UPPSALA
UNIVERSITET

Evolutionary investigations of transcriptome data in two flycatcher species

Severin Uebbing

Degree project in biology, Master of science (2 years), 2011
Examensarbete i biologi 45 hp till masterexamen, 2011
Biology Education Centre and Department of Evolutionary Biology, Institution of Ecology and Genetics, Uppsala University
Supervisors: Axel Künstner and Hans Ellegren

Contents

Abbreviations

cDNA – complementary DNA
DE – differential expression, differentially expressed
FDR – false discovery rate
NGS – next-generation sequencing
PCR – polymerase chain reaction
RPKM – reads per kilobase of exon model per million mapped reads

1 Abstract

The collared flycatcher (*Ficedula albicollis*) and the pied flycatcher (*F. hypoleuca*) represent a sister species model which has been studied for several decades especially at Uppsala University in terms of evolutionary ecology (e.g., sexual selection, speciation and hybridisation). Their sympatric occurrence on two Swedish islands in the Baltic sea and their tendency to hybridise where they co-occur make them an interesting model system for the investigation of speciation and hybridisation.

Differences between recently diverging species may as well be due to regulatory changes in expression patterns as to mutations in coding sequence. While the latter has been in focus a lot in previous research, far less is known about the former. This is even more so for differential expression between species, and investigations on this topic might lead to illuminating insights into the speciation process and the species' differences.

The advent of next-generation sequencing techniques has led to an increasing amount of genome and transcriptome sequencing efforts. For the two flycatcher species, genome and transcriptome data have recently been acquired using Illumina sequencing. RNA from several tissues has been sequenced from several individuals of both species. The genome sequence of a collared flycatcher is currently being assembled.

In this study, I investigated different properties of the transcriptomes of collared and pied flycatchers. First, I tested different methods for their performance in the detection of differential expression, using means of estimating the quality of the results that I deduced by myself. I further tested properties of the results of the method I found to perform best, `baySeq`, and analysed different general properties of our data, i.e. patterns of differential gene expression, dosage compensation, sex-biased gene expression, and tissue specificity.

I found that there is a substantial amount of differential gene expression between the species. The amount of genes that differ in expression level varies between tissues. Sex-biased gene expression occurs and differs between tissues and species. Genes with patterns of differential gene expression tend to be more tissue specific than genes that occur in more general patterns. As could be shown for birds before, I did not find a pattern of general dosage compensation. Male flycatchers tended to show a balanced autosomal-to-Z ratio whereas female birds had a general expression level of Z-linked genes that scaled at half that of autosomes.

Transcriptome sequencing may assist in investigations on speciation, sex-related expression patterns, and further general aspects of genome and transcriptome evolution. The ample amount of different possible analyses shows that transcriptome sequencing provides a valuable approach to gather further insights into the buildup of an organism's genome and its function.

2 Introduction

2.1 Pied and collared flycatchers – an ecological model system for mating systems, speciation, and hybridisation

Collared flycatcher (*Ficedula albicollis*) and pied flycatcher (*F. hypoleuca*) form a sister species model for ecological and evolutionary research that has been intensively studied especially in Uppsala since the early 1980ies (Alatalo *et al.* 1981, Qvarnström *et al.* 2010). The two species diverged probably less than 1 million years ago (Qvarnström *et al.* 2010) during the Pleistocene, but occur in widely overlapping ranges in Central and Eastern Europe at present. However, in mainland Sweden only pied flycatchers occur. About 150 years ago, collared flycatchers colonised Gotland and formed a hybrid zone on the Swedish island (Lundberg & Alatalo 1992) that even expanded further to Öland during the 1960ies (Qvarnström *et al.* 2009).

Where the species distributions overlap, hybridisation occurs regularly although fitness of hybrids is significantly reduced (Alatalo *et al.* 1982, Svedin *et al.* 2008). Current theory suggests that certain loci, that have been significantly differentiated in this young sister species pair, provide hybrid incompatibilities (Backström *et al.* 2010). Getting the full genome sequence from a species of this model system would present a worthwhile step in getting closer to knowing the mating barriers that prevent these species to fully admix.

2.2 Sequencing the genome of collared flycatchers using next-generation sequencing technique

Just a few years ago, whole genome sequencing projects used conventional Sanger sequencing technique which consumed huge amounts of money and workload, at orders of magnitude that were far from feasible for single university units. With the recent advent of next-generation sequencing (NGS) methods as the Roche/454 Genome Sequencer FLX (Margulies *et al.* 2005) and the Illumina Genome Analyzer (Bentley *et al.* 2008), this has lapsed.

In a recent attempt to acquire the whole genome sequence of a male collared flycatcher individual, almost 2 billion Illumina sequencing reads and more than 6.3 million 454 sequencing reads have been gathered. Along with that, about 635 million Illumina reads of transcriptome data have been sequenced. The RNA came from nine different tissues

from five male and five female pied flycatchers from the area around Uppsala and from five male and five female collared flycatchers from Öland. The tissues include brain, kidney, liver, lung, muscle, skin, ovary, and testis, as well as eight embryos from each species.

The goal of my thesis work was to investigate properties of our flycatcher species' transcriptome, as there are differential expression (DE) between species, tissues, and sexes, general patterns of sex-biased expression and dosage compensation, or tissue-specificity in expression patterns. A first step in doing so was to find and evaluate suitable methods for these analyses (see below, section 2.3).

2.3 Differential gene expression analysis

Despite the ongoing discussion on the relative importance of protein coding vs. regulatory mutations in evolution (see e.g. Hoekstra & Coyne 2007), there are reasonable arguments (Wilson *et al.* 1974) and a growing amount of evidence (Hoekstra & Coyne 2007) that differences in expression level play a major role in adaptation and (incipient) speciation. Recent studies have shown that both established (Enard *et al.* 2002) and young (Wolf *et al.* 2010) sister species show substantial differences in expression patterns and that transcriptome studies can provide valuable insights into how species diverge.

Recently, several methods have been developed that allow to infer DE between different groups of samples using lists of read counts from transcriptome libraries sequenced with NGS techniques (Oshlack *et al.* 2010). The advantage of specially developed statistical testing methods over simple comparisons using statistics like the fold change in expression, lies in the incorporation of variance and statistical significance into the testing procedure. Differences in mean expression of single genes with heavily inflated variances could appear to be large, yet accounting for variance in the testing procedure would reveal their statistical insignificance. Incorporating the mean expression into testing allows to decide on statistical significance for lowly expressed genes which suffer from a lack of statistical power due to their small sample sizes.

My first task was thus to test and evaluate these methods for their performance. This was not straight-forward as the parameters of the distribution of NGS reads are largely unknown or known to be non-trivial. Their discrete nature suggests that read counts follow a binomial or Poisson distribution. However, it has been shown that for biological replicates, read count distributions are overdispersed and therefore modelling of counts using a Poisson or binomial distribution would lead to a large number of false positives (Robinson & Smyth 2008). Instead, the use of overdispersed log-linear (beta-binomial; Baggerly *et al.* 2003) and overdispersed logistic models (gamma-Poisson or negative binomial; Lu *et al.* 2005) have been suggested.

Furthermore, the count variance of a transcript differs depending on library size and mean expression. This is e.g. because variances are often inflated, especially for small count numbers. That in turn leads to statistical power issues and it has been suggested that this has to be accounted for (Robinson & Smyth 2008). Several different models have been proposed to address this problem (Robinson & Smyth 2007, 2008, Anders &

Huber 2010).

The methods I tested for their performance in the detection of DE address these problems in different ways. `DEGseq` (Wang *et al.* 2010a, 2010b) models read counts as Poisson distributed and tests for DE using different methods, e.g. likelihood ratio test, Fisher's exact test, or plot based methods which calculate a Z score from a comparison of mean concentration and expression ratio between conditions. The two main methods recommended by the authors are plot based testing methods using a random sampling model (MARS) or a model that assumes technical replicates (MATR). By default data is not normalised, but options allow to normalise by median expression or using a LOWESS normalisation. As mentioned, modelling read count data as being Poisson distributed does not account for overdispersion and later methods therefore use different approaches. `DEGseq` is further the only tested method which does not take replicates into account, just as technical replicates when using the MATR method. In combination with the fact that the Poisson distribution model does not account for overdispersion from biological replicates one would expect an increased number of false positives detected by this method in our biologically replicated dataset.

For modelling read count data, `edgeR` (Robinson *et al.* 2010) uses a negative binomial regression and for estimating the distribution parameters it uses a moderated conditional maximum likelihood approach (Robinson & Smyth 2008). In order to test for DE, the authors deduce an 'exact test' procedure from Fisher's exact test (Robinson & Smyth 2007). Furthermore, the programme implements a normalisation method according to the overall mean fold change between sample groups while disregarding genes that are unique to one group (Robinson & Oshlack 2010).

`DESeq` (Anders & Huber 2010) uses a simple library size normalisation and a negative binomial distribution for modelling read counts. The variance parameter is linked to the mean read count and these simplifications lead to a very fast, yet claimed to be appropriate, performance of the method.

Hardacstle & Kelly (2010) proposed a Bayesian inference method for detecting signatures of DE in a method called `baySeq`. Read counts can be modelled as Poisson distributed with gamma distributed gene-wise means or negative binomially distributed. The parameters for the distributions, which have to be considered as priors for the Bayesian framework, are estimated using a quasi-likelihood approach. Data is normalised by library size. This is the only method which allows for more complicated experimental designs than pairwise comparisons between groups, e.g. including of more than two sample groups which can show all possible combinations of expression patterns (i.e. $A = B = C; A = B \neq C; A \neq B = C; A = C \neq B; A \neq B \neq C$).

As the fraction of truly DE genes was unknown, I first had to develop some criteria for evaluating the methods. Therefore, I conceived certain assumptions to be met by the programmes: (1) A large overlap of genes declared to be DE between different methods would indicate that they consistently find the same genes and would put trust in methods that show a particularly large overlap. (2) The total number of significant DE genes should be in a reasonable range. This is, naturally, a rather fuzzy and weak criterion. However, a result may seem rather improbable if, e.g. the same non-reproductive tissue in the same species between sexes is said to have more than, say, 30% of DE genes, as

the tissues perform the same function in both sexes. (3) DE genes should show a larger fold change than non-DE genes. DE genes should have (4) the same mean expression level and (5) variance as non-DE genes within the same group of samples. And there should be no bias for (6) longer or shorter genes or (7) GC content for the detection of DE. In order to test these assumptions, I performed statistical and graphical analyses using `R`.

There are, however, some problems with these assumptions. Genes with higher read counts tend to have larger sample sizes which in turn leads to greater statistical power (Oshlack *et al.* 2010). A bias for genes with higher mean expression should therefore be expected. This in turn has an impact on two other of the aforementioned criteria. During the preparation of sequencing libraries, cDNA is sheared and size selected. Longer genes will then have more pieces of the selected size in the pool than shorter genes just due to their greater lengths. Thus, longer genes will be sequenced to a greater depths than shorter genes. On the other hand, longer genes show on average lower expression levels than shorter ones (Urrutia & Hurst 2003). And it has been shown that extremely GC rich or poor gene regions are being sequenced less frequently by Illumina sequencing technique, most likely due to different chemical affinities during PCR steps (Bentley *et al.* 2008).

2.4 The extent and mechanism of dosage compensation in birds is largely unknown

Sex determining mechanisms and sex chromosomes have evolved independently multiple times (see e.g. Ellegren 2011), and every such instance can potentially come along with fundamental differences in its realisation. Theory predicts that the evolution of sex chromosomes is accompanied by degradation and multiple gene loss of the non-recombining, heterogametic sex chromosome, i.e. the Y chromosome in XX/XY male heterogametic systems and the W chromosome in female heterogametic ZZ/ZW systems (Muller 1950). This would inevitably lead to dosage inequalities between males and females as the homogametic sex possesses two copies of the larger and gene-richer homogametic sex chromosome (X or Z chromosome, respectively). To compensate for these inequalities, dosage compensation mechanisms have evolved (Muller 1932; reviewed in Straub & Becker 2007). In the XX/XY chromosome system as it is realised in mammals, one of the two female X chromosomes gets inactivated randomly (Ohno *et al.* 1959, Lyon 1961, Beutler *et al.* 1962; reviewed in Leeb & Wutz 2010) in order to compensate for the unequal amount of gene product between males and females. In *Drosophila*, the male X chromosome gets up-regulated to the level of female XX expression (Hamada *et al.* 2005), and a different mechanism causes a similar result in *Caenorhabditis* (Meyer & Casson 1986).

On the existence, extent, and mechanism of dosage compensation in ZZ/ZW systems far less is known. Recent studies suggested that a general dosage compensation between males and females is not existing in birds (Ellegren *et al.* 2007, Itoh *et al.* 2007, Wolf

& Bryk 2011), in silk worms (Zha *et al.* 2009, but see Walters & Hardcastle 2011), and in schistosome parasites (Vicoso & Bachtrog 2011). It rather seems like dosage compensation of Z-linked genes, where necessary, is facilitated locally on a gene-by-gene basis in birds (Mank & Ellegren 2009). However, these studies have been performed using either 454 sequencing (Wolf & Bryk 2011) – which is less appropriate for quantitative transcriptomics studies due to the smaller sample sizes of the method compared to the here used Illumina NGS technique – or microarrays (all other mentioned studies) – which are less suitable for comparisons of different genes with each other due to different binding affinities of the probes (Xiong *et al.* 2010).

There are, in theory, two possible ways of dosage compensation; 1) the homogametic chromosome is downregulated to levels of the single heterogametic chromosome: this would mean that sex chromosome expression is only half that of autosomes. 2) Expression of the heterogametic chromosome is being upregulated to expression levels of the homogametic chromosome. The most commonly quoted theory of how unequal sex chromosomes evolve states that, during the process of decay of the heterogametic sex chromosome, its sister chromosome increases its expression level until dosage is compensated (Muller 1950, Charlesworth 1978). I.e., ultimately, the second of the aforementioned possibilities should become realised. For a long time it has been thought that this is indeed the case, and all published studies seemed to agree with that (see Mank 2009). However, recent investigations in human (Xiong *et al.* 2010) as well as the mentioned studies in ZW systems provided contradicting results.

2.5 Determining the sex of the embryo samples

The embryo samples in the dataset were of unknown gender because they were not determined when sequencing libraries were created from the eggs. With our transcriptome data, there are in principle two possible ways of retrieving this information. One can, first, perform a `blast` search for W-linked genes in order to find out which embryos are female, as only the genomes of female birds harbour W-linked genes. Having identified some females this way, one can, secondly, use male-to-female expression ratios of Z-linked genes. This can only work given that there is no dosage compensation between male and female embryos, what has been shown to be the case in chicken (Ellegren *et al.* 2007).

2.6 Tissue specificity of DE genes

It has been shown that sex-biased genes are on average more tissue specific than genes that are equally expressed between sexes (Mank *et al.* 2008). This seems reasonable because fewer tissues are affected by expression changes of a tissue specific gene compared to a ubiquitous gene. Thus a specific gene might be less constrained to changes. However, this mechanism is not necessarily restricted to sex-biased gene expression but might as well play a role in differential gene expression between species. If a gene is less

constrained to a given expression level it should be both more liberal to changes between sexes and in terms of evolutionary divergence. In order to investigate these patterns in our data, I correlated τ, a measure for a gene's tissue specificity, to log fold change values of genes DE between both sexes and species.

The used measure for tissue specificity, τ (Yanai *et al.* 2004), ranges from 0 (equal expression in all tissues analysed) to 1 (expressed specificly in a single tissue). However, as absolute equal expression in all tissues is an unnatural ideal, there are basically no τ-values below 0.2.

