.

Mis-expression in F1 hybrids, defined as an expression level in hybrids being either lower or higher than in any of the parental species, was found to be common, especially in brain, heart, kidney, and liver. This result suggested that Dobzhansky-Muller hybrid incompatibilities were widespread in these tissues. Interestingly, while testis showed the highest number of DE genes among tissues, it showed no clear evidence of mis-expression. It was therefore

unlikely that incompatible interactions between regulatory changes explained the observed sterility in F1 hybrids.

When combining DE and ASE analysis, we found a significant association between DE and ASE in both parental species as well as in F1 hybrids for kidney and liver. However, for testis, DE was only associated with ASE in F1 hybrids. In addition, by using conserved non-coding elements (CNEs) located either 5 kb upstream or in introns as proxy for *cis*-regulatory regions, we found that genes with fixed CNE differences had significantly higher expression divergence between species than genes without fixed CNE differences, specifically in testis. Taken together, those evidence demonstrated a critical role of tissue-specific regulatory mechanisms in gene expression evolution. While mis-expression in F1 hybrids was widespread, there was no such evidence for testis. Instead, *cis*-regulatory changes might underlie the rapid expression divergence in testis.

Paper IV − *Cis*-regulatory variation and allele-specific expression in the collared flycatcher (*Ficedula albicollis*) genome

The evolutionary significance of genetic variation in *cis*-regulatory elements (cis-regulatory variation) ha*s* become increasingly clear with recent genome-scale analyses (Crowley *et al.* 2015; He *et al.* 2012; Kita *et al.* 2017; Mack *et al.* 2016; Metzger *et al.* 2016; Santos *et al.* 2014; Steige *et al.* 2017). Such variation has been shown to provide a rich source of material for phenotypic changes, adaptation, and speciation (Mack & Nachman 2017; Signor & Nuzhdin 2018).

However, evolutionary processes and molecular mechanisms underlying *cis*-regulatory variation remain to be explored for most organisms. Like sequence variation in coding region, *cis*-regulatory variation is expected to be affected by mutation, drift, selection and recombination. While weak purifying selection has been found to act on a large amount of *cis*-regulatory variation in *C. elegans*, *Drosophila*, plants, and human (Fay & Wittkopp 2008; Josephs *et al.* 2017; Naidoo *et al.* 2018; Rockman *et al.* 2010), the relative importance of drift and selection on *cis*-regulatory variation remains largely unknown for most organisms.

Moreover, the molecular determinants for whether a gene contains *cis*-regulatory variation and shows ASE is also unclear. In yeast, it has been found that genes that are dosage-sensitive experience stronger constraint on expression variation than genes that are not dosage-sensitive, due to the role of protein-

protein interaction in a regulatory network (Birchler & Veitia 2012; Lehner 2008; Metzger *et al.* 2015). In plants, genes whose expression level are influenced by *cis*-regulatory variation, are also found to have an excess of *trans*-posable element insertions (Hollister & Gaut 2009; Hollister *et al.* 2011; Steige *et al.* 2015). Those findings suggested that gene-specific functions and features may determine whether a gene can be influenced by *cis*-regulatory variation and show ASE. However, in *C. elegans*, background selection is found to be the predominant force that shapes the evolution of regulatory sequences (Rockman *et al.* 2010). In that case, the determinants can be features from much board, genome-scale background, such as local mutation rate, the density of target sites for selection, and local recombination rate. The relative importance of gene-specific functions versus genome-scale features is also unclear for most organisms.

Additionally, while sequence variation in coding sequence can be grouped into nonsynonymous and synonymous substitution, it is difficult to distinguish *cis*-regulatory variation from neutral variation at non-coding DNA sequences (Rockman & Wray 2002; Wittkopp & Kalay 2012). Such difficulty may explain why we have a limited understanding of *cis*-regulatory evolution.

In this study, we used conserved non-coding elements (CNEs) that immediately flank genes as a proxy for *cis*-regulatory elements, and accordingly, used sequence variation in CNEs as a proxy for *cis*-regulatory variation. Depending on whether a gene showed ASE or not, we combined genes into an ASE and a control group. We then tested if the strength of selection in *cis*-regulatory elements and genes differed between the ASE and control groups by using DFE-alpha v2.16 (Eyre-Walker & Keightley 2009).

We found that *cis*-regulatory variation for genes that showed ASE were subject to weaker purifying selection and less frequent positive selection, compared to genes that did not show ASE. In contrast, we did not find such a difference between the two groups for sequence variation in coding sequence. These results indicated that patterns of selection acting on regulatory and coding sequences could be uncoupled and, as a result, the evolution of gene expression and coding sequences could occur independently.

We further found that gene-specific features and functions, such as the number of protein-protein interactions, gene expression level, polymorphic transposable elements, and tissue specificity, other than the genome-scale features in the background, were stronger predictors for ASE (model accuracy 0.69). In addition, our results also showed that gene expression was regulated under tissue-specific manner, where tissues differed in which genes showed ASE, the proportions of ASE, patterns of selection acting on *cis*-regulatory elements, and determinants of ASE.

