Characterisation of the house sparrow (*Passer domesticus*) transcriptome: a resource for molecular ecology and immunogenetics

ROBERT EKBLOM[1,2*], PAUL WENNEKES[1], GAVIN J. HORSBURGH[2], TERRY BURKE[2]

1) Uppsala University, Department of Ecology and Genetics, Evolutionary Biology Centre, Norbyvägen 18 D, SE-75236, Uppsala, Sweden
2) University of Sheffield, Department of Animal and Plant Sciences, Sheffield S10 2TN, UK

* **Corresponding author:** robert.ekblom@ebc.uu.se
Tel: +46 (0)18-471 6468
Fax: +46 (0)18-471 6310

Key words:
RNA-Seq, Bird, SNP, microsatellite, MHC, Transcriptome

Running head:
The House Sparrow Transcriptome

Submitted as "Resource Article" to Molecular Ecology Resources
Total word count: 7762

Abstract

The house sparrow (*Passer domesticus*) is an important model species in ecology and evolution. However, until recently genomic resources for molecular ecological projects have been lacking in this species. Here we present transcriptome sequencing data (RNA-Seq) from three different house sparrow tissues (spleen, blood and bursa). These tissues were specifically chosen to obtain a diverse representation of expressed genes and to maximise the yield of immune-related gene functions. After *de-novo* assembly, 15,250 contigs were identified, representing sequence data from a total of 8,756 known avian genes (as inferred from the closely related zebra finch). The transcriptome assembly contain sequence data from nine manually annotated MHC genes, including an almost complete MHC class I coding sequence. There were 407, 303 and 68 genes overexpressed in spleen, blood and bursa, respectively. Gene ontology terms related to ribosomal function were associated with overexpression in spleen and oxygen transport functions with overexpression in blood. In addition to the transcript sequences, we provide 327 gene-linked microsatellites (SSRs) with sufficient flanking sequences for primer design, and 3,177 single nucleotide polymorphisms (SNPs) within genes, that can be used in follow-up molecular ecology studies of this ecological well-studied species.

Introduction

The house sparrow (*Passer domesticus*) is one of the world's most widespread and well-known bird species. Originally inhabiting the Western Palearctic (Ericson *et al.* 1997), its close proximity to humans enabled it to colonise all five continents, either by natural dispersal or by human-mediated introductions, mainly in the nineteenth century, followed by rapid range expansions (Anderson 2006). The aggregated, sedentary life style, close association with human settlements and propensity to breed in nest boxes have made the house sparrow an ideal model species for ecological studies. Research areas where studies on this species have contributed significantly include behavioural ecology (Barnard 1980), sexual selection (Møller 1987), endocrinology (Hegner & Wingfield 1987), population biology (Summers-Smith 1988), immunoecology (Gonzalez *et al.* 1999), physiology of plumage coloration (Evans *et al.* 2000) and hybrid speciation (Hermansen *et al.* 2011). Long-term ecological studies of wild house sparrow populations are currently being conducted on Lundy Island, England (Griffith *et al.* 1999; Schroeder *et al.* 2013), in an island metapopulation off the coast of Norway (Jensen *et al.* 2013; Jensen *et al.* 2004), at the University of Kentucky Agricultural Experiment Station, Kentucky (Westneat *et al.* 2004; Westneat *et al.* 2011), near Norman, Oklahoma (Mock *et al.* 2009; Schwagmeyer *et al.* 2002), around Veszprém, Hungary (Liker & Bókony 2009) and at the Centre d'Etude Biologique de Chize´ in France (Bonneaud *et al.* 2004a).