Further, I used log fold change values for this analysis instead of, e.g. `baySeq` posterior probabilities. I did so because posterior probabilities (or p-values from other methods) depict the *probability* of DE. They focus rather on statistical significance than on actual strength of a change. This approach is useful for making solid inferences on a gene's likelihood of being DE, but disregard to some extent the amount a gene is DE. Fold change values, on the other hand, depict the *strength* of DE. Therefore, they present the more appropriate measure of DE for this analysis.

3 Materials and Methods

3.1 Preparations: Steps done prior to my project work

Five male and five female pied flycatchers, as well as eight eggs, were sampled just outside Uppsala. Five male and five female collared flycatchers, as well as eight eggs, were collected on Öland. Tissues were prepared in all the same way and stored in RNAlater at -80°C. Libraries were prepared using the standard Illumina mRNA sequencing protocol with the addition of taggs according to the Illumina multiplexed paired-end sequencing protocol. Steps include: isolation of poly-A mRNA, fragmentation and size selection; transcription into cDNA; ligation of six bp adapter sequence taggs and purification of ligation products; PCR amplification. 12 samples were run on a single Illumina HiSeq2000 lane, including one whole tissue from one species and two samples from another tissue to fill up a lane. Sequencing resulted in over 630 million 100 bp transcriptome reads in total. Read duplicates were removed and reads were quality filtered prior to assembly and analysis.

Reads are sequenced from different positions of different genes, following a random sampling process. (Note, however, that it has been shown that GC content has a non-negligible influence on the likelihood of being sampled in the sequencing process; Bentley *et al.* 2008.) In order to get gene count tables, one needs to summarise all single reads that map to the same gene. Therefore, gene models are required to which the sequenced reads can be mapped. Because the collared flycatcher genome sequence was still in progress of being assembled, gene models were taken from the zebra finch genome (Warren *et al.* 2010). The lineages towards zebra finch and flycatchers seperated approximately 40 million years ago (Barker *et al.* 2004, Cracraft & Barker 2009) and gene sequence divergence estimates are at approximately 17% (A. Künster, pers. comm.), i.e. the sequences are conserved to some extent, yet showing some amount of divergence. Collared flycatcher genome reads were mapped to these using `BWA` version `0.5.9` (Li & Durbin 2009) in order to retrieve flycatcher-specific gene models. By mapping the transcriptome reads of each library to these gene models read count tables were created that I used for the downstream analyses of my own project work.

3.2 Hardware

Most analyses were performed using a Lenovo ThinkPad R500 64 bits with an Intel Core2 Duo CPU at 2.0 GHz and 2 Gb memory running `Linux Kubuntu`. Memory and/or

processing intense operations were run on the SNIC-UPPMAX computer system `Kalkyl`. It runs `Scientific Linux` and provides nodes with two Quad-core Intel Xeon 5520 Nehalem 2.26 GHz, 8 Mb cache processors, and 24, 48, or 72 Gb memory, respectively.

3.3 Testing methods for the detection of differential gene expression

I tried to test the four different methods using two different approaches, i.e. first, creating simulated data with a known fraction of DE genes, and second, comparing their performance on real data.

As mentioned in section 2.3, current literature suggests a negative binomial read count distribution with a dispersion parameter to be modelled depending on gene expression mean for each gene (see e.g. Hardcastle & Kelly 2010). Thus, I tried to simulate NGS read count data using the open source statistical programming language `R` version `2.10.1` (R Development Core Team 2009). `R` provides functions for modelling data following a given distribution including a negative binomial distribution that samples randomly values of such a model given a mean and a dispersion parameter. However, as already mentioned, the dispersion of NGS read count data varies with expression level of a gene which further complicates proper modelling.

In a next step, I tested all four programmes for their performance on real data. Before testing the methods, I had to decide on certain parameters that were used throughout the analyses. I tested `DEGseq` version `1.2.2` for its methods MARS and MATR and decided to use MARS results for comparisons with other programmes, because the MATR method consistently resulted in very few significant genes (usually below 20) and because the replicate type of our data is not technical (i.e. from the same biological sample, on a different run) but biological (from two different biological samples). All other parameters were kept as default. The default cut off is at a p-value of 0.001.

I tested both, a common dispersion parameter and a tag wise estimated dispersion parameter for analysis with `edgeR` version `2.2.3`. The tag wise parameter was consistently more conservative as one might expect because increased variance of single genes compared to the average is not mistaken for DE. In this line of argumentation I used the tag wise model for comparisons with other methods. The `prior.n` parameter which is used to 'squeeze' dispersion towards a common value was taken as default and recommended (i.e. 10). All other options were kept as default, as was the cut off at an false discovery rate (FDR) of 5% after the method of Benjamini & Hochberg (1995).

`DESeq` version `1.5.1` was used entirely according to default settings. The default cut off is at an FDR of 10% after the method of Benjamini & Hochberg (1995).

When using `baySeq` version `1.6.0`, I used the negative binomial distribution approach in order to account for overdispersion where present. This led to longer computing times and the necessity to estimate additional priors for the distribution parameters. However, runs on experimental data using a Poisson distribution model gave similar results as with a negative binomial model. I used 10,000 iterations for finding priors. As recommended

by the authors, I took a posterior probability for DE of 0.3 as cut off for a gene being DE. This corresponds to a posterior probability of 0.7 for non-DE and might therefore seem to be irrational in the first place. This is even more so as posterior probabilities for a gene always sum up to 1, i.e. a gene with a posterior probability of DE between 0.3 and 0.5 has a higher posterior probability for not being DE than being DE. However, as one can show (see fig. 4.6 after Hardcastle & Kelly 2010), there appear basically no genes without a difference in expression level between groups above a posterior probability for DE of 0.2. Taking a cut off value of 0.3 might therefore lead to increased sensitivity of the method even for lowly expressed genes with less statistical significance but clear difference in expression.

3.4 Dosage compensation

For analyses on genome-wide and chromosome-wise expression patterns, I used `R` for statistical and graphical analyses. First, I normalised data using RPKM (reads per kilobase of exon model per million mapped reads) normalisation (Mortazavi *et al.* 2008) in order to control for differences in sequencing depth and gene length:

$$R_{i,j} = \frac{10^9 \cdot c_{i,j}}{N_i \cdot l_j}$$

where $c_{i,j}$ is the number of reads mapped on the contigs for gene j in library i, N_i is the total library size for library i and l_j is the length of gene j.

Furthermore, it has been shown that GC content can have an influence on our estimate of expression level through different sequencing efficiency (Bentley *et al.* 2008) and differences in GC content could therefore lead to biased results. In order to control for these effects, I used a linear model with $\log_2$ male-to-female gene expression ratio as response, gene GC content as a covariate, and chromosome as a factor with 31 levels (chromosomes 1A, 1B, 2 – 4, 4A, 5 – 15, 17 – 26, 28, Z, linkage group 22, and Un). Chromosomes 16 and 27 as well as linkage groups 2 and 5 were removed because they comprised less than four genes each. After weighting the chromosomes by their respective number of genes and multiple testing correction, only GC content, Z chromosome, and the intercept of the correlation were highly significant. I then transformed the data by subtracting the effects of GC content and intercept from the expression values.

With these corrected estimates, I analysed male-to-female ratio on a chromosome-wise basis for all tissues in both species. This would uncover imbalanced male-to-female expression for the Z chromosome. In addition to that, I analysed autosomal vs. Z chromosome expression for both sexes, each tissue and both species. This shows if there is a pattern of dosage compensation between autosomal and Z-linked expression level.

3.5 Determining the sex of the embryo samples

The sexes of the embryos that have been sampled for this study were not known when they were collected. Furthermore, there were no W-linked genes in the dataset, because the W chromosome sequence is not assembled for zebra finch. Because we have deduced gene models from the zebra finch genome assembly, there are no W-linked genes in the dataset. So in order to retrieve gender information on the embryo libraries, I created `blast` databases from all of our embryo raw read libraries and used `blastn` version `2.2.23+` (Zhang *et al.* 2000) to search for alignments with chicken sequences for *HINTW* and *CHDW*, two genes located on the female W chromosome. Blasting this way (our reads as databases and the gene sequences as queries) is more computational efficient and produces equivalent results (not shown). Significant hits in this search would only be expected from libraries derived from female embryos. I verified hits by blasting them and their respective paired reads against Entrez's online nr/nt database using `megablast`[1].

Afterwards I verified my findings by calculating male-to-female gene expression ratios between embryo tissues that had no hit with any of the aforementioned W-linked genes and those that did have hits. The expression ratios for Z-linked genes would then be expected to be positive, i.e. one would expect higher mean gene expression, in male embryos (see section 3.4).

3.6 Tissue specificity of DE genes

In order to investigate tissue specificity patterns in our data, I plotted gene tissue specificity, τ, against log fold change values of DE genes between both sexes and species. I needed to transform the expression data several times; first, expression data cover a very broad range, mainly due to an extremely long tail towards high expression levels. This is commonly encountered by a $\log_2$ transformation. Expression levels are then centred around 0. This is, however, not suitable for this analysis, because high negative and positive fold changes equal each other out. The use of absolute fold change values solves this problem, leaving the data bound to 0. In order to finally achieve approximately normally distributed data, I used yet another $\log_2$ transformation.

τ is calculated as:

$$\tau = \frac{\sum_{i=1}^{N}(1 - \frac{x_i}{x_{max}})}{N-1}$$

where N is the number of tissues analysed, x_i the RPKM value for gene x in tissue i, and x_{max} the highest RPKM value measured for gene x over all tissues.

[1] `http://blast.ncbi.nlm.nih.gov/Blast.cgi`

4 Results

4.1 Methods for the detection of differential gene expression

4.1.1 Approaches to test methods for the detection of DE using simulated datasets

Read counts were simulated according to data one would retrieve from Illumina NGS using the `R` function `rnbinom` which simulates sample data from a negative binomial distribution. It takes values for the number of observations n, mean μ, and the dispersion parameter *size* as parameters for simulating data. I used empirical values from our flycatcher data for μ (≈ 30) and *size* (≈ 0.1, for $size = (\mu + \mu^2)/var$). Graphical analysis using quantile-quantile (QQ) plots showed that my simulated data departed from real data in that variance for small values was overestimated and variance for high values was underestimated. My attempts to correct for either bias resulted in worse behaviour for the other. However, when simulating datasets with a simulated fraction of DE genes (either 10%, 20%, or 30%), both `baySeq` and `edgeR` methods commonly estimated a fraction of genes being DE close to the true values and with less than 10% false positives. This result may indicate that the methods perform well, but because I cannot be sure that the distribution of read counts as simulated was close to natural conditions, results might not be comparable to analyses of real data.

To summarise, simulated data did not lead to satisfying results, whereas I was able to find means for testing the different methods for the detection of DE genes using real data (see below).

4.1.2 Testing methods for the detection of DE on real data

I tested the four different methods according to the criteria mentioned in section 3.3 using comparisons of expression in all different tissues between both species, as well as between sexes for both species seperately. Shown are the results for the comparison of brain tissue between ten collared and ten pied flycatchers. All other comparisons were very similar and led to consistent results.

A large overlap between different methods would provide reliability for methods that find consistently the same genes to be DE. A graphical analysis of the overlap between methods is given by a venn diagram (fig. 4.1). The largest group in this diagram consist of genes being unique to `DEGseq` (893). The second largest group consists of genes being common to all methods (634), showing a generally overlap between methods. Further

large groups include genes unique to `baySeq` (225), genes shared between `baySeq`, `DESeq`, and `DEGseq` (202), and those shared between `baySeq` and `DEGseq` (155), and `DESeq` and `DEGseq` (153). Other categories were substantially smaller, often zero or close to.

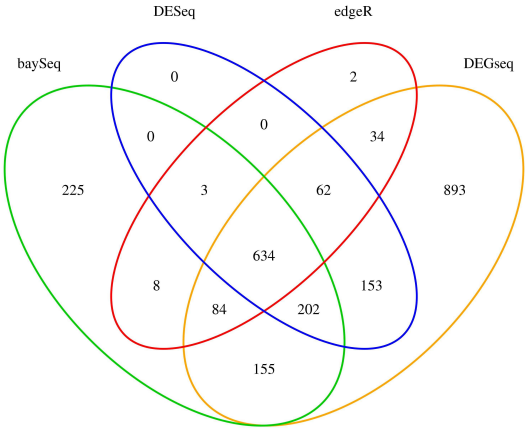


Figure 4.1: **Venn diagram showing overlap and differences between the four different methods tested.** Shown are the results for the detection of DE between pied and collared flycatchers in brain tissue for 10 individuals each.

Different methods find not only different genes but also different amounts of genes to be DE. Table 4.1 shows the numbers of genes declared to be DE by the different methods for seven different tissues (brain, kidney, liver, lung, muscle, skin, and embryo) between 10 pied and 10 collared flycatchers. There are between 6.3% and 33.7% genes found to be DE depending on the tissue analysed and the method used. `DEGseq` declared consistently higher fractions of genes to be DE, whereas `DEseq` and `edgeR` were rather conservative; `baySeq` was somewhat in between. Differences in the fraction of DE genes between tissues investigated are not as pronounced as between methods applied, despite yet following rather clear trends. E.g. brain and muscle tissues had always rather low numbers of significant genes.

One should expect truly DE genes to have on average a more pronounced change in expression level than non-DE genes. I tested the different methods for this assumption by a comparison of the density distribution of expression fold change for genes declared to be DE by a given method with the whole dataset of genes (fig. 4.2). First of all, all methods found genes with a higher expression difference average than the whole genome dataset. `DEGseq`, however, declared also genes to be DE which did not show any difference in expression. Finally, it appeared that `edgeR` needed a bigger fold change of expression in order to detect a pattern of DE or, in other words, it seems to be more conservative than the other programmes.

There should be no bias for genes with higher expression in being DE. Therefore, I compared the results of the four tested methods with the whole dataset in terms of their distribution of expression level mean (fig. 4.3a). All methods provided results with on average higher mean. `DEGseq` and `DESeq` showed a more pronounced bias for genes with higher expression level. Although both `baySeq` and `edgeR` showed comparably little

Table 4.1: **Numbers of genes found to be DE by the different methods between pied and collared flycatchers.** Numbers in brackets behind tissue names indicate total numbers of genes in the different tissues; numbers in brackets behind genes found to be DE by the different methods show percentage DE genes of total genes in a tissue. Note general patterns, e.g. very high numbers for `DEGseq` or consistently low numbers for brain tissue.

Tissue	`baySeq`	`DEGseq`	`DESeq`	`edgeR`
Brain (13,196)	1311 (9.9%)	2217 (16.8%)	1054 (8.0%)	827 (6.3%)
Embryo (13,767)	3890 (28.3%)	4642 (33.7%)	2189 (15.9%)	1263 (9.2%)
Kidney (13,457)	3506 (26.1%)	4240 (31.5%)	1714 (12.7%)	1493 (11.1%)
Liver (12,084)	3165 (26.2%)	3375 (27.9%)	1881 (15.6%)	1654 (13.7%)
Lung (13,630)	2240 (16.4%)	4094 (30.0%)	1134 (8.3%)	1447 (10.6%)
Muscle (12,225)	1701 (13.9%)	2868 (23.5%)	1018 (8.3%)	941 (7.7%)
Skin (13,873)	3318 (23.9%)	4662 (33.6%)	1117 (8.1%)	1491 (10.7%)

bias for genes with higher expression level, only `baySeq` picked up genes with very low expression and furthermore resembled better the shape of the density distribution of the whole dataset than other methods.