Collectively, in this study, we investigated evolutionary processes and molecular mechanisms underlying cis-regulatory variation. With the technological advances in many areas, the evolution of cis-regulatory variation can be studied in more organisms in more details, and ultimately, a comprehensive picture of gene expression evolution will emerge in the near future.

Conclusions and future prospects

Evolutionary change in gene expression is an important contribution to phenotypic variation, adaptation, and speciation. Understanding the molecular basis and the forces governing its change are major research topics in biology. A powerful approach to study regulatory evolution is the analysis of allele-specific expression (ASE). A large majority of developed methods for ASE detection are designed for studying model organisms from lab settings, whereas there is a paucity of methods for studies using natural populations. The difficulty of sampling identical wild replicates, a highly frequent polymorphism in the natural population, and a lack of necessary genomic knowledge for natural organisms can contribute to the methodological bias.

In this thesis, two methods for ASE detection are developed, which enable the analysis of ASE in natural organisms. The first method is called Bayesian NB approach and is among one of the first methods that allows for ASE detection at an individual level with the estimation of overdispersion. Relying on the given phasing knowledge, Bayesian NB approach aggregates multiple SNPs, estimates overdispersion, and tests for ASE using negative binomial distribution with an approximate empirical Bayesian solution. A known phasing is critical for Bayesian NB approach, however inferring phasing from a few wild individuals is difficult.

The second method is called RPASE (Read-backed Phasing-based ASE detection), which is developed with a main advantage that no phasing knowledge is required in advance anymore. RPASE preforms physical phasing based on Genome Analysis Toolkit's read-backed phasing, which has a high accuracy with individual phasing resolution. However, due to a nature of physical phasing, which ultimately results from linear transcription of individual alleles, such phasing induces a dependency of read counts. RPASE addresses the dependency issue by designing a tailored extra-binomial exact test, which dissects the dependency by reads stratification and in the meanwhile estimates overdispersion and tests for ASE.

The performance of both methods is assessed by simulating a variety of datasets over many key features. They are also applied to public datasets and compared with other ASE detection method. Those evaluations indicate that

both methods are valid approaches with good ASE detection power and robust control for variation.

PRASE is used later to investigate the molecular mechanisms and evolutionary significance underling gene expression evolution. Incompatible interactions between divergent regulatory sequences are observed extensively in F1 hybrids, indicating a widespread *cis-trans* compensation during gene expression evolution. Tissue-specific regulatory mechanism is found to be critical. Testis, in particular, shows a different pattern compared to other tissues. It has a highest expression divergence among tissues but no clear signature of misexpression. Instead, *cis*-regulatory changes appear to play an important role. These results illustrate that selective constraints associated with pleiotropic effects can be circumvented by a tissue-specific manner of regulatory evolution.

In addition, *cis*-regulatory sequences of genes with ASE show a larger proportion of slightly deleterious mutations and weaker signatures of positive selection than genes without ASE. No such differences are observed for coding sequences. Therefore, patterns of selection that act upon regulatory and coding sequences can be uncoupled and occur independently.

In the future, I think, machine learning has the potential to dramatically further our understanding about evolution. With rapidly decreasing costs of high-throughput sequencing and developments of massively parallel technologies, we produce ever increasing amounts of data at a higher sequencing depth and from more dimensions. These dimensions include genome, transcriptome, proteome, epigenome, phenotypic measurements, ecological properties, and so on. To investigate evolutionary questions at an unprecedented scale and to provide a comprehensive view, there is a need to combine all data together. Traditional statistical approaches, like those that are mentioned in this thesis, are far from sufficient when facing large data sets. Machine learning approaches, such as support vector machine, decision tree, neural network, and deep learning are beginning to be applied in many fields of biology. Ongoing methodological developments and emerging applications promise an exciting future.

Svensk Sammanfattning

Evolutionära förändringar i genuttryck kan bidra till fenotypisk variation, anpassning och artbildning. Att förstå den molekylära bakgrunden till dessa förändringar och de evolutionära krafter som styr dem är viktiga biologiska forskningsämnen. Ett effektivt sätt att studera hur genreglering evolverar är att analysera allelspecifikt uttryck (förkortat ASE, allele-specific expression). De flesta metoder för ASE-detektering är utformade för att studera modellorganismer i labmiljö, medan det finns få tillgängliga metoder för naturliga populationer. Svårigheten med replikat, hög genetisk variation och brist på nödvändiga genomiska redskap i naturliga populationer är bidragande orsaker till detta.