Although surveys of genetic variation have been a vital part of the research on house sparrows since the dawn of ecological genetics (Cole & Parkin 1981; Wetton *et al.* 1987), even recent genetic studies have mainly been limited to a small number of microsatellite markers (Garnier *et al.* 2009; Griffith *et al.* 2007) or sequencing of specific candidate genes, for example MHC

(Bonneaud *et al.* 2004b) and a few other immune-genes (Martin *et al.* 2011). Only very recently have genomic resources such as SNP (single nucleotide polymorphisms) markers (Hagen *et al.* 2013) become available. With the advent of massively parallel sequencing, genomic data are now becoming available for a large number of species. In particular, transcriptome sequencing (RNA-Seq; Mortazavi *et al.* 2008; Wang *et al.* 2009) is revolutionising genetic studies of non-model organisms (Ekblom & Galindo 2011; Vera *et al.* 2008; Vijay *et al.* 2013). Transcriptome resources have already been developed for a number of bird species including some work in the house sparrow (Hagen *et al.* 2013) and related species (Balakrishnan *et al.* 2013; Peterson *et al.* 2012). Here, we aim to contribute to the ongoing development of genomic resources in house sparrows (Hagen *et al.* 2013) by sequencing the transcriptomes of several tissues in order to quantify tissue-specific expression levels as well as identifying gene linked SNPs and microsatellites.

Much recent and current research on house sparrow is focused on immunoecology, in particular the links between sexual selection and immune function (Evans *et al.* 2000; Gonzalez *et al.* 1999). Genes involved in the immune system are often found to have evolved rapidly, probably as a result of a host-pathogen co-evolutionary arms race (Axelsson *et al.* 2008; Ekblom *et al.* 2010b; Nielsen *et al.* 2005). The strong signs of selection, together with a high degree of genetic variation in some immune genes, have made them an appropriate target for studying functional genetic variation. Some of the genes of the major histocompatibility complex (MHC), in particular, have been studied in many vertebrate species (Edwards & Hedrick 1998; Piertney & Oliver 2006; Spurgin & Richardson 2010). Genetic variation in MHC genes has also been investigated in various house sparrow populations (Bonneaud *et al.* 2004b; Borg *et al.* 2011). The MHC class I gene has been studied in particular detail, for

example in relation to life history (Bonneaud *et al.* 2004a), immunocompetence (Bonneaud *et al.* 2005), geographical genetic structure (Loiseau *et al.* 2009), avian malaria infection (Loiseau *et al.* 2011) and molecular evolution (Karlsson & Westerdahl 2013).

In this study, we aim to describe gene expression in specific tissues in which immune-related genes are known to be active. Blood includes transcriptionally active lymphocytes in which a variety of immune functions are regulated. RNA-Seq of whole blood (as used here) or "buffy coat" (the specific fraction of blood containing lymphocytes) should thus yield data on many different immune genes. The identification of genes expressed in blood has an added value, as this is the tissue that is most often routinely sampled during long-term field projects in birds (McDonald & Griffith 2011). If the expression of interesting candidate genes can be captured using this tissue this would then open up possibilities for large-scale transcriptomic studies of traits such as behaviour and life-history variation in a wide variety of organisms. The spleen functions as the main lymphoid organ, filtering out effete lymphocytes and erythrocytes, and responding actively to blood borne antigens (Roitt 1997). The avian spleen also plays a major role in the production, maturation and storage of lymphocytes (Smith & Hunt 2004). The bursa (or Bursa of Fabricius) is a tissue type specific to birds and is responsible for the maturation of b-cells (Click 1983; Glick *et al.* 1956), a function mainly performed by bone marrow in mammals. It is only transcriptionally active during development and is degenerated in adult birds. We included it in this study because it has previously been reported to be the primary tissue of transcription for many immune related genes (Ekblom *et al.* 2010b).

Material and methods

Sampling and RNA extraction

Whole blood (0.5 ml) and bursa samples came from one 9-day old chick and spleen was sampled from one adult male obtained from a natural population in the vicinity of Sheffield, England (ethical permission from Natural England, Licence number: 20092529). All samples were placed in RNAlater (Ambion Inc.) immediately after