As with the mean expression level, the variance in expression (i.e. the within-group variance) should be equally distributed for DE and for non-DE genes. Figure 4.3b compares the variance distribution of the results of the different methods with all genes in the dataset. All methods showed slight preferences for genes with higher variance. The results of `DEGseq` and `DESeq` had higher variances than the other two methods. `baySeq` was most sensitive for genes with very small variance.

Several issues are known of how gene length can have an influence on expression level and, through this, on statistical power while testing for DE. However, there is no obvious reason why one would expect longer genes being more likely to be DE than shorter ones. I compared the gene length distributions of genes declared to be DE by the different methods to the whole tissue dataset (fig. 4.4a). The density distributions of all tested methods followed the distribution of the whole dataset quite well so that I decided to use statistical significance testing to get better estimates for the fit of the distributions. Because the $\log_2$ gene length distributions followed a normal distribution quite well, I compared the means of the distributions with the sample mean of the whole dataset using two-sided t-tests. Except for `DEGseq` ($p = 0.0611$), the means of the $\log_2$ length distributions of all methods were significantly different from the length distribution of all genes (`baySeq` $p < 10^{-9}$, `DESeq` $p < 10^{-4}$, `edgeR` $p < 10^{-15}$). I controlled for potential influences of non-normality using non-parametric Mann-Whitney tests but the results were the same. The differences in mean were – for the three deviating methods – always skewed towards smaller gene lengths.

Finally, there should be no bias in GC content for the methods' capabilities of the detection of DE. As mentioned before, extreme GC content can have a negative influence on the amount of sequencing reads produced. Figure 4.4b shows GC content density distributions of genes found to be DE by the different methods in comparison with

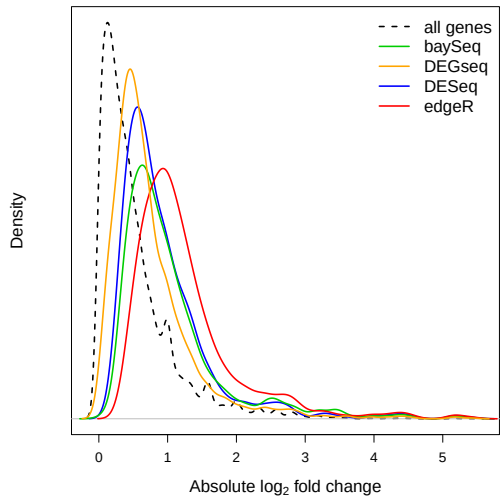
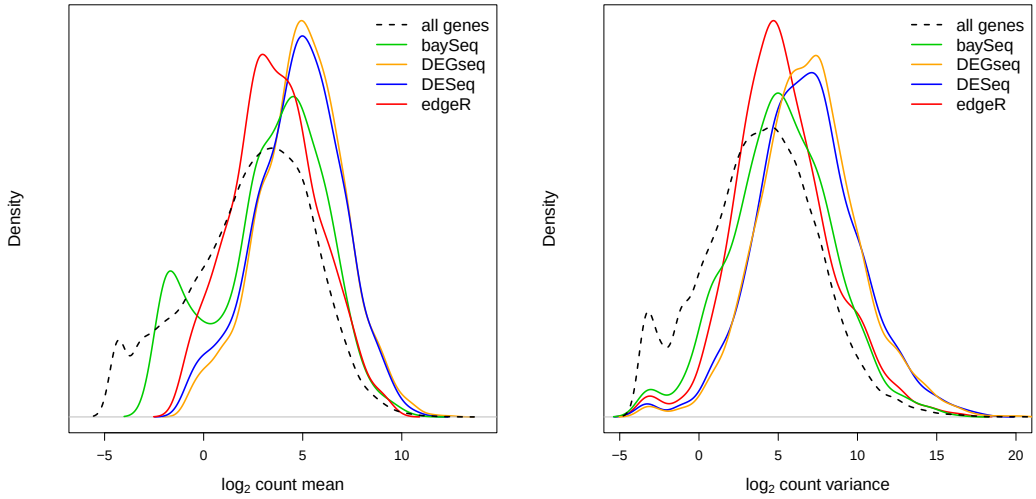


Figure 4.2: **Testing different methods according to expression fold change.** Comparison of the density distributions of the absolute $\log_2$ fold change of genes declared to be DE by the four different methods tested. The dashed line shows the density distribution of all genes in the dataset including DE and non-DE genes.

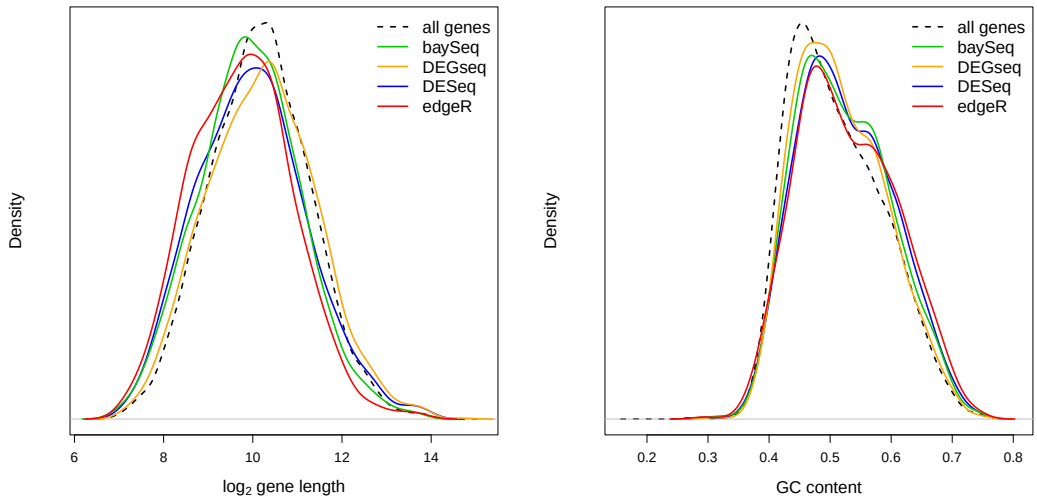
the whole dataset. Although the distributions were quite close to the distribution of the complete dataset, all methods showed a slight bias towards higher GC contents. Furthermore, the distributions seemed to be bimodal with a larger peak at a GC content of approximately 0.45 and a smaller one at 0.55 GC. The distributions of the genes found to be DE by the different methods showed this second peak more pronounced than the dataset of all genes, due to their shift towards higher GC content values. As this already suggests, Mann-Whitney tests for differences in mean GC distribution between DE genes and the whole dataset were highly significant for all methods ($p < 10^{-4}$).

In order to explain the results from figure 4.4, I checked for a correlation between GC content and gene length as found earlier by Jjingo and colleagues (2011). Indeed, there was a significant negative correlation between these properties ($t = -15.3, p < 10^{-51}$, fig. 4.5). A bias for smaller genes could therefore explain a bias for higher GC content.



(a) (b)

Figure 4.3: **Testing methods for bias in read count mean and variance. (a)** Comparison of the density distributions of the mean read counts of genes declared to be DE by the four different methods tested. **(b)** Comparison of the density distributions of the read count variances. Variance distributions for DE genes are shown for collared flycatcher expression variance only, as variance between the two groups would inflate this statistic. The dashed curves shows the density distribution of all genes in the dataset including DE and non-DE genes.



(a) (b)

Figure 4.4: **Testing methods for bias in gene length and GC content. (a)** Comparison of the density distributions of the $\log_2$ lengths of genes declared to be DE by the four different methods tested. **(b)** Comparison of the GC content distributions. The dashed curves shows the density distribution of all genes in the dataset including DE and non-DE genes.

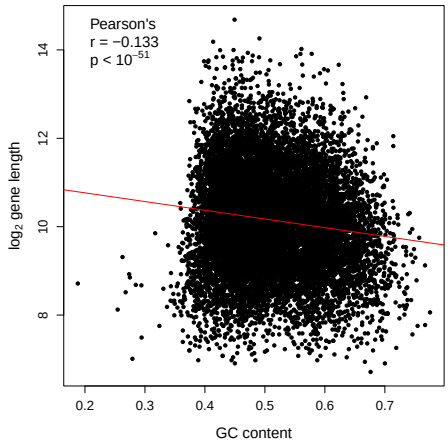


Figure 4.5: **Relationship of gene length and GC content.** Gene length and GC content were negatively correlated with each other ($t = 15.3, p < 10^{-51}$).

4.1.3 Properties of differential gene expression detection by `baySeq`

I picked `baySeq` as the method of choice because it seemed to give the least biased results in all comparisons made (see discussion, section 5.1.2). In order to get insights into the way differential gene expression is being estimated by this method, I performed analyses to explore further properties of the data. I created volcano plots, displaying `baySeq`'s posterior probabilities for DE against $\log_2$ fold change values (fig. 4.6). Genes with a posterior probability higher than 0.2 usually had a clear fold change in expression or, in other words, there appeared basically no genes with no difference in gene expression at values higher 0.2. I can therefore agree with the authors that a cutoff of 0.3 is appropriate. In some tissues, however, i.e. embryo, kidney, skin, and ovary, there appeared very few genes in the region of no fold change in expression and yet were assigned a high probability for DE. Figure 4.7b suggests a reason for this observation (see below). Most plots appeared approximately symmetrical, i.e. there was an equal amount of genes found to be DE between the two groups. However, for kidney, liver, and ovary tissues, approximately two thirds of the DE genes were biased towards collared flycatchers; for testis, three fourths of the DE genes were biased towards pied flycatchers.

The next question then was, what gene properties could potentially influence the detection of DE by `baySeq`. One approach was to compare the position of a gene in the plots of figure 4.6 with its expression level. In order to do so, I colour coded the expression level of a gene. The plot for brain tissue is given as an example (fig. 4.7a). Highly expressed genes were assigned high posterior probabilities in case they showed a recognisable pattern of DE and very low probabilities in case they were equally expressed. Not as highly expressed genes needed to show more pronounced fold changes in order to be declared DE. Genes that showed strong fold changes in expression but were not assigned to be DE usually exhibited rather low expression levels. Very lowly expressed genes clustered at probability values of approximately 0.02 to 0.04 as no clear decision could be made.

Another possible influence could come from different levels of variance in gene expression. Figure 4.7b is similar to figure 4.7a, only the colour codes for variance in expression level instead of mean expression. Embryo has been chosen as example tissue because it illustrates that genes with high posterior probability for DE but no actual mean difference in expression level between groups showed high variances in expression.

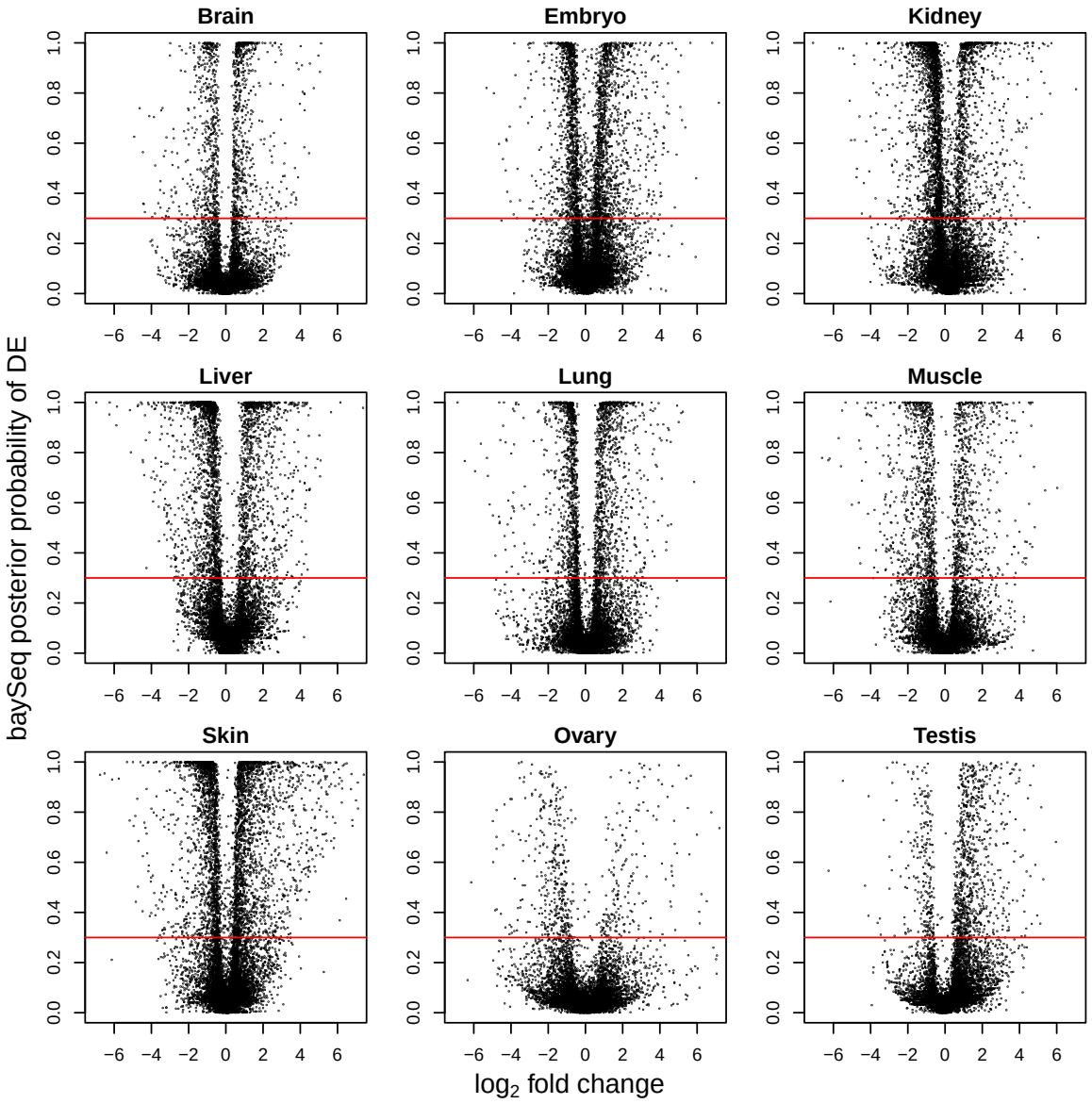
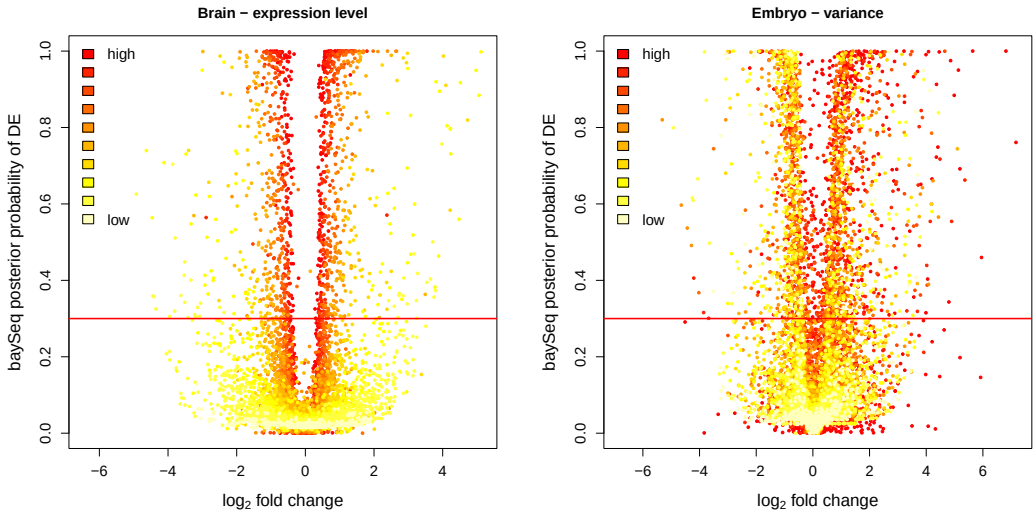


Figure 4.6: **`baySeq` posterior probabilities of DE against $\log_2$ fold change in expression.** The different plots give an expression of how well `baySeq` managed to estimate DE in the different tissues. The abscissae show $\log_2$ fold change values between the two different samples. Genes with positive values are expressed higher in pied flycatchers, negative values indicate higher expression in collared flycatchers; a $\log_2$ fold change of 0 means there is no DE between groups. The ordinates show `baySeq` posterior probability values for DE; they range from 0 (low probability of DE) to 1 (high probability of DE). The cutoff value of 0.3 posterior probability for DE as suggested by Hardcastle & Kelly (2010) is shown by a horizontal red line. The comparisons are for 10 collared vs. 10 pied flycatchers except for gonads (five vs. five) and embryo (eight collared vs. six pied flycatchers).