I denna avhandling utvecklas två metoder för ASE-detektering som möjliggör analys av ASE i naturliga populationer. Den första metoden kallas Bayesian NB approach och är en av de första metoderna för ASE-detektering på individnivå som uppskattar överdispersion. Givet att de individuella kromosomernas sekvens (fas) är känd, fogar metoden samman flera enbaspolymorfier, uppskattar överdispersion och testar för ASE genom att använda en negativ binomialfördelning med en approximativ empirisk Bayesiansk lösning. Korrekta kromosomfaser är kritiska för denna metod, men det är svårt att etablera fasen för polymorfier då analysen baseras på ett begränsat antal vilda individer.

Den andra metoden kallas RPASE (Read-backed Phasing-based ASE detection). Metoden har fördelen att den inte kräver kännedom om kromosomfaserna på förhand. RPASE utför fysisk fasning baserat på Genome Analysis Toolkit's ”read-backed phasing”, som har hög precision på kromosomfasning på individnivå. Eftersom kromosomfasning i slutändan är en linjär transkription av individuella alleler på många olika positioner, medför detta ett beroende av antalet sekvensläsningar som finns tillgängligt. RPASE hanterar detta genom att inkludera ett extra-binomialt exakt test som kvantifierar beroendet genom stratifiering av läsningar och under tiden uppskattar överdispersion och testar för ASE.

Båda metodernas prestanda utvärderas genom simulering av en rad dataset med olika egenskaper. Metoderna tillämpas också på publika dataset och jämförs med andra ASE-detekteringsmetoder. Dessa utvärderingar indikerar att

båda metoderna är användbara för att detektera ASE och att de är robusta genom att de inkluderar kontroll för variation.

RPASE används senare för att undersöka molekylära mekanismer och evolutionära konsekvenser av förändringar i genuttryck. Inkompatibla interaktioner mellan divergerande regulatoriska sekvenser observeras i stor utsträckning i F1-hybrider, vilket indikerar en utbredd cis-trans-kompensation under evolutionen av genuttryck. Vävnadsspecifika regulatoriska mekanismer visade sig vara kritiska. Speciellt testikelvävnad visade ett avvikande mönster jämfört med andra vävnader. I testiklar observerades störst variation i genuttryck men inget tydligt tecken på feluttryck. I stället verkade cis-regulatoriska förändringar spela en viktig roll. Dessa resultat illustrerar att selektiva begränsningar förknippade med pleiotropa effekter kan undvikas genom att regulatorisk evolution kan vara vävnadsspecifik.

Dessutom visar analyserna att cis-regulatoriska sekvenser hos gener med ASE har en större andel ofördelaktiga mutationer och svagare tecken på positiv selektion än gener utan ASE. Inga sådana skillnader observerades för de kodande sekvenserna. Därför kan selektionskrafter som verkar på regulatoriska respektive kodande sekvenser vara okopplade och verka oberoende av varandra.

Translated to Swedish by Linnéa Smeds and Niclas Backström

Acknowledgements

I would like to thank my supervisor, Hans Ellegren. I am very grateful that you welcomed me to your lab and have me as your Ph.D. student. You gave me the opportunity to think and discover on my own while being there as support. You provided a great research environment, and I have learned an incredible amount from our group meetings and journal clubs. Also, thank you for your patience when teaching me how to write and publish paper.

I would like to thank my co-supervisor, Yudi Pawitan. Thank you for being my co-supervisor and accompany me on this journey. You introduced me to the exciting area of biostatistics and made me realize that becoming an R expert is very important. I am always amazed by your breath and depth of understanding of statistical models. Your book inspires me all the time, and I wish I know all kinds of likelihood as well as you do. Douglas Scofield, thank you for helping me in my most stressful and hopeless time. Without you, RPASE could definitely not be a proper package and a publication. You always encourage me and give me confidence, which means a lot to me. Carina Mugal, thank you for enlightening me on projects when I was totally lost. I really admire your scientific spirit, your hard work, and your precision at work. Thank you for supporting me and helping me develop my ideas.

Many thanks for other people in the lab. You all contribute to such a great research environment one could only dream of. Takeshi, Alex, Tanja, and Kim, I could always bother you with my questions and your answers were always more worth then I was asking for. Special thanks to people who worked on papers included in this thesis. Severin, thank you for introducing me to allele-specific expression. Rory, thank you for helping me with the data, and for always being available and kind. Niclas and people from Anna Qvarnström's lab, thank you for your work in the field. I also want to thank my office mates Cosima, Ioana, and Zaenab, it has been very nice to have you around.

Linnéa, Homa and Paulina, thank you so much for being my colleagues and friends. Thank you for sharing my pressure and helping me find a way out. You always stand by my side, support me, cheer me up, and make me feel strong again and again. I have learned so many things from you. Hweiyen, thank you for understanding me. You listen to me and give me constructive

suggestions all the time. Ludo, we started this together and grew up together. Thank you for your encouragement and your sense of humor. I also want to give special thanks to Fan Yang Wallentin. Thank you for being my professor and my friend. You taught me not only structure equal model but also how to make a good life. I want to thank Li Sen for introducing me to EBC, Chen Jun, for discussing about allele-specific expression and beyond, and Toby, for all your books.