(a) (b)

Figure 4.7: `baySeq` **posterior probabilities of DE against $\log_2$ fold change in expression with colour indicating two different gene properties.** **(a)** Colour indicating expression level. All genes were categorised into deciles (10-quantiles) according to their mean expression over all libraries in both groups. Deciles were assigned different colours according to their expression levels, dark red for high to light yellow for low expression. Mean expression values ranged approximately from 10^{-2} to 10^{4}. Only the plot for brain tissue is shown, but other tissues were of similar appearance (see fig. S.1). **(b)** Colour showing expression variance. Genes were categorised according to their mean expression variance within the two groups. Deciles were assigned different colours according to their expression variances, dark red for high to light yellow for low variance. Mean variance values ranged from below 10^{-2} to about 10^{2}. The plot for embryo tissue is shown as an example but other tissues were of analogous appearance (see fig. S.2). The cutoff value of 0.3 posterior probability for DE is shown by a horizontal red line.

One possible strength of baySeq compared to the other tested methods is that it can handle more complex models than simple two group comparisons. In order to get an estimate of how well this works, I analysed the expression data of five tissues with multiple group comparisons. I estimated models of DE between five female collared flycatchers vs. all 15 others (i.e. five male collared, five female pied, and five male pied flycatchers), female pied flycatchers vs. all others, male collared flycatchers vs. all others, male pied flycatchers vs. all others, all 10 males vs. all 10 females, all 10 collared vs. all 10 pied flycatchers, and no DE between groups, i.e. equal expression in all 4 different groups. For each given model, baySeq returns a posterior probability value. All posterior probabilities for a gene add up to 1. It is, therefore, possible that more than one model is assigned a higher posterior probability than 0.3, although this is chosen to be the significance cutoff. I.e. more than one model can turn out to be significant (although unlikely). The results for this analysis are shown in table 4.2. Differences between species resulted in the highest numbers of DE genes. However, genes that were significant for all other models sum up to roughly the same amount. Genes that showed sex-biased expression in the same pattern for both species, i.e. in the contrast between 10 males and 10 females of both species, were rather few.

Table 4.2: baySeq **run with multiple models.** Contrasts denote which comparison has been done in the respective model, where 'collared ♀' means a comparison between five female collared flycatchers and all other 15 libraries. Numbers show the genes found to match that expression model with a baySeq posterior probability of 0.3 or better for five different tissues; in numbers in parentheses show percent of total genes in a tissue. Gonad tissues were excluded because comparisons between genders are not reasonable, embryo was excluded because of its high variance and skin because pied female skin samples were of bad quality at that point.

Contrast	Brain	Kidney	Liver	Lung	Muscle
Collared ♀	154 (1.2%)	155 (1.2%)	186 (1.5%)	138 (1.0%)	353 (2.9%)
Pied ♀	126 (1.0%)	1139 (8.5%)	263 (2.2%)	878 (6.4%)	150 (1.2%)
Collared ♂	120 (0.9%)	104 (0.8%)	529 (4.4%)	319 (2.3%)	133 (1.1%)
Pied ♂	226 (1.7%)	696 (5.2%)	266 (2.2%)	363 (2.7%)	190 (1.6%)
Gender	139 (1.1%)	142 (1.1%)	154 (1.3%)	565 (4.1%)	181 (1.5%)
Species	934 (7.1%)	2326 (17.3%)	2321 (19.2%)	1365 (10.0%)	1137 (9.3%)

I tried to get an estimate of the influence of replication on the detected amount of differential gene expression. In a first attempt, I plotted the increase of genes detected to be DE by baySeq against the number of biological replicates used (fig. 4.8). The correlation turned out to be linear and did not seem to level off when reaching 10 replicates, the level of replication per species of our data.

In a second attempt, I performed a jackknife analysis where I randomly sampled 100 times five pied flycatcher samples and five collared flycatcher samples without replacement out of the 10 available libraries and estimated the number of DE genes using baySeq. There were 419 – 1179 genes found with a mean of 737 genes. Their distribution appeared to be bimodal with a bigger peak at about 550 genes and a smaller

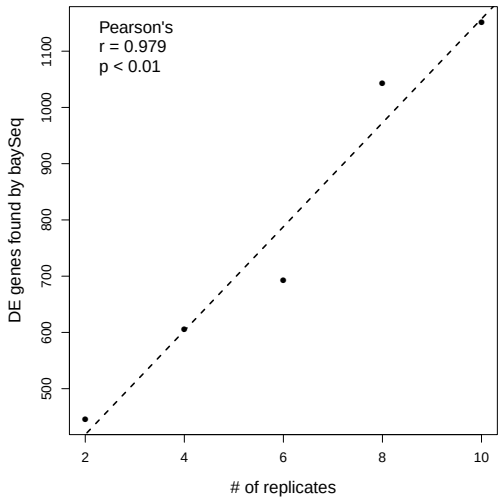


Figure 4.8: **The number of genes detected to be DE by `baySeq` increases linearly.** I estimated the number of DE genes between pied and collared flycatchers in brain tissue using an increasing amount of biological replicates. I plotted the mean of two different runs for each of the replicate number categories except for 10, where I could only use all our libraries once. A linear model was highly significant ($t = 14.52, p < 0.001$) for the effect of replicate number on the amount of DE genes detected.

peak at approximately 900 genes (fig. 4.9). I checked if differences in the gender distribution between jackknife replicates could account for the bimodality but tests were non-significant (although distributions showed a trend towards the expected direction).

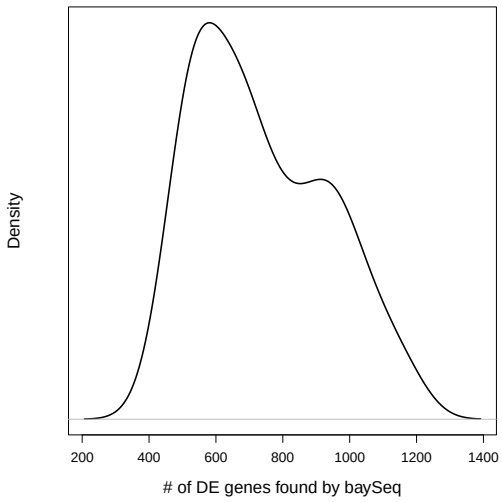


Figure 4.9: **Density distribution of the number of DE genes in a jackknife analysis.** I used `baySeq` to estimate the amount of DE genes when sampling five replicates of each species randomly for 100 times.

4.2 Dosage compensation

In order to investigate general trends in sex-linked expression patterns, I looked at broad scale expression level differences between autosomes and the Z chromosome. I calculated gene-wise male-to-female $\log_2$ expression ratios and compared them by chromosome (fig. 4.10). Autosomes were showing equal expression in males and females, whereas Z chromosome expression was significantly different between males and females. Median $\log_2$ male-to-female ratio was 0.58 for the Z chromosome, according to a 1.49 fold expression in males compared to females. These results were consistent for all different tissues analysed in both species. Also, density distributions of gene expression profiles from autosomes were significantly different from those of the Z chromosome (fig. 4.11).

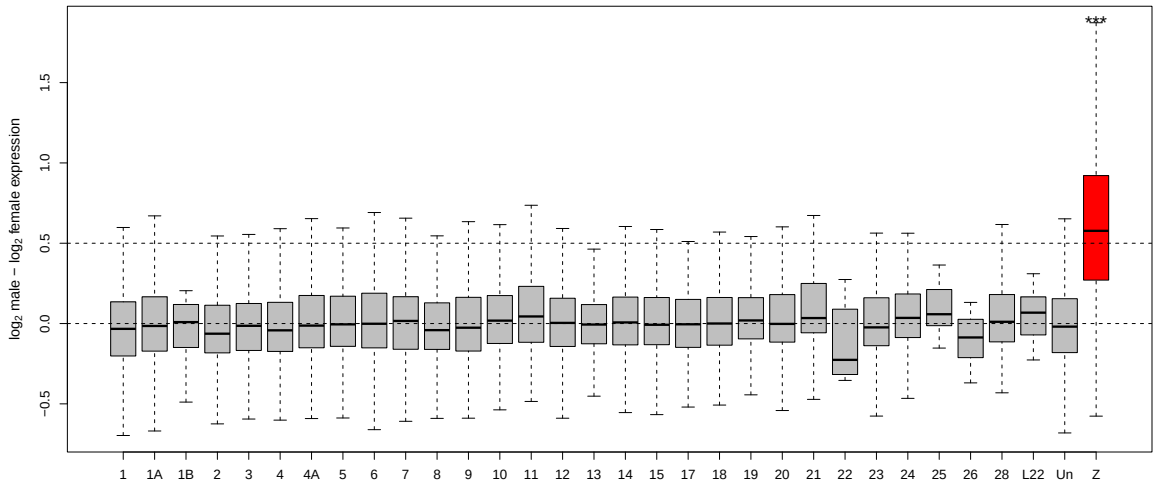


Figure 4.10: **Log_2 male-to-female expression ratios** over all sequenced tissues in pied and collared flycatchers by chromosome. The autosomes tended to show equal expression as their median $\log_2$ difference was around 0, whereas the $\log_2$ difference median of the Z chromosome was at 0.58 (95% confidence interval: -0.34 – 1.41, Mann-Whitney $p < 10^{-15}$, shown by asterisks in plot). Estimates are corrected for GC content and weighted by number of genes per chromosome.

In a second approach I compared absolute expression profiles of autosomal genes with that of the Z chromosome in both males and females for all tissues in both species. I found that female expression level for the Z chromosome always scaled clearly below autosomes (range 0.53 fold that of autosomal expression level for brain in collared flycatchers to 0.73 for embryo in pied flycatchers; Mann-Whitney q-value $< 10^{-6}$, FDR correction after Storey & Tibshirani 2003), indicating a lack of complete female dosage compensation (fig. 4.12). On the other hand, male Z expression ranged from 0.69 (testis in collared flycatchers) to 1.10 (embryos of pied flycatchers) of the level of autosomes. Brain Z chromosome expression level in collared flycatchers and testis Z chromosome expression in both species were significantly lower then the autosomal level (Mann-Whitney $q < 0.05$ and $q < 0.0001$, respectively).

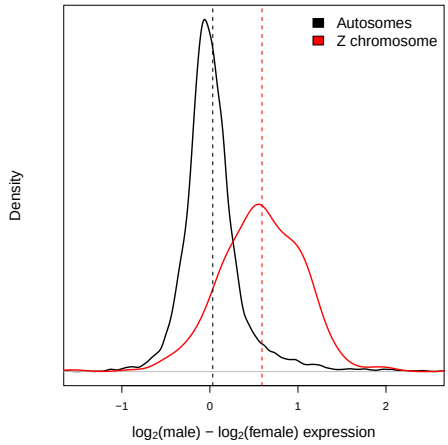


Figure 4.11: **Density distribution of the $\log_2$ male-to-female ratio** for genes expressed in all sequenced tissues of pied and collared flycatchers. A Mann-Whitney test was highly significant ($p < 10^{-15}$). Estimates are corrected for GC content and number of genes per chromosome.

Having detected differences in expression level between males and females, I wanted to investigate which fraction of genes showed patterns of DE and if these genes showed any distinct common features. Therefore, I used `baySeq` to infer DE between the sexes in all different tissues for both species (tab. 4.3). Between 1.7% and 8.9% of the genes in a library were found to be DE (212 – 979 genes). There was one clear outlier, skin tissue in pied flycatchers with 16.3% (i.e., 2242 genes) (see discussion, section 5.2). Sex-biased genes were always significantly enriched for Z-linked genes (binomial test, $p < 10^{-3}$ after FDR correction for multiple testing according to Benjamini & Hochberg 1995). When analysing the fraction of male-biased genes, I distinguished between autosomal and Z-linked genes to avoid the influence of the presence of only one active Z in females compared to males. I calculated the expected number of male-biased autosomal genes as half the number of autosomal sex-biased genes; the expected number of Z-linked, male-biased genes was calculated as half the total number of sex-biased Z-linked genes weighted by the mean expression ratio of Z-linked genes between males and females. As one would expect, Z-linked, sex-biased genes were significantly enriched for biases towards males ($p < 10^{-3}$ for all tissues except skin in pied flycatchers). However, although several tests were significant for the majority of the comparisons among autosomal, Z-linked genes, some libraries showed significant skews towards male-bias, some towards female-bias.

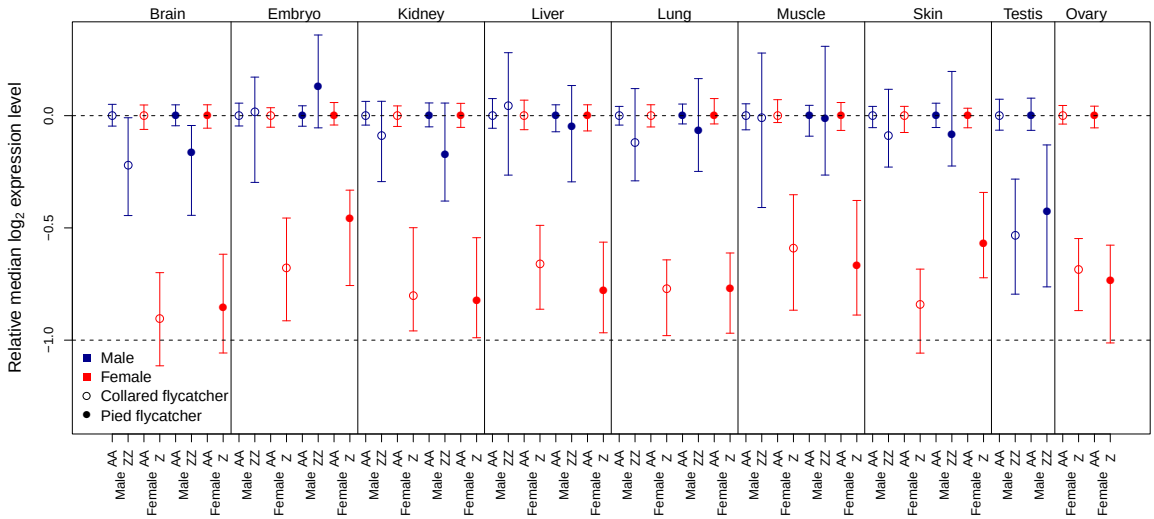


Figure 4.12: **Median expression levels of autosomes and Z chromosome**, relative to autosomal median expression level, for the nine different tissues. Male values are indicated in blue, female values in red, collared flycatchers are shown by open, pied flycatchers by filled circles. Error bars show bias corrected 95% confidence intervals for the median as determined by 10,000 bootstrap replicates. Estimates are corrected for GC content and number of genes per chromosome.