Outside university, I want to thank people who worked in CSSAU. It was a big part of my life beyond Ph.D. We make this organization great, and local media wrote a lot about us. All the shows and activities we organized definitely gave our hundreds of CSSAU members a great time. I enjoyed all the moments with you during my time. There is not enough space to thank each of you, but all of you truly made my life vivid.

Shi Chengxi, you are one of the smartest people I know. Although we argued almost about everything, you are a real model for me. Tian yuan, Xu Gang and Liu Jing, thank you for taking care of me from my first day in Sweden. My friends from BMC and ÅNGSTRÖM, thank you for letting me in. I always remember all the delicious food you cooked, interesting topics we had, and card game we played.

My dear, Meng, this Ph.D. would not be possible without you. Thank you for loving me, understanding me, tolerating me and supporting me. You are great husband and father. Thank you for taking care of our son for many holidays and weekends alone when I work. My little son Ruiqi, you are the best present I've ever had. Thank you for brightening all my days.

特别感谢我的爸爸妈妈！我在一个如此幸福的家庭长大，爸爸妈妈如同朋友一般的陪伴，给了我许多独立思考的空间和自己做选择的权利，谢谢你们对我无条件的支持。感谢妈妈对我求知欲的培养，让我有机会发现并展现自己的力量。感谢爸爸让我学会时刻保持一颗好奇心并且不断向前探索。最后，谢谢你们如此的爱我，我也爱你们！

References

Anders S, Pyl PT, Huber W (2015) HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169.

Arnoult L, Su KFY, Manoel D*, et al.* (2013) Emergence and diversification of fly pigmentation through evolution of a gene regulatory module. *Science* **339**, 1423-1426.

Bell GDM, Kane NC, Rieseberg LH, Adams KL (2013) RNA-Seq analysis of allele-specific expression, hybrid effects, and regulatory divergence in hybrids compared with their parents from natural populations. *Genome Biology and Evolution* **5**, 1309-1323.

Bersaglieri T, Sabeti PC, Patterson N*, et al.* (2004) Genetic signatures of strong recent positive selection at the lactase gene. *The American Journal of Human Genetics* **74**, 1111-1120.

Birchler JA, Veitia RA (2012) Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proceedings of the National Academy of Sciences* **109**, 14746-14753.

Blekhman R, Oshlack A, Chabot AE, Smyth GK, Gilad Y (2008) Gene regulation in primates evolves under tissue-specific selection pressures. *PLoS genetics* **4**, e1000271.

Bolívar P, Mugal CF, Nater A, Ellegren H (2015) Recombination rate variation modulates gene sequence evolution mainly via GC-biased gene conversion, not Hill–Robertson interference, in an avian system. *Molecular biology and evolution* **33**, 216-227.

Bolívar P, Mugal CF, Rossi M*, et al.* (2018) Biased inference of selection due to GC-biased gene conversion and the rate of protein evolution in flycatchers when accounting for it. *Molecular biology and evolution* **35**, 2475-2486.

Brawand D, Soumillon M, Necsulea A*, et al.* (2011) The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343.

Browning BL, Browning SR (2007) Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. *Genetic epidemiology* **31**, 365-375.

Burri R, Nater A, Kawakami T*, 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**, 1656-1665.

Castel SE, Levy-Moonshine A, Mohammadi P, Banks E, Lappalainen T (2015) Tools and best practices for data processing in allelic expression analysis. *Genome biology* **16**, 195.

Chen J, Nolte V, Schlötterer C (2015) Temperature stress mediates decanalization and dominance of gene expression in *Drosophila melanogaster*. *PLoS genetics* **11**, e1004883.