Table 4.3: **Numbers of genes found to be sex-biased using `baySeq`.** C or P after tissue name refer to collared and pied flycatcher species, respectively. Numbers in parentheses indicate total gene number in the respective tissue's dataset. DE indicates differential expression, in this case sex-biased expression between males and females of the same species. 'Z-linked' shows Z-linked genes found to be DE. '♂-bias^{A}' are male-biased, autosomal DE genes; '♂-bias^{Z}' are male-biased, Z-linked DE genes. Expected (exp.) values are estimated according to the fraction of Z-linked genes in the whole dataset for Z-linked DE genes; according to a random binomial distribution of 0.5 for autosomal male-biased genes; and accordingly, just weighted by the tissue-wide male Z to female Z expression ratio, for Z-linked male-biased genes. Asterisks indicate significance levels of 0.05 (*), 0.01 (**), and 0.001 (***) in a two-tailed binomial test after correction for multiple testing according to Benjamini & Hochberg (1995).

Tissue (# genes)	DE (%)	Z-linked (exp.)	♂-bias^{A} (exp.)	♂-bias^{Z} (exp.)
Brain C (12,666)	212 (1.7%)	88 (10)***	28 (50)***	85 (66)***
Brain P (12,658)	239 (1.9%)	102 (12)***	50 (58)	101 (77)***
Embryo C (13,172)	236 (1.8%)	95 (11)***	40 (54)*	94 (72)***
Embryo P (13,481)	1201 (8.9%)	147 (55)***	696 (450)***	134 (104)***
Kidney C (13,123)	356 (2.7%)	147 (17)***	44 (88)***	146 (110)***
Kidney P (12,771)	979 (7.7%)	167 (44)***	453 (340)***	166 (100)***
Liver C (11,421)	443 (3.9%)	100 (20)***	128 (140)	98 (64)***
Liver P (11,232)	504 (4.5%)	105 (24)***	159 (180)*	102 (71)***
Lung C (13,029)	427 (3.3%)	191 (21)***	156 (102)***	191 (131)***
Lung P (13,412)	919 (6.9%)	257 (44)***	451 (292)***	256 (202)***
Muscle C (11,760)	394 (3.4%)	114 (18)***	92 (115)**	110 (85)***
Muscle P (11,538)	290 (2.5%)	86 (14)***	68 (88)**	81 (69)***
Skin C (13,392)	336 (2.5%)	152 (15)***	83 (74)	151 (121)***
Skin P (13,781)	2242 (16.3%)	164 (104)***	398 (924)***	126 (134)

4.3 Determining the sex of the embryo samples

In order to sex our embryo libraries, I blasted the W-linked chicken *HINTW* and *CHDW* genes against databases created from our embryo libraries. There were no hits for *HINTW*, but several libraries had a hit against *CHDW*, i.e. collared_02, collared_03, collared_07, collared_09, and pied_23. The fact that there was only one hit found in each positive library and none for *HINTW* indicates that these genes are only very lowly expressed and therefore likely difficult to pick up at all.

I used the web based version of `megablast`[1] to verify the hits by blasting them and their respective paired reads against Entrez's nr/nt database. All these reads, also those paired reads that had no hits in the `blast` search before, mapped highly significant to avian *CHDW* sequences (*e*-values of around 10^{-40}; hits for other genes were at *e*-values of 10^{-35} for CHD1 or worse). Reads that had no hits before mapped to regions out of the sequence used for blasting in the first step, e.g. to 3' or 5' untranslated regions.

In order to verify the inferred genders from the `blast` search, I compared the expression level of an individual embryo to those of the individuals that had hits for W-linked genes (fig. 4.13 and 4.14). Autosomal individual-to-female ratios should always be equally, but Z-linked individual-to-female expression ratios should be positive for male individuals. For collared flycatchers, this picture held nicely; it became clear that collared_06 was female, whereas collared_04, collared_05, and collared_08 were male. For pied flycatchers, the status was not that clear; there was a lot more variance in chromosomal expression level. This was not due to the fact that I had only one clear female individual to compare to (tested a posteriori, not shown), but most likely due to additional variance in the pied flycatcher embryo libraries. However, due to their strongly negative individual-to-female Z ratio, pied_12 – pied_15 were most likely female; pied_22 had a well pronounced positive individual-to-female Z ratio, and was most likely male; pied_18 and pied_24 were of ambiguous appearance.

Because the pattern for pied flycatchers was not very clear, I created one further plot for sexing the pied embryo individuals, showing the sum of all Z-linked RPKM normalised by the RPKM sum of all genes in the library (fig. 4.15). The so far known picture is being confirmed as pied_24 showed very high, and pied_12 – pied_15 very low values, at a level of that of pied_23; pied_18 and pied_22 scaled somewhere in between, clearly distinct from both, pied_24 and the female pied_23, but in the approximate range of the male collared flycatchers.

[1] `http://blast.ncbi.nlm.nih.gov/Blast.cgi`

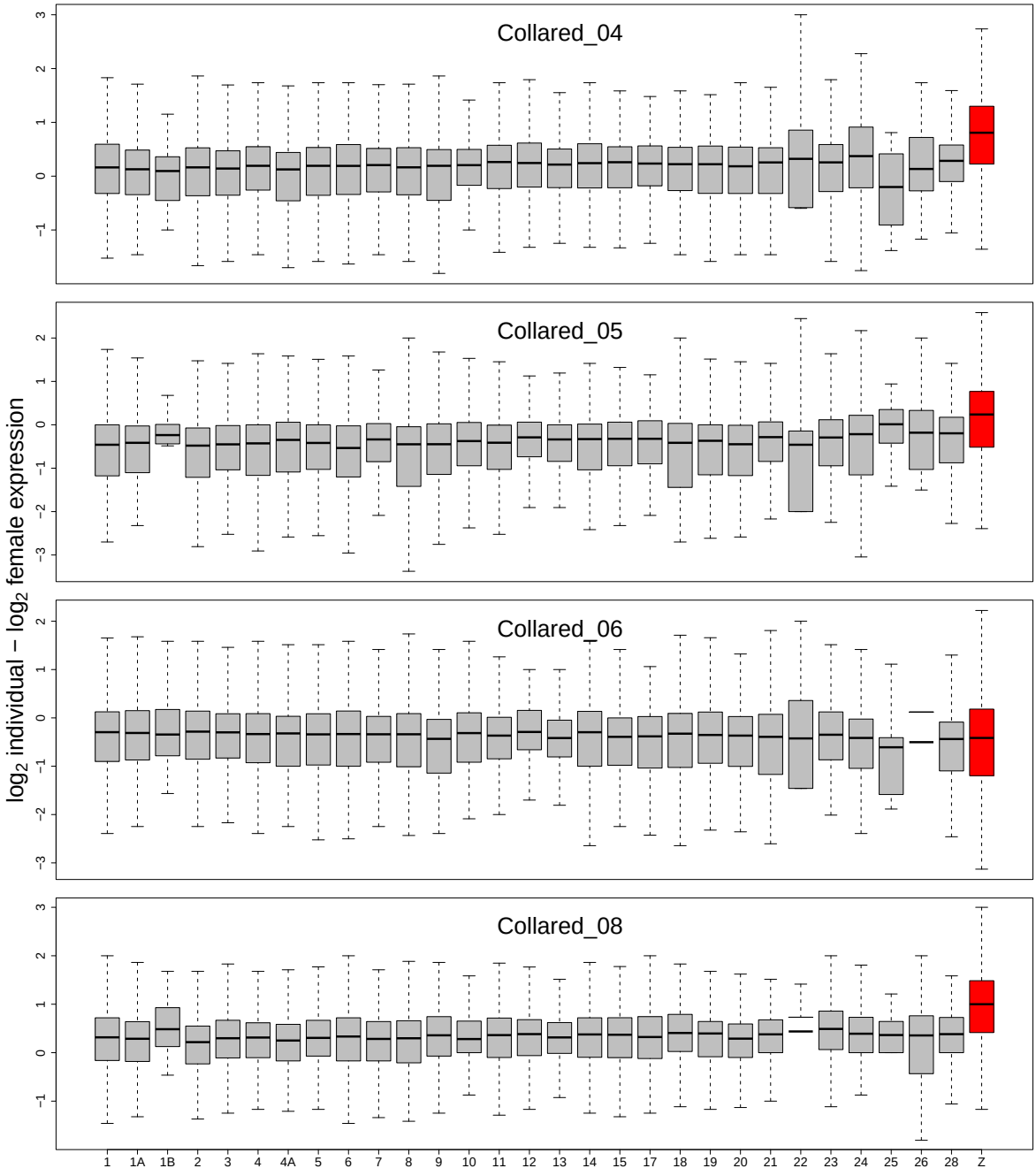


Figure 4.13: **Sexing of collared flycatcher embryo samples.** The $\log_2$ individual-to-female expression ratios are plotted by chromosome. The individuals are the ones to be sexed, females are those known to be female after blasting (i.e. collared_02, collared_03, collared_07, and collared_09). The Z chromosome expression ratio indicates nicely a individual's gender.

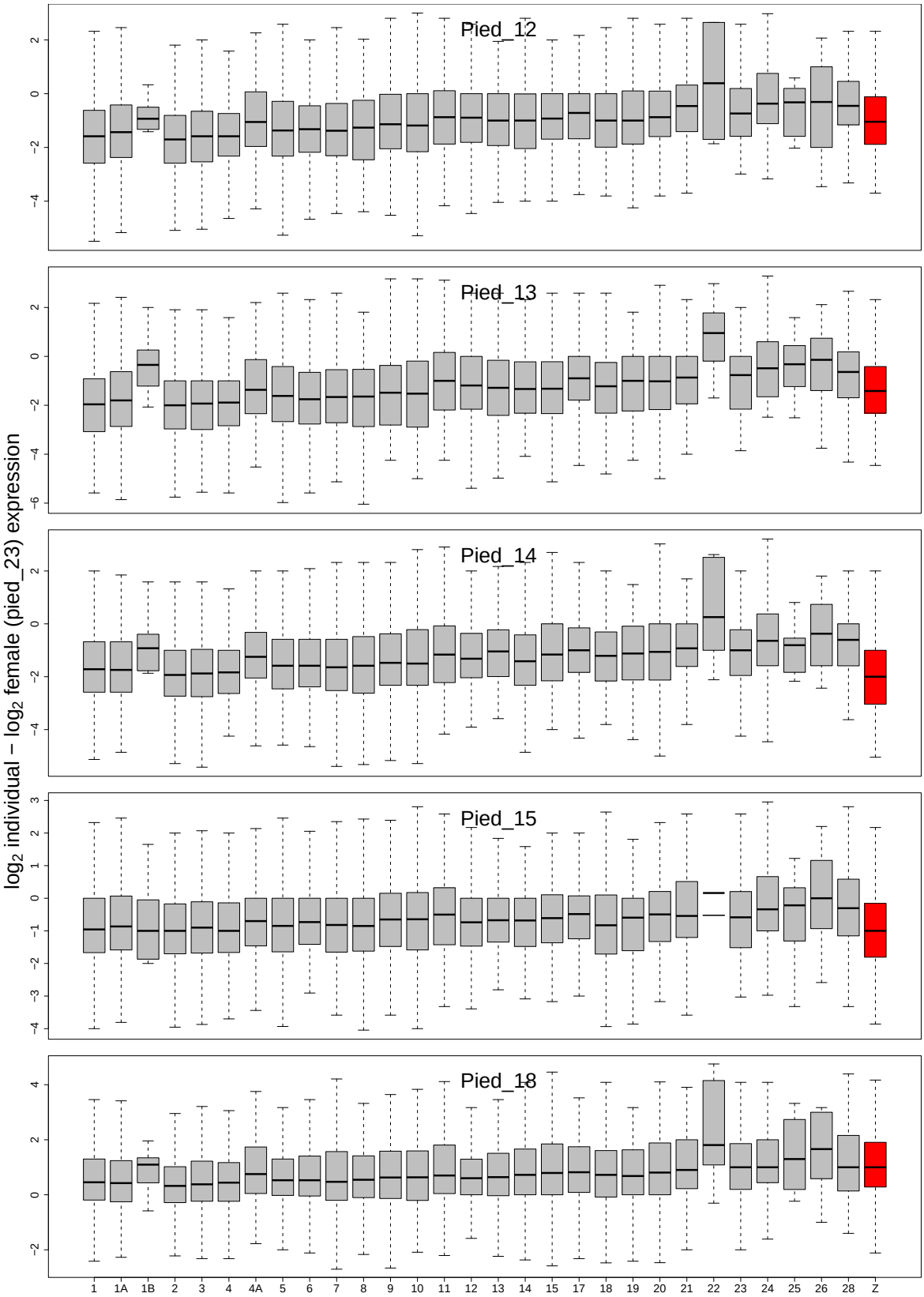


Figure 4.14: **Sexing of pied flycatcher embryo samples.** (continued on next page)

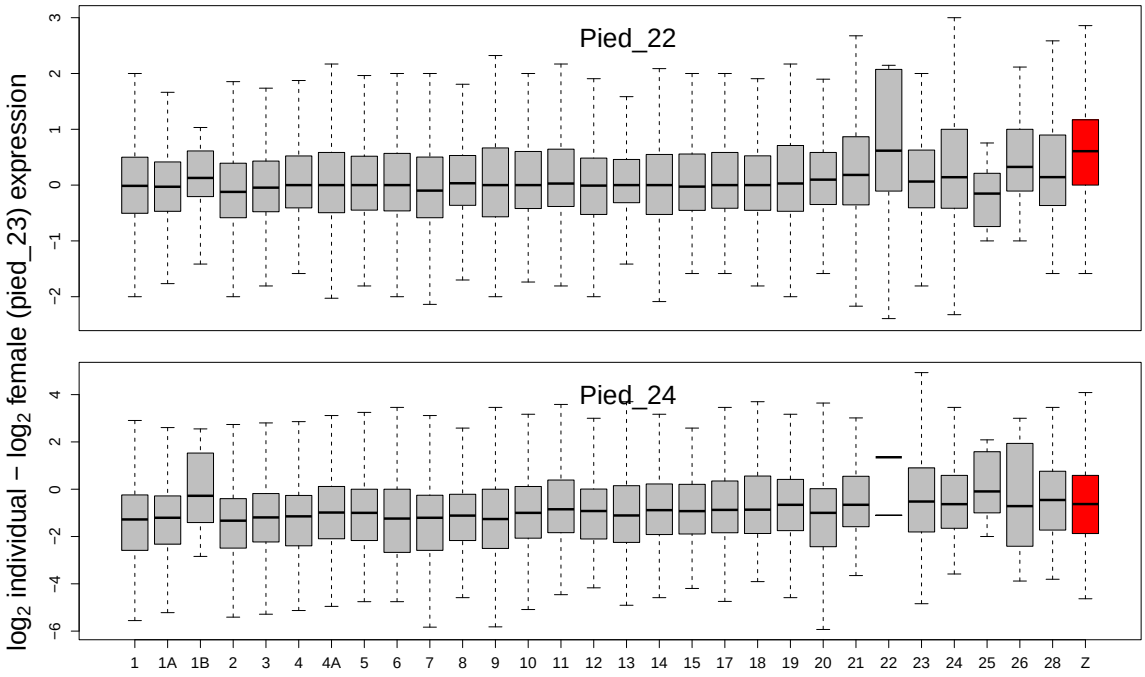


Figure 4.14: **Sexing of pied flycatcher embryo samples.** The $\log_2$ individual-to-female (i.e. pied_23) expression ratios are plotted by chromosome. The different individuals are to be sexed. In contrast to collared flycatchers, the pied flycatcher embryos in our dataset have a lot more variance in chromosomal expression level.