Clark RM, Wagler TN, Quijada P, Doebley J (2006) A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture. *Nature genetics* **38**, 594.
Consortium G (2017) Genetic effects on gene expression across human tissues. *Nature* **550**, 204.
Coolon JD, McManus CJ, Stevenson KR, Graveley BR, Wittkopp PJ (2014) Tempo and mode of regulatory evolution in Drosophila. *Genome Research* **24**, 797-808.
Craig RJ, Suh A, Wang M, Ellegren H (2018) Natural selection beyond genes: Identification and analyses of evolutionarily conserved elements in the genome of the collared flycatcher (*Ficedula albicollis*). *Molecular ecology* **27**, 476-492.
Cresko WA, Amores A, Wilson C*, et al.* (2004) Parallel genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback populations. *Proceedings of the National Academy of Sciences* **101**, 6050-6055.
Crowley JJ, Zhabotynsky V, Sun W*, et al.* (2015) Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. *Nature Genetics* **47**, 353-360.
Davidson JH, Balakrishnan CN (2016) Gene regulatory evolution during speciation in a songbird. *G3: Genes| Genomes| Genetics* **6**, 1357-1364.
De La Calle-Mustienes E, Feijóo CG, Manzanares M*, et al.* (2005) A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. *Genome research* **15**, 1061-1072.
Degner JF, Marioni JC, Pai AA*, et al.* (2009) Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**, 3207-3212.
DePristo MA, Banks E, Poplin R*, et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**, 491-498.
Dobin A, Davis CA, Schlesinger F*, et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21.
Drapeau MD, Cyran SA, Viering MM, Geyer PK, Long AD (2006) A *cis*-regulatory sequence within the yellow locus of *Drosophila melanogaster* required for normal male mating success. *Genetics* **172**, 1009-1030.
Edsgärd D, Iglesias MJ, Reilly S-J*, et al.* (2016) GeneiASE: Detection of condition-dependent and static allele-specific expression from RNA-seq data without haplotype information. *Scientific Reports* **6**, 1.
Ellegren H, Smeds L, Burri R*, et al.* (2012) The genomic landscape of species divergence in Ficedula flycatchers. *Nature* **491**, 756-760.
Emerson J, Hsieh L-C, Sung H-M*, et al.* (2010) Natural selection on *cis* and *trans* regulation in yeasts. *Genome research* **20**, 826-836.
Eyre-Walker A, Keightley PD (2009) Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Molecular biology and evolution* **26**, 2097-2108.
Fay JC, Wittkopp PJ (2008) Evaluating the role of natural selection in the evolution of gene regulation. *Heredity* **100**, 191-199.
Fraser HB (2011) Genome‐wide approaches to the study of adaptive gene expression evolution: Systematic studies of evolutionary adaptations involving gene expression will allow many fundamental questions in evolutionary biology to be addressed. *Bioessays* **33**, 469-477.
Gilad Y, Oshlack A, Rifkin SA (2006) Natural selection on gene expression. *Trends in Genetics* **22**, 456-461.

Glaser-Schmitt A, Parsch J (2018) Functional characterization of adaptive variation within a *cis*-regulatory element influencing *Drosophila melanogaster* growth. *PLoS biology* **16**, e2004538.
Glaser-Schmitt A, Zečić A, Parsch J (2018) Gene regulatory variation in *Drosophila melanogaster* renal tissue. *Genetics* **210**, 287-301.
Glassberg EC, Gao Z, Harpak A, Lan X, Pritchard JK (2019) Evidence for Weak Selective Constraint on Human Gene Expression. *Genetics* **211**, 757-772.
Gomes S, Civetta A (2015) Hybrid male sterility and genome-wide misexpression of male reproductive proteases. *Scientific reports* **5**, 11976.
Gompel N, Carroll SB (2003) Genetic mechanisms and constraints governing the evolution of correlated traits in drosophilid flies. *Nature* **424**, 931.
Gompel N, Prud'homme B, Wittkopp PJ, Kassner VA, Carroll SB (2005) Chance caught on the wing: *cis*-regulatory evolution and the origin of pigment patterns in *Drosophila*. *Nature* **433**, 481.
Goncalves A, Leigh-Brown S, Thybert D*, et al.* (2012) Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. *Genome Research* **22**, 2376-2384.
Good JM, Giger T, Dean MD, Nachman MW (2010) Widespread over-expression of the X chromosome in sterile F1 hybrid mice. *PLoS genetics* **6**, e1001148.
Graze RM, McIntyre LM, Main BJ, Wayne ML, Nuzhdin SV (2009) Regulatory divergence in *Drosophila melanogaster* and *D. simulans*, a genomewide analysis of allele-specific expression. *Genetics* **183**, 547-561.
Graze RM, Novelo LL, Amin V*, et al.* (2012) Allelic imbalance in Drosophila hybrid heads: exons, isoforms, and evolution. *Molecular Biology and Evolution* **29**, 1521-1532.
Guerrero RF, Posto AL, Moyle LC, Hahn MW (2016) Genome‐wide patterns of regulatory divergence revealed by introgression lines. *Evolution* **70**, 696-706.
Haerty W, Singh RS (2006) Gene regulation divergence is a major contributor to the evolution of Dobzhansky–Muller incompatibilities between species of *Drosophila*. *Molecular Biology and Evolution* **23**, 1707-1714.
Harmston N, Barešić A, Lenhard B (2013) The mystery of extreme non-coding conservation. *Phil. Trans. R. Soc. B* **368**, 20130021.
He D, Choi A, Pipatsrisawat K, Darwiche A, Eskin E (2010) Optimal algorithms for haplotype assembly from whole-genome sequence data. *Bioinformatics* **26**, i183-i190.
He F, Zhang X, Hu J*, et al.* (2012) Genome-wide analysis of cis-regulatory divergence between species in the Arabidopsis genus. *Molecular Biology and Evolution* **29**, 3385-3395.
Hollister JD, Gaut BS (2009) Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome research* **34**, 564.
Hollister JD, Smith LM, Guo Y-L*, et al.* (2011) Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*. *Proceedings of the National Academy of Sciences USA* **7**, 8222.
Hollocher H, Hatcher JL, Dyreson EG (2000) Genetic and developmental analysis of abdominal pigmentation differences across species in the *Drosophila* dunni subgroup. *Evolution* **54**, 2057-2071.
Johnson NA, Porter AH (2000) Rapid speciation via parallel, directional selection on regulatory genetic pathways. *Journal of Theoretical Biology* **205**, 527-542.