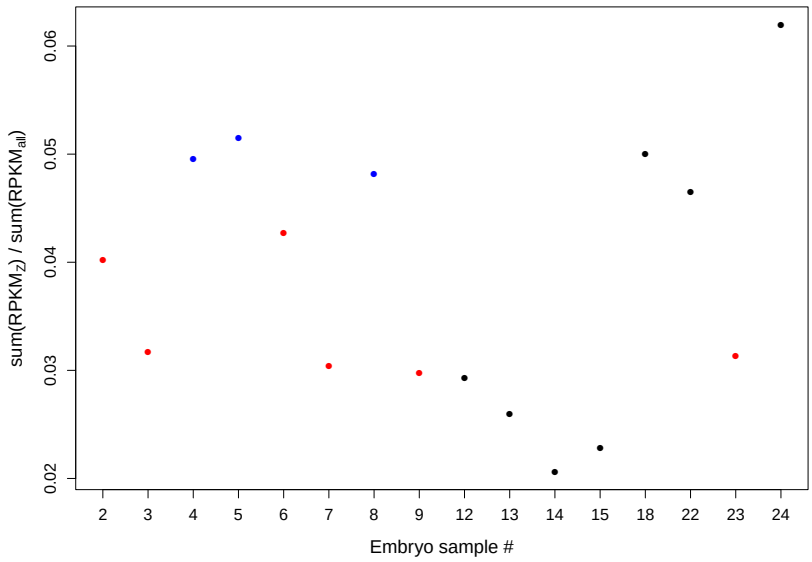
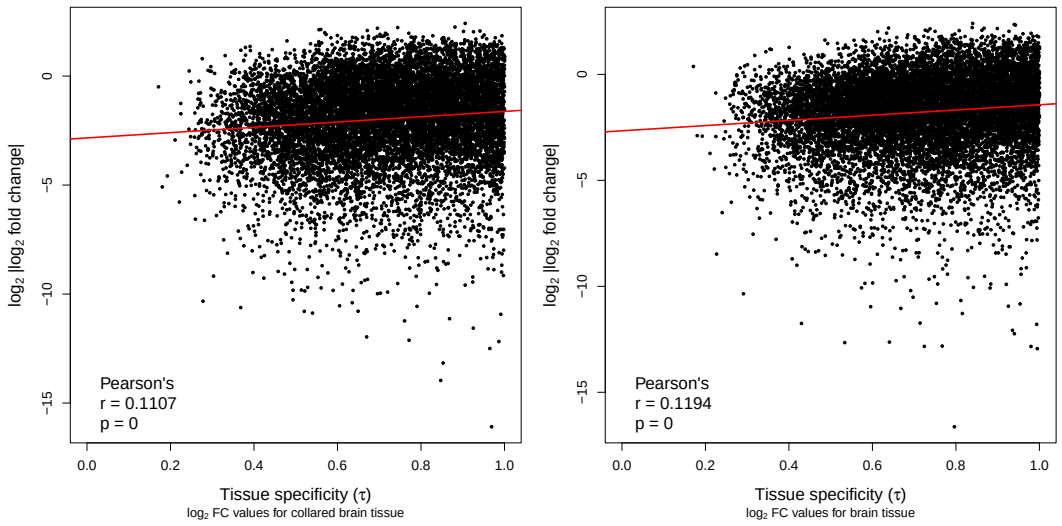


Figure 4.15: **Sexing of flycatcher embryo samples.** Plotted are the sums of all Z-linked RPKM expression values divided by the sum of all (including autosomal) RPKM values of a library. Hitherto sexed individuals are plotted blue for males and red for females.

4.4 Tissue specificity of DE genes

I plotted τ, a measure for tissue specificity, against log fold change values for sex-biased expression in collared brain tissue and found a significant positive correlation ($p < 10^{-16}$) between the strength of DE and tissue specificity (fig. 4.16a). The same was found for DE between species ($p < 10^{-16}$, fig. 4.16b). Analyses were consistent for other tissues and both species (not shown).



(a) (b)

Figure 4.16: **Tissue specificity of DE genes.** The strength of differential gene expression turned out to be significantly correlated with a gene's tissue specificity. **(a)** Sex-biased gene expression in brain tissue of collared flycatchers; $t = 11.9, p < 10^{-16}$. **(b)** DE genes in brain tissue between collared and pied flycatchers; $t = 13.2, p < 10^{-16}$.

5 Discussion

5.1 Testing methods for the detection of differential gene expression

5.1.1 Approaches to test methods for the detection of DE using simulated datasets

My attempts to simulate data with a known fraction of DE genes resulted little satisfying. This was in part due to the fact that very little is known about the actual properties of the density distribution of gene expression and the resulting distribution of Illumina NGS sampled short read counts.

It is commonly, yet tentatively assumed that the underlying distributions are negative binomial (Lu *et al.* 2005, Robinson & Smyth 2007). Further, it has been noted that modelling read counts in order to make inferences about DE in a negative binomial model is quite robust to violations. However, when going the other way and trying to model read count data with a distribution that resembles our real data for the purpose of testing methods for the detection of DE genes, the situation is different. The methods should be tested on simulated data that is as similar to real data as possible. In case there are crucial intrinsic differences between simulated and real data, a comparison between results cannot be assumed to be meaningful.

I tried to simulate data guided by what was known in the literature and according to some properties of a subset of our real data. Comparisons of these with real data consistently showed problems in both the lower and the upper range of expression level. Finally, when I got access to a full dataset of real read count data, I developed means of testing the methods on real data which seemed to work quite well and I followed that approach instead.

5.1.2 Testing methods for the detection of DE on real data

In order to test the performance of the evaluated methods on real read count data, I compared several properties of the genes found to be DE by the different methods including total number, fold change, mean expression, expression variance, gene length, and GC content.

As for the number of genes, the different methods found between 6.3% – 13.7% (`edgeR`) and 16.8% – 33.7% (`DEGseq`) of the genes to be DE in the different tissues. `DESeq` and `edgeR` were rather conservative and found between 6.3% and 15.9%; `baySeq` found 9.9% – 28.3% of the genes to be DE and `DEGseq` 16.8% – 33.7%. Theory predicts biological

read count data to be overdispersed compared to a Poisson or binomial distribution (Robinson & Smyth 2008) and that the additional variance in overdispersed read count data leads to an overestimated amount of DE genes when modelled as Poisson or binomially distributed. An inflated fraction of false positives would thus expected from the results of `DEGseq` as this method does employ a Poisson distribution for modelling the distribution of the data.

For all methods, DE gene numbers were consistently smallest for brain tissue; further, muscle tissue had always a lower number of DE genes. Embryos, on the other hand, had the most genes found for all methods but `edgeR`. If one would rank the tissues according to the numbers of DE genes found, the order would be quite conserved among methods. Finding such consistent patterns puts some trust into the general capability of the methods to find DE genes.

This last point can be studied in further detail by inspecting the overlap of genes found by different methods. I analysed this by creating a venn diagram comparing the four methods (fig. 4.1). The largest number in this plot was that of genes uniquely found by `DEGseq`, showing the method's large amount of genes detected. This could be due to the fact that the method models read count distributions as Poisson distributed, which leads to a high expected amount of false positives due to overdispersion. The second largest number, on the other hand, was the overlap of all four methods showing that there is quite a large a 'core set' of genes with a clear pattern of DE. Other gene numbers in this diagram were smaller, some zero or close to, some around 150 to 225 for categories including the methods `baySeq`, `DESeq`, and `DEGseq`; `edgeR` turned out to be quite conservative.

Given the observation that quite an amount of DE genes were found by some of the methods but not all, the question arises what gene properties make up the differences in detection. In order to answer this question, I first had a look on the magnitude of fold change DE genes actually show, compared to the genome-wide distribution. It appeared that `baySeq` and `edgeR` covered a broader range of different fold change levels, with `edgeR` finding genes with a higher mean fold change than `baySeq`. `DEGseq`, on the other hand, found genes in a rather restricted range of fold change values and at rather small differences in expression in general, although still higher on average than the genome-wide distribution (fig. 4.2). However, it was the only method declaring a substantial amount of genes to be significantly DE that showed no true mean change in expression between groups. I.e., they might to quite some amount consist of false positives. `DESeq` performed somewhat in between, finding genes in a medium broad spectrum of fold change values with a similar distribution pattern as `baySeq`. Reconsidering the conservative behaviour of `edgeR` and `DESeq` compared to the other two methods, it makes an intuitive sense that the distribution width was relatively narrow for `DESeq` and broad for `baySeq`, but not that it was narrow for `DEGseq` and broad for `edgeR`. The narrow distribution of genes found by `DEGseq` might indicate that the capability to detect differential gene expression is restricted by some means for this method.

Second, I compared the distributions of the detected genes' expression means with the genome-wide distribution (fig. 4.3a). Here, it was apparent that lowly expressed genes were rarely detected as DE. This is obvious, because small read sample sizes due

to low expression lead to reduced statistical power. However, DEGseq and DESeq were more affected by this problem than the other methods and found only genes in a small, confined region of highly expressed genes to be significantly DE. baySeq, at the other end of the spectrum, covered a broad range of different mean expression levels. The genes detected by edgeR fell in between as they lacked the very lowly expressed range, but were not particularly skewed for high expression as genes found by DEGseq and DESeq.

Furthermore, I looked at the distribution of expression variance of detected genes, again compared to the genome-wide average (fig. 4.3b). I used variance values only for collared flycatchers in order to avoid the influence of additional variance from the difference in expression itself. In this case, one would expect DE genes to have, on average, a similar variance distribution as the whole genome dataset. All methods showed a bias for higher variance values, yet some more than others. DEGseq and DESeq showed a rather pronounced distribution at high variance values, whereas genes found by baySeq and edgeR were only slightly dragged into that direction. For edgeR, this resulted in a smaller distribution range with less detected genes at very small variance levels, but also no strong skew for high variance levels. baySeq, on the other hand, picked up genes in a broader range of variance levels including genes with rather low variances, but still not too much skew towards high variances. I would argue that this criterion is rather crucial because it appears sound to me that one could mistake high variance for patterns of actual DE. A good modelling of the read count variance distribution – a problem which has been stated to be rather complicated and not yet been solved to satisfaction (see e.g. Robinson & Smyth 2008) – is thus what is needed to perform well in this perspective.

Finally, I compared the detected genes' distributions of gene length and GC content with the genome-wide distribution in order to find potential biases for these properties (fig. 4.4). For both comparisons, all methods were very close to the genome-wide distribution. However, except for the gene length distribution of genes found by DEGseq, the distributions were always significantly different from the distribution in the whole dataset. Speaking of gene length, distributions were slightly skewed to genes with shorter lengths (fig. 4.4a). This could be expected as it has been observed before that shorter genes have on average higher expression levels (Urrutia & Hurst 2003) and, as mentioned before, higher gene expression leads to larger sample sizes and greater statistical significance. In the comparison of gene GC content distributions, detected genes showed on average slightly higher GC contents which became especially prominent as this led to the increase of a lower, second peak in the bimodal GC content distribution (fig. 4.4b). Although I cannot provide a readily apparent explanation for the bimodality of the GC content distribution, there is a possible explanation for the fact that DE genes tend to have higher GC contents than the genomic average. As previously shown (Jjingo *et al.* 2011), I found a negative correlation between gene GC content and gene length (fig. 4.5). A bias for shorter genes could therefore in turn lead to a bias for higher GC content.

To sum up these comparisons, one can say that DEGseq and DESeq showed some biases for gene properties – other than pure different expression level – more pronounced than the other two methods. Overall, it seems to me that baySeq performed slightly better than the other methods, in terms that it was comparatively little biased but still

appeared to be sensitive in a broad range to all investigated gene properties. I showed the results for analyses of differential expression between the species and for brain tissue. However, I retrieved the same results when testing sex-biased expression between sexes of the same species in all different tissues (not shown).

5.1.3 Properties of differential gene expression detection by `baySeq`

Having decided to use `baySeq` as the method of choice for the analysis of DE, I wanted to get more insights into how `baySeq` estimates DE, i.e. what gene properties (i.e. mainly expression level and variance) influence the detection of DE. E.g., as seen before, genes with higher mean expression level were more likely to be detected as DE (given they show a difference in expression between groups) than genes that were less abundantly expressed, most likely due to statistical power reasons. First, I created volcano plots, where I plotted `baySeq` posterior probabilities against $\log_2$ fold change values in order to see if posterior probabilities come along with true differences in expression level between groups. Figure 4.6 shows this comparison for all tissues tested comparing the two species with each other. All tissues showed a clear pattern with only very few genes in an area of no difference in expression level (i.e. fold change is 0) and a significant `baySeq` posterior probability for DE (i.e. above the cutoff value of 0.3). However, there were some genes in that area for embryo, kidney, and skin tissues. As one can see in figures 4.7b and S.2, these genes usually showed high variances in expression (see discussion below). One further observation was that some plots were rather symmetrical with equal amounts of genes biased towards each species, while other were not.

Colouring single genes in these volcano plots can help by adding additional information on the placement of single genes on given points in the plotting area. Specifically, I used a colour range from light yellow to dark red in order to show low to high mean expression levels (fig. 4.7a) or low to high expression variances (fig. 4.7b); the first plots visualise the influence of mean expression level on the capability of `baySeq` in the detection of DE, the second the influence of expression variance on the detection of DE.

It appeared that genes with high mean expression level were rather clearly assigned to either being not DE, i.e. they had a posterior probability of very close to 0; or they were assigned to being DE, having high posterior probabilities and not necessarily big, but substantial fold changes (fig. 4.7a). Lowly expressed genes, however, needed to show more pronounced changes in expression between groups in order to being assigned posterior probabilities above the cutoff and they hence tended to be located below the cutoff. This indicates that sample size, in terms of gene expression level, is a factor influencing the ability of `baySeq` to detect DE, as one would expect. Furthermore, it shows that quite a substantial amount of genes is expressed so lowly that there are statistical problems in detecting their status of expression difference between groups.

Furthermore, genes situated in the plot area of no true mean fold change but posterior probabilities above the cutoff always showed high expression variances (fig. 4.7b) and rather high mean expression values (fig. 4.7a). So if a gene has a high expression mean

and a large variance, these patterns can possibly be mistaken for patterns of true DE. This shows that high variances comprise an issue that can confound patterns of DE. On the other hand, only a few tissues showed genes localised in the area of no fold change while having high posterior probabilities. In other words, only for some of our tissues, this really creates some sort of problem, whereas other tissues seemed to be less affected by differences in expression variance. Genes with lower variances were found enriched on both sides of the area of no fold change, very well colocalising with highly expressed, DE genes. The majority of genes with low variances, however, showed little patterns of DE and were thus also assigned low posterior probabilities.

There might be some reasons for why three of the tissues, embryo, kidney, and skin, had some genes placed in the area of no true DE but high posterior probabilities. As can be seen in figure 4.7b, high expression variances are the main reason for placement in this plotting area. In the case of the embryos, a good guess for why there are such high variances could be that the time of sampling was different for the individual eggs as they were collected from wild populations and exact information on age and incubation time were relatively bad known. Further, as expression patterns can change rather quickly during embryonic development (Mank & Ellegren 2009), one might expect having sampled different stages of embryonic development in the different eggs what would then lead to high expression variances which in turn caused problems in detecting patterns of DE.