Josephs EB, Wright SI, Stinchcombe JR, Schoen DJ (2017) The relationship between selection, network connectivity, and regulatory variation within a population of *Capsella grandiflora*. *Genome biology and evolution* **9**, 1099-1109.
Kawakami T, Mugal CF, Suh A*, et al.* (2017) Whole‐genome patterns of linkage disequilibrium across flycatcher populations clarify the causes and consequences of fine‐scale recombination rate variation in birds. *Molecular ecology* **26**, 4158-4172.
Kawakami T, Smeds L, Backström N*, et al.* (2014) A high‐density linkage map enables a second‐generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution. *Molecular ecology* **23**, 4035-4058.
Kim D, Pertea G, Trapnell C*, et al.* (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* **14**, R36.
King MC, Wilson AC (1975) Evolution at 2 levels in humans and chimpanzees. *Science* **188**, 107-116.
Kita R, Venkataram S, Zhou Y, Fraser HB (2017) High-resolution mapping of *cis*-regulatory variation in budding yeast. *Proceedings of the National Academy of Sciences USA* **1**, 421.
Kopp A, Duncan I, Carroll SB (2000) Genetic control and evolution of sexually dimorphic characters in *Drosophila*. *Nature* **408**, 553.
Kopp A, True JR (2002) Evolution of male sexual characters in the oriental *Drosophila* melanogaster species group. *Evolution & development* **4**, 278-291.
Landry CR, Wittkopp PJ, Taubes CH*, et al.* (2005) Compensatory *cis-trans* evolution and the dysregulation of gene expression in interspecific hybrids of *Drosophila*. *Genetics* **171**, 1813-1822.
Lehner B (2008) Selection to minimise noise in living systems and its implications for the evolution of gene expression. *Molecular systems biology* **4**, 170.
Lemos B, Araripe LO, Fontanillas P, Hartl DL (2008) Dominance and the evolutionary accumulation of *cis*- and *trans*-effects on gene expression. *Proceedings of the National Academy of Sciences USA* **105**, 14471-14476.
Leon-Novelo LG, McIntyre LM, Fear JM, Graze RM (2014) A flexible Bayesian method for detecting allelic imbalance in RNA-seq data. *BMC Genomics* **15**, 920.
Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* **12**, 323.
Li B, Qing T, Zhu J*, et al.* (2017) A comprehensive mouse transcriptomic bodymap across 17 tissues by RNA-seq. *Scientific Reports* **7**, 4200.
Li H, Handsaker B, Wysoker A*, et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079.
Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**, 550.
Mack KL, Campbell P, Nachman MW (2016) Gene regulation and speciation in house mice. *Genome research* **26**, 451-461.
Mack KL, Nachman MW (2017) Gene regulation and speciation. *Trends in Genetics* **33**, 68-80.
Mack KL, Phifer-Rixey M, Harr B, Nachman MW (2019) Gene expression networks across nultiple tissues are associated with rates of molecular evolution in wild house mice. *Genes* **10**, 225.
Marcellini S, Simpson P (2006) Two or four bristles: functional evolution of an enhancer of scute in Drosophilidae. *PLoS biology* **4**, e386.
Marques DA, Lucek K, Meier JI*, et al.* (2016) Genomics of rapid incipient speciation in sympatric threespine stickleback. *Plos Genetics* **12**, e1005887.