For kidney and skin tissues, there is no such readily apparent reason why something similar should be true in their cases. We had to repeat some pied female skin samples because their runs produced very few reads. However, getting new, finally successful sequencing results did not change that picture. It is, of course, possible that both the observed high variances as well as the high posterior probability of DE are genuine features of the respective genes. In that case, `baySeq` would decide for a pattern of DE because differences in expression suggest a pattern of DE in all but a few libraries and these exceptional libraries only would deviate so much that the mean expression difference between groups seemed balanced by chance. However, this explanation seems rather unlikely to me since this issue only appeared in a few libraries. As long as there is no other apparent reason for this behaviour I recommend treating these tissues with care.

A similar reasoning may be valid for the asymmetrical appearance of the volcano plots in gonad tissues. Catching and sampling birds from the two species in different reproductive states could lead to differences in expression levels. As the two species do not have the same timing in breeding behaviour but only overlap slightly (Qvarnström *et al.* 2009), it is difficult to sample specimens at the same time point in their reproductive cycle and it has probably not been achieved exactly. However, this explanation does not hold as well for the other skewed tissues.

Rather, skewed expression patterns as observed in my volcano plots might depict biases during sample normalisation. The samples have been normalised according to library size. RPKM normalisation cannot improve the picture as the same genes are compared with each other and therefore there cannot be a bias due to different gene lengths. One could try trimmed fold change normalisation procedures as suggested by Robinson &

Oshlack (2010), which deals exactly with that property. It appears sound that too large skews towards one sample group might display normalisation problems. Eliminating all sample variation in the total amount of differential expression, on the other hand, might mean to manipulate the data too heavily; a skewed fold change pattern could after all depict both, natural patterns or sample normalisation problems. The fact that the bottom of the volcano plot distributions, i.e. the genes below a posterior probability of approximately 0.3, are centred at zero for most cases (also for fold change-skewed plots) might question the need for fold change normalisation.

A feature that is unique to `baySeq` compared to the other methods is that more complex models than simple two-group comparisons can be done. In order to test this feature, I ran `baySeq` with seven different models, including no DE between groups, DE between species but not sexes, between sexes but not species, and between each of the sexes in each of the species against all respective other three groups (e.g. male pied flycatchers against all other groups). The results are shown in table 4.2. Most genes were found for the comparison between species, less for the other contrasts, although some models for some tissues sometimes showed somewhat higher values. One cannot really infer a lot about the reliability of the results because there is nothing appropriate to compare the results with. However, focusing on the species comparison, the general trend from the two-group comparison between the species persists, with brain tissue having fewer genes detected and kidney and liver tissues more. Furthermore, what one could expect from such a comparison, is that, compared to two-group comparisons, there are less genes found in general. This is because in a two-group comparison strong contrasts between one of the four groups compared to the other ones would confound comparisons between two of the groups taken together (e.g. when contrasting the sexes and thereby merging the species). This becomes apparent when comparing the relatively few sex-biased genes in this analysis with the analysis of sex-biased expression in a single two-group comparison. In biological terms speaking, there are a few sex-biased genes that are biased in both species. But there are substantially more sex-biased genes within one species that do not show the same sex-bias pattern in the other species. It seems to be the case that the species have diverged to some extent in terms of sex-biased expression patterns.

It can be shown that higher replication leads to more genes found to be DE (fig. 4.8). This is expected, as more replicates can help estimating a gene-wise variance level and lead to more stable estimates of sample mean which helps discriminating between distinct groups. However, as one can see, increase in statistical power does not level off when reaching our sampling level of 10 biological replicates for each species. Even more samples would therefore mean more genes found to be DE. And so far there is no indication if a saturation level could be reached soon. Given that the numbers of DE genes found are already at 10% to 30% using our method, this seems quite astonishing.

As different sets of libraries returned different results using this Bayesian method, I performed a jackknife analysis in order to investigate how much the number of detected genes varies when analysing with different libraries. Out of the whole dataset of 10 pied and 10 collared flycatchers, I sampled randomly 100 times five samples each and used `baySeq` to estimate the amount of DE genes. The results varied quite substantially and

the distribution of the detected genes turned out to be bimodal. My first guess was that this bimodality was due to uneven gender distributions in the sampling – because of the 10 samples from each species, there were each five males and females and sampling could therefore by chance create uneven comparisons. But testing for this did not give any significant results, although the expected trend was observed. Larger sample sizes might help with this issue, but because `baySeq` is very computing intensive, I did not try that.

5.2 Dosage compensation

Plotting $\log_2$ male-to-female ratios for each chromosome revealed that for the Z chromosome – in contrast to the autosomes, which were balanced – expression levels were higher in males than in females (fig. 4.10 and 4.11). These results were consistent for all tissues compared (excluding gonads, because ovary and testis are not homologous and thus not comparable). I.e., male and female flycatchers do have different expression levels for the same set of Z-linked genes, apparently without experiencing selection for compensating for the difference. This is in contrast to what theory predicts and what has been observed in mammals, flies, and nematodes (Mank 2009). The case in lepidopterans is still being discussed (see Walters & Hardcastle 2011). On the other hand, it corroborates results from previous studies in birds (Ellegren *et al.* 2007, Itoh *et al.* 2007, Wolf & Bryk 2011). Furthermore, Illumina NGS technique as used here is better suited for genome-wide quantitative transcriptomics than 454 sequencing (as used by Wolf & Bryk 2011) or microarrays (as used by the other two mentioned studies, as well as by Walters & Hardcastle 2011). So this study provides further confirmation to the previous results using a technique that is likely to be even more powerful for this type of analysis.

I further compared autosomal and Z-linked expression level in order to check for dosage imbalance between autosomes and the Z chromosome (fig. 4.12). The results confirmed the previous analysis in that female Z expression was substantially below autosomal expression, whereas male Z expression was at the same level as autosomal expression. However, this does not necessarily hold for gender-specific tissues. Sex-specific tissues provide the rare possibility of relaxed released conflict. Global upregulation of Z-linked genes in ovary tissue would not have any implication for the expression in males and vice versa. However, in ovaries, the pattern was the same as before; Z-linked expression level was reduced compared to autosomal expression. In testes, in contrast, male Z chromosome expression was significantly reduced. Instead of the possible dosage compensation in the female sex-specific tissue, I found a dosage imbalance in the male-specific tissue. It seems that the released sexual conflict in this tissue led to an unequal expression ratio between Z chromosome and autosomes in testis tissues contrary to what one might expect.

Z chromosome expression in brain tissue of collared flycatchers was also significantly below that of autosomal expression. However, as the p-value was only slightly significant, I would handle this result with care and treat it rather as a false positive until it is

confirmed elsewhere.

The finding of unequal global gene doses between male and female birds provokes further questions. Are all genes affected uniformly? If not, which ones are? Are all genes expressed towards a male-bias or are there exceptions? What fraction of male-biased genes is Z-linked? In order to address these questions, I looked for patterns of sex-biased expression using `baySeq` (tab. 4.3). A fraction of 1.7% to 8.9% of genes were detected to be sex-biased; furthermore, there was a clear outlier (skin tissue in pied flycatchers) with 16.3% of the genes assigned to be sex-biased. This might be due to the bad quality of some of the skin libraries of female pied flycatchers. These libraries have been sequenced again, but some problems seem still to persist.

Besides that, there are some patterns nicely apparent, i.e., first, that sex-biased genes were significantly enriched for Z-linked genes. Second, Z-linked genes were significantly more male-biased than expected, even when controlling for the general dosage imbalance between males and females. There was no such a clear pattern for autosomal, sex-biased genes: Although some autosomal genes were significantly biased towards either male or female expression, there was no clear and immediately visible pattern apparent.

5.3 Determining the sex of the embryo samples

I blasted the chicken sequences of two W-linked genes, *HINTW* and *CHDW*, against the short read libraries of our embryo data. This resulted in hits for single reads or single reads pairs in one pied and four collared flycatcher libraries. The results were highly significant and were confirmed by blasting the reads that produced alignments to Entrez's nr/nt dataset. Thus, I can say with confidence that the samples that produced the hits were female.

However, since hits were very few and as producing no hit against the female gene fragments does not necessarily mean that an individual is male, I used two different plots for estimating male-to-female Z ratios. Males would be expected to have a higher male-to-female ratio for Z-linked genes as they harbour two Z chromosomes instead of only one in females. And indeed, figure 4.13 showed nicely that collared_06 was female, although it did not produce a hit in the `blast` search; the other three collared samples that did not produce hits were clearly identifiable as males.

Due to higher variances in their chromosome-wise expression patterns, the genders were not as apparent for pied flycatchers. However, using an additional plot, I am quite confident now to have identified pied_18, pied_22, and pied_24 as males, the other individuals as females. I used these classifications later for analyses of sex-biased expression patterns (see section 5.2).

5.4 Tissue specificity of DE genes

I confirmed previous findings (Mank *et al.* 2008) that sex-biased genes are on average more tissue specific than equally expressed genes. Furthermore, I showed that the same

is true for genes that show differences in expression between collared and pied flycatchers. This makes sense as less ubiquitously expressed genes have less constraints to change; a pattern which is true for changes both between sexes and between young sister species. In turn, this means that the most pronounced differences between sexes and between young diverging species must be expected to be confined to single tissues. In other words, when screening for candidate genes that show different expression patterns between recently diverging species, it is a good advice to include as many tissues in the analysis as possible.

5.5 Conclusions

5.5.1 NGS and differential gene expression

I tested four different methods for their performance in detecting patterns of DE and evaluated their results in a qualitative matter, leading to quality estimates that are useful for biologists working with these methods. The methods were tested for potential confounding biases in detecting DE genes and compared them relative to each other. My results suggested that one method, `baySeq`, performed better than the other methods in basically all aspects tested as it showed the least biased results. Furthermore, it is more flexible as more complex models than two-group comparisons can be performed. There are, however, still some crucial points to consider; some genes are declared to be DE while lacking a true difference in expression between groups, the issue of normalisation is not yet cleared to satisfaction, and the level of replication as present in this study does not seem to be sufficient to find a complete set of DE genes.

5.5.2 On species and sexes

When investigating transcriptome patterns in the collared flycatcher / pied flycatcher model system, I focused on two main topics; differential expression between the two sister species and sex-biased expression.

I have shown that there were indeed expression differences between species, at a level higher than that between sexes of the same species, showing that the species have diverged in expression patterns. Different tissues showed different expression patterns for different genes and at different levels. The number of DE genes are always at levels of several hundred, and these may well include false positives of unknown number. So in order to find genes that are involved in speciation and divergence between the species, further studies will be necessary. One approach could be to plot DE genes on the genome sequence in order to find local cassettes of genes that are DE between species and could in concordance play a role in speciation.

To conclude, this study may very well provide a valuable entrance point into further studies of this ecological model and I hope that it will provide a great deal of help for future transcriptome analyses in our lab.

6 Acknowledgements

I like to close by thanking all the people that helped me with my work on this project.

First of all, I thank Hans Ellegren for accepting me as a master student and providing me with this interesting thesis project. I am glad for being part of the flycatcher genome sequencing efforts. I further thank him for his project supervision and guidance in a difficult new area of research.

I would like to express my sincere thanks to Axel Künstner for his hand-on supervision and fruitful discussions on all conceivable topics connected to my project work. Thank you very much for proofreading this thesis!

I like to thank all the people involved in the project: Hannu Mäkinen prepared all the libraries for sequencing. Linnéa Smeds and Axel preprocessed the reads by trimming and quality-filtering. Linnéa, Axel, and Páll Ólason assembled the genome and Páll created the flycatcher-specific zebra finch gene models. Axel summarised the single reads to gene count numbers and provided me with the data.

Further, I thank all people in Hans' lab for the many fruitful discussions that keep broadening my knowledge in this field of biology and beyond. It is a great pleasure to work with you and I always appreciate the ample opportunities for discussions.

Last but not least I thank my parents for all their patience and support during my stay in Uppsala. Thank you for helping me on my way so far and with my decision for continuing my studies as a PhD student.

7 References

Alatalo RV, Carlson A, Lundberg A, & Ulfstrand S (**1981**) The conflict between male polygamy and female monogamy: the case of the pied flycatcher *Ficedula hypoleuca*. *American Naturalist* **117**:738-753.

Alatalo RV, Gustafsson L, & Lundberg A (**1982**) Hybridization and breeding success of collared and pied flycatchers on the island of Gotland. *The Auk* **99**:285-291.

Anders S & Huber W (**2010**) Differential expression analysis for sequence count data. *Genome Biology* **11**:R106.

Backström N, Lindell J, Zhang Y, Palkopoulou E, Qvarnström A, Sætre G-P, & Ellegren H (**2010**) A high-density scan of the Z chromosome in *Ficedula* flycatchers reveals candidate loci for diversifying selection. *Evolution* **64**:3461-3475.

Baggerly KA, Deng L, Morris JS, & Aldaz CM (**2003**) Differential expression in SAGE: accounting for normal between-library variation. *Bioinformatics* **19**:1477-1483.

Barker FK, Cibois A, Schikler P, Feinstein J, & Cracraft J (**2004**) Phylogeny and diversification of the largest avian radiation. *Proceedings of the National Academy of Sciences of the USA* **101**:11040-11045.

Benjamini Y & Hochberg Y (**1995**) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* **57**:289-300.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR *et al.* (**2008**) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**:53-59.

Beutler E, Yeh M, & Fairbanks VF (**1962**) The normal human female as a mosaic of X-chromosome activity: studies using the gene for G-6-PD-deficiency as a marker. *Proceedings of the National Academy of Sciences of the USA* **48**:9-16.

Charlesworth B (**1978**) Model for evolution of Y chromosomes and dosage compensation. *Proceedings of the National Academy of Sciences of the USA* **75**:5618-5622.

Cracraft J & Barker FK (**2009**) Passerine birds (Passeriformes). *In:* Hedges SB & Kumar S (eds.) The Timetree of Life, pp. 423-431. Oxford University Press, Oxford, UK.

Ellegren H (**2011**) Sex-chromosome evolution: recent progress and the influence of male and female heterogamety. *Nature Reviews Genetics* **12**:157-166.

Ellegren H, Hultin-Rosenberg L, Brunström B, Dencker L, Kultima K, & Scholz B (**2007**) Faced with inequality: chicken do not have a general dosage compensation of sex-linked genes. *BMC Biology* **5**:40.

Enard W, Khaitovich P, Klose J, Zöllner S, Heissig F, Giavalisco P, Nieselt-Struwe K, Muchmore E, Varki A, Ravid R, Doxiadis GM, Bontrop RE, & Pääbo S (**2002**) Intra- and interspecific variation in primate gene expression patterns. *Science* **296**:340-343.

Hamada FN, Park PJ, Gordadze PR, & Kuroda MI (**2005**) Global regulation of X chromosomal genes by the MSL complex in *Drosophila melanogaster*. *Genes & Development* **19**:2289-2294.

Hardcastle TJ & Kelly KA (**2010**) `baySeq`: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* **11**:422.

Hoekstra HE & Coyne JA (**2007**) The locus of evolution: evo devo and the genetics of adaptation. *Evolution* **61**:995-1016.

Itoh Y, Melamed E, Yang X, Kampf K, Wang S, Yehya N, Van Nas A, Replogle K, Band MR, Clayton DF, Schadt EE, Lusis AJ, & Arnold AP (**2007**) Dosage compensation is less effective in birds than in mammals. *Journal of Biology* **6**:2.

Jjingo D, Huda A, Gundapuneni M, Mariño-Ramírez L, & King Jordan I (**2011**) Effect of the transposable element environment of human genes on gene length and expression. *Genome Biology and Evolution* **3**:259-271.

Leeb M & Wutz A (**2010**) Mechanistic concepts in X inactivation underlying dosage compensation in mammals. *Heredity* **105**:64-70.

Li H & Durbin R (**2009**) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* **25**:1754-1760.

Lu J, Tomfohr JK, & Kepler TB (**2005**) Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach. *BMC Bioinformatics* **6**:165.

Lyon MF (**1961**) Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature* **190**:372-373.

Lundberg A & Alatalo RV (**1992**) The pied flycatcher. Poyser, London, UK.

Mank JE (**2009**) The W, X, Y and Z of sex-chromosome dosage compensation. *Trends in Genetics* **25**:226-233.

Mank JE & Ellegren H (**2009**) All dosage compensation is local: gene-by-gene regulation of sex-biased expression on the chicken Z chromosome. *Heredity* **102**:312-320.

Mank JE, Hultin-Rosenberg L, Zwahlen M, & Ellegren H (**2008**) Pleiotropic constraint hampers the resolution of sexual antagonism in vertebrate gene expression. *American Naturalist* **171**:35-43.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z *et al.* (**2005**) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**:376-380.

Meyer BJ & Casson LP (**1986**) *Caenorhabditis elegans* compensates for the difference in X chromosome dosage between the sexes by regulating transcript levels. *Cell* **47**:871-881.

Mortazavi A, Williams BA, McCue K, Schaeffer L, & Wold B (**2008**) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**:621-628.

Muller HJ (**1932**) Further studies on the nature and causes of gene mutations. In *Proceedings of the 6th International Congress of Genetics*, Vol. 1 (ed. DF Jones), pp. 213-215. Banta, Menasha, WI.

Muller HJ (**1950**) Evidence of the precision of genetic adaptation. *Harvey Lectures* **43**:165-229.

Ohno S, Kaplan WD, & Kinosita R (**1959**) Formation of the sex chromatin by a single X-chromosome in liver cells of *Rattus norvegicus*. *Experimental Cell Research* **18**:415-418.

Oshlack A, Robinson MD, & Young MD (**2010**) From RNA-seq reads to differential expression results. *Genome Biology* **11**:220.

Qvarnström A, Rice AM, & Ellegren H (**2010**) Speciation in *Ficedula* flycatchers. *Philosophical Transactions of the Royal Society B* **365**:1841-1852.

Qvarnström A, Wiley C, Svedin N, & Vallin N (**2009**) Life-history divergence facilitates regional coexistence of competing *Ficedula* flycatchers. *Ecology* **90**:1948-1957.

R Development Core Team (**2009**) `R`: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Robinson MD, McCarthy DJ, & Smyth GK (**2010**) `edgeR`: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**:139-140.

Robinson MD & Oshlack A (**2010**) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* **11**:R25.

Robinson MD & Smyth GK (**2007**) Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **23**:2881-2887.

Robinson MD & Smyth GK (**2008**) Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* **9**:321-332.

Storey JD & Tibshirani R (**2003**) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the USA* **100**:9440-9445.

Straub T & Becker PB (**2007**) Dosage compensation: the beginning and end of generalization. *Nature Reviews Genetics* **8**:47-57.

Svedin N, Wiley C, Veen T, Gustafsson L, & Qvarnström A (**2008**) Natural and sexual selection against hybrid flycatchers. *Proceedings of the Royal Society B* **275**:735-744.

Urrutia AO & Hurst LD (**2003**) The signature of selection mediated by expression on human genes. *Genome Research* **13**:2260-2264.

Vicoso B & Bachtrog D (**2011**) Lack of global dosage compensation in *Schistosoma mansoni*, a female-heterogametic parasite. *Genome Biology and Evolution* **3**:230-235.

Walters JR & Hardcastle TJ (**2011**) Getting a full dose? Reconsidering sex chromosome dosage compensation in the silkworm, *Bombyx mori*. *Genome Biology and Evolution*. first published online, doi: `10.1093/gbe/evr036`.

Wang L, Feng Z, Wang X, Wang X, & Zhang X (**2010**a) `DEGseq`: an `R` package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* **26**:136-138.

Wang Z, Dong D, Ru B, Young RL, Han N, Guo T, Zhang S (**2010**b) Digital gene expression tag profiling of bat digits provides robust candidates contributing to wing formation. *BMC Genomics* **11**:619.

Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Künstner A, Searle S, White S, Vilella AJ, Fairley S *et al.* (**2010**) The genome of a songbird. *Nature* **464**:757-762.

Wilson AC, Maxson LR, & Sarich VM (**1974**) Two types of molecular evolution. evidence from studies of interspecific hybridization. *Proceedings of the National Academy of Sciences of the USA* **71**:2843-2847.

Wolf JBW, Bayer T, Haubold B, Schilhabel M, Rosenstiel P, & Tautz D (**2010**) Nucleotide divergence vs. gene expression differentiation: comparative transcriptome sequencing in natural isolates from the carrion crow and its hybrid zone with the hooded crow. *Molecular Ecology* **19s**:162-175.

Wolf JBW & Bryk J (**2011**) General lack of global dosage compensation in ZZ/ZW systems? Broadening the perspective with RNA-seq. *BMC Genomics* **12**:91.

Xiong Y, Chen X, Chen Z, Wang X, Shi S, Wang X, Zhang J, & He X (**2010**) RNA sequencing shows no dosage compensation of the active X-chromosome. *Nature Genetics* **42**:1043-1047.

Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E, Lancet D, & Shmueli O (**2004**) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**:650-659.

Zha X, Xia Q, Duan J, Wang C, He N, Xiang Z (**2009**) Dosage analysis of Z chromosome genes using microarray in silkworm, *Bombyx mori*. *Insect Biochemistry and Molecular Biology* **39**:315-321.

Zhang Z, Schwartz S, Wagner L, & Miller W (**2000**) A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology* **7**:203-214.

S Supplement

Figure legends

Figure S.1: `baySeq` posterior probabilities of DE against $\log_2$ fold change in expression with colour indicating expression level. Genes with positive values are expressed higher in pied flycatchers, negative values indicate higher expression in collared flycatchers; a $\log_2$ fold change of 0 means there is no DE between groups. `baySeq` posterior probabilities for DE range from 0 (low probability of DE) to 1 (high probability of DE). All genes were categorised into deciles (10-quantiles) according to their mean expression over all libraries in both groups. Deciles were assigned different colours according to their expression levels, dark red for high to light yellow for low expression. Mean expression values ranged approximately from 10^{-2} to 10^4. The cutoff value of 0.3 posterior probability for DE is shown by a horizontal red line. The comparisons are for 10 collared vs. 10 pied flycatchers except for gonads (five vs. five) and embryo (eight vs. eight).

Figure S.2: `baySeq` posterior probabilities of DE against $\log_2$ fold change in expression with colour indicating expression variance. Genes with positive values are expressed higher in pied flycatchers, negative values indicate higher expression in collared flycatchers; a $\log_2$ fold change of 0 means there is no DE between groups. `baySeq` posterior probabilities for DE range from 0 (low probability of DE) to 1 (high probability of DE). All genes were categorised into deciles (10-quantiles) according to their mean expression variance within the two groups. Deciles were assigned different colours according to their expression variances, dark red for high to light yellow for low variance. Mean variance values ranged from below 10^{-2} to about 10^2. The cutoff value of 0.3 posterior probability for DE is shown by a horizontal red line. The comparisons are for 10 collared vs. 10 pied flycatchers except for gonads (five vs. five) and embryo (eight vs. eight).

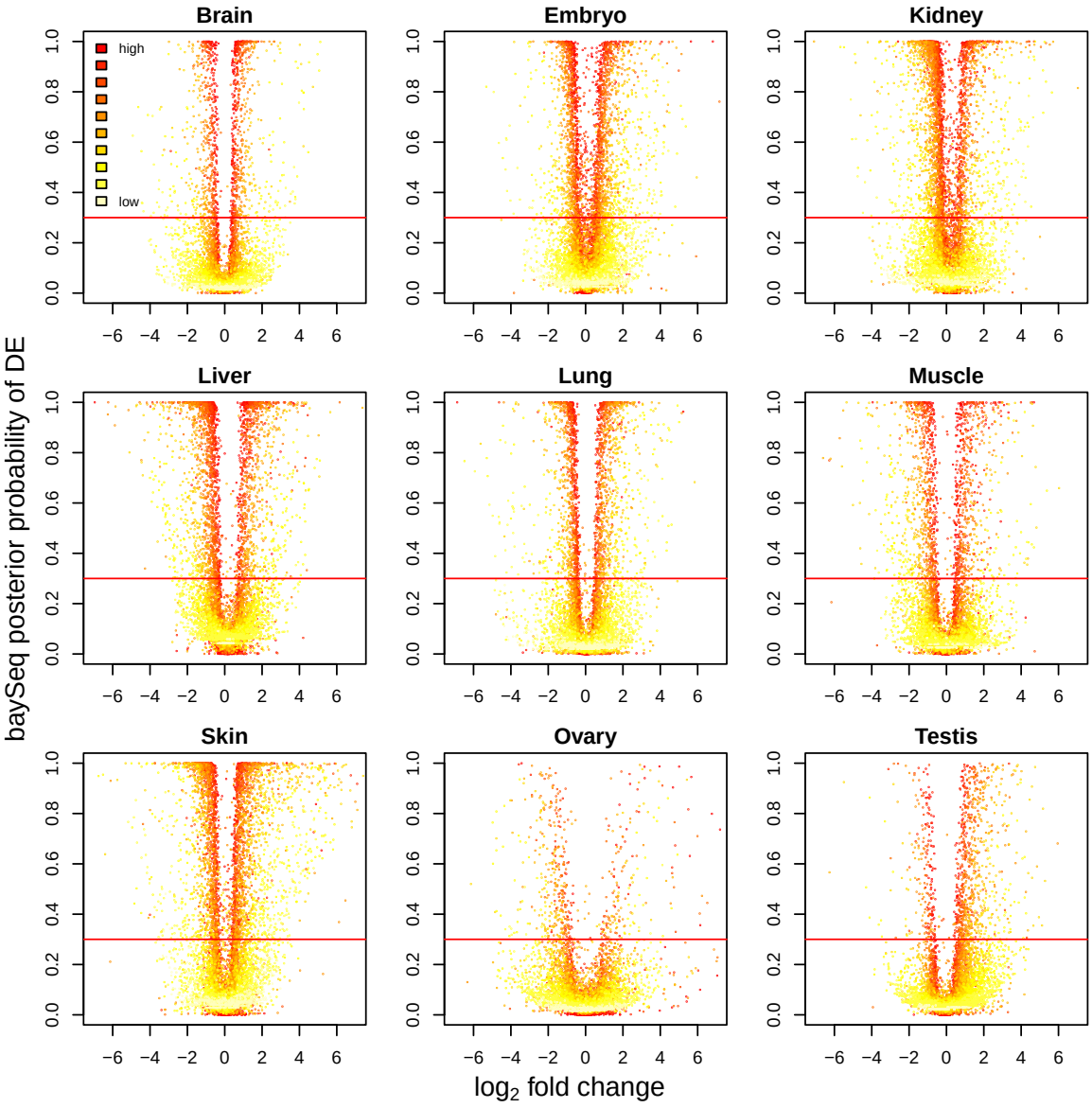


Figure S.1: `baySeq` **posterior probabilities of DE against $\log_2$ fold change in expression with colour indicating expression level.**

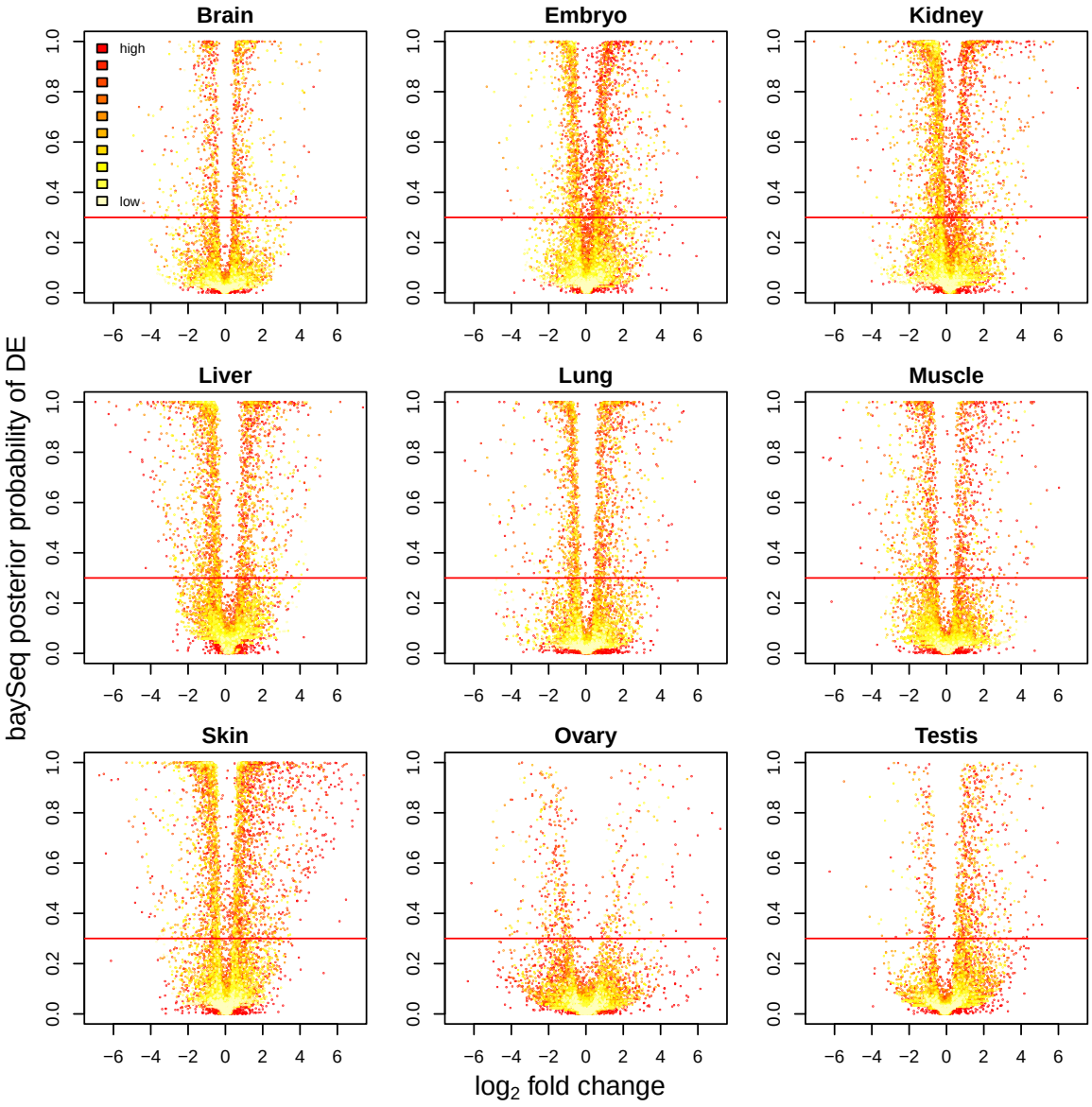


Figure S.2: `baySeq` **posterior probabilities of DE against $\log_2$ fold change in expression with colour indicating expression variance.**