Mayba O, Gilbert HN, Liu J*, et al.* (2014) MBASED: allele-specific expression detection in cancer tissues and cell lines. *Genome Biology* **15:405**.
McCoy RC, Wakefield J, Akey JM (2017) Impacts of neanderthal-introgressed sequences on the landscape of human gene expression. *Cell* **168**, 916-927.
McFarlane SE, Sirkiä PM, Ålund M, Qvarnström A (2016) Hybrid dysfunction expressed as elevated metabolic rate in male Ficedula flycatchers. *PloS one* **11**, e0161547.
McKenna A, Hanna M, Banks E*, et al.* (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297-1303.
McManus CJ, Coolon JD, Duff MO*, et al.* (2010) Regulatory divergence in *Drosophila* revealed by mRNA-seq. *Genome Research* **20**, 816-825.
Menelaou A, Marchini J (2013) Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold. *Bioinformatics* **29**, 84-91.
Metzger BP, Duveau F, Yuan DC*, et al.* (2016) Contrasting frequencies and effects of *cis*-and *trans*-regulatory mutations affecting gene expression. *Molecular biology and evolution*, msw011.
Metzger BP, Yuan DC, Gruber JD, Duveau F, Wittkopp PJ (2015) Selection on noise constrains variation in a eukaryotic promoter. *Nature* **521**, 344.
Michalak P, Noor MA (2003) Genome-wide patterns of expression in *Drosophila* pure species and hybrid males. *Molecular biology and evolution* **20**, 1070-1076.
Morris MRJ, Richard R, Leder EH*, et al.* (2014) Gene expression plasticity evolves in response to colonization of freshwater lakes in threespine stickleback. *Molecular Ecology* **23**, 3226-3240.
Nadachowska-Brzyska K, Burri R, Olason PI*, et al.* (2013) Demographic divergence history of pied flycatcher and collared flycatcher inferred from whole-genome re-sequencing data. *Plos Genetics* **9**.
Naidoo T, Sjödin P, Schlebusch C, Jakobsson M (2018) Patterns of variation in *cis*-regulatory regions: examining evidence of purifying selection. *BMC genomics* **19**, 95.
Olds LC, Sibley E (2003) Lactase persistence DNA variant enhances lactase promoter activity in vitro: functional role as a *cis* regulatory element. *Human Molecular Genetics* **12**, 2333-2340.
Pastinen T (2010) Genome-wide allele-specific analysis: insights into regulatory variation. *Nature Reviews Genetics* **11**, 533-538.
Pawitan Y (2001) *In all likelihood: statistical modelling and inference using likelihood* Oxford University Press.
Pennacchio LA, Ahituv N, Moses AM*, et al.* (2006) *In vivo* enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499.
Polychronopoulos D, King JW, Nash AJ, Tan G, Lenhard B (2017) Conserved non-coding elements: developmental gene regulation meets genome organization. *Nucleic acids research* **45**, 12611-12624.
Price TA, Wedell N (2008) Selfish genetic elements and sexual selection: their impact on male fertility. *Genetica* **132**, 295.
Prud'homme B, Gompel N, Carroll SB (2007) Emerging principles of regulatory evolution. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 8605-8612.
Quinn A, Juneja P, Jiggins FM (2014) Estimates of allele-specific expression in *Drosophila* with a single genome sequence and RNA-seq data. *Bioinformatics* **30**, 2603-2610.

Qvarnström A, Rice AM, Ellegren H (2010) Speciation in *Ficedula* flycatchers. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**, 1841-1852.
Ranz JM, Namgyal K, Gibson G, Hartl DL (2004) Anomalies in the expression profile of interspecific hybrids of *Drosophila melanogaster* and *Drosophila simulans*. *Genome research* **14**, 373-379.
Robinson MD, Smyth GK (2007) Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **23**, 2881-2887.
Rockman MV, Skrovanek SS, Kruglyak L (2010) Selection at linked sites shapes heritable phenotypic variation in *C. elegans*. *Science* **330**, 372-376.
Rockman MV, Wray GA (2002) Abundant raw material for *cis*-regulatory evolution in humans. *Molecular biology and evolution* **19**, 1991-2004.
Romero IG, Ruvinsky I, Gilad Y (2012) Comparative studies of gene expression and the evolution of gene regulation. *Nature Reviews Genetics* **13**, 505.
Rozowsky J, Abyzov A, Wang J*, et al.* (2011) AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Molecular Systems Biology* **7**, 522.
SÆTRE GP, BORGE T, MOUM T (2001) A new bird species? The taxonomic status of 'the Atlas Flycatcher'assessed from DNA sequence analysis. *Ibis* **143**, 494-497.
Sanges R, Hadzhiev Y, Gueroult-Bellone M*, et al.* (2013) Highly conserved elements discovered in vertebrates are present in non-syntenic loci of tunicates, act as enhancers and can be transcribed during development. *Nucleic acids research* **41**, 3600-3618.
Santos ME, Braasch I, Boileau N*, et al.* (2014) The evolution of cichlid fish egg-spots is linked with a cis-regulatory change. *Nature communications* **5**, 5149-5149.
Schaefke B, Emerson J, Wang T-Y*, et al.* (2013) Inheritance of gene expression level and selective constraints on *trans*-and *cis*-regulatory changes in yeast. *Molecular biology and evolution* **30**, 2121-2133.
Shapiro MD, Marks ME, Peichel CL*, et al.* (2004) Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* **428**, 717.
Shen Y, Yue F, McCleary DF*, et al.* (2012) A map of the *cis*-regulatory sequences in the mouse genome. *Nature* **488**, 116.
Siepel A, Bejerano G, Pedersen JS*, et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research* **15**, 1034-1050.
Signor SA, Nuzhdin SV (2018) The Evolution of Gene Expression in *cis* and *trans*. *Trends in Genetics* **34**, 489-570.
Snyder MW, Adey A, Kitzman JO, Shendure J (2015) Haplotype-resolved genome sequencing: experimental methods and applications. *Nature Reviews Genetics* **16**, 344-358.
Steige KA, Laenen B, Reimegård J, Scofield DG, Slotte T (2017) Genomic analysis reveals major determinants of *cis*-regulatory variation in Capsella grandiflora. *Proceedings of the National Academy of Sciences USA* **114**, 1087-1092.
Steige KA, Reimegård J, Koenig D, Scofield DG, Slotte T (2015) *Cis*-regulatory changes associated with a recent mating system shift and floral adaptation in *Capsella*. *Molecular biology and evolution* **32**, 2501-2514.
Stevenson KR, Coolon JD, Wittkopp PJ (2013) Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome. *Bmc Genomics* **14**.
Sundararajan V, Civetta A (2011) Male sex interspecies divergence and down regulation of expression of spermatogenesis genes in *Drosophila* sterile hybrids. *Journal of molecular evolution* **72**, 80-89.

Svedin N, Wiley C, Veen T, Gustafsson L, Qvarnström A (2008) Natural and sexual selection against hybrid flycatchers. *Proceedings of the Royal Society B: Biological Sciences* **275**, 735-744.
Swallow DM (2003) Genetics of lactase persistence and lactose intolerance. *Annual review of genetics* **37**, 197-219.
Takahasi KR, Matsuo T, Takano-Shimizu-Kouno T (2011) Two types of *cis-trans* compensation in the evolution of transcriptional regulation. *Proceedings of the National Academy of Sciences USA* **108**, 15276-15281.
Tirosh I, Reikhav S, Levy AA, Barkai N (2009) A yeast hybrid provides insight into the evolution of gene expression regulation. *Science* **324**, 659-662.
Tishkoff SA, Reed FA, Ranciaro A*, et al.* (2007) Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics* **39**, 31.
Tulchinsky AY, Johnson NA, Watt WB, Porter AH (2014) Hybrid incompatibility arises in a sequence-based bioenergetic model of transcription factor binding. *Genetics* **198**, 1155-1166.
Tung J, Zhou X, Alberts SC, Stephens M, Gilad Y (2015) The genetic architecture of gene expression levels in wild baboons. *Elife* **4**, e04729.
Turner LM, White MA, Tautz D, Payseur BA (2014) Genomic networks of hybrid sterility. *PLoS genetics* **10**, e1004162.
Uebbing S, Künstner A, Mäkinen H*, et al.* (2016) Divergence in gene expression within and between two closely related flycatcher species. *Molecular ecology* **25**, 2015-2028.
Uebbing S, Konzer A, Xu L*, et al.* (2015) Quantitative mass spectrometry reveals partial translational regulation for dosage compensation in chicken. *Molecular biology and evolution* **32**, 2716-2725.
Van de Geijn B, McVicker G, Gila Y, Pritchard JK (2015) WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nature Methods* **12**, 1061-1063.
Van der Auwera GA, Carneiro MO, Hartl C*, et al.* (2013) From FastQ data to high‐confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics* **43**, 11.10. 11-11.10. 33.
Verta JP, Landry CR, MacKay J (2016) Dissection of expression‐quantitative trait locus and allele specificity using a haploid/diploid plant system–insights into compensatory evolution of transcriptional regulation within populations. *New Phytologist* **10.1111**, nph.13888.
Verzijden MN, Ten Cate C, Servedio MR*, et al.* (2012) The impact of learning on sexual selection and speciation. *Trends in ecology & evolution* **27**, 511-519.
Visel A, Prabhakar S, Akiyama JA*, et al.* (2008) Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nature genetics* **40**, 158.
Visel A, Rubin EM, Pennacchio LA (2009) Genomic views of distant-acting enhancers. *Nature* **461**, 199.
Wang M, Uebbing S, Ellegren H (2017) Bayesian inference of allele-specific gene expression indicates abundant *cis*-regulatory variation in natural flycatcher populations. *Genome biology and evolution* **9**, 1266-1279.
Wang M, Uebbing S, Pawitan Y, Scofield DG (2018) RPASE: individual based allele‐specific expression detection without *prior* knowledge of haplotype phase. *Molecular ecology resources* **18**, 811-819.
Wittkopp PJ, Haerum BK, Clark AG (2004) Evolutionary changes in *cis* and *trans* gene regulation. *Nature* **430**, 85-88.

Wittkopp PJ, Haerum BK, Clark AG (2008) Regulatory changes underlying expression differences within and between Drosophila species. *Nature Genetics* **40**, 346-350.

Wittkopp PJ, Kalay G (2012) *Cis*-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics* **13**, 59.

Wray GA (2007) The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics* **8**, 206-216.

Wu TD, Nacu S (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873-881.

Yan H, Yuan WS, Velculescu VE, Vogelstein B, Kinzler KW (2002) Allelic variation in human gene expression. *Science* **297**, 1143-1143.

Yanai I, Benjamin H, Shmoish M*, et al.* (2005) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650-659.

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution* **24**, 1586-1591.

Zhuang Y, Adams KL (2007) Extensive allelic variation in gene expression in Populus F1 hybrids. *Genetics* **177**, 1987-1996.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1823

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-384493

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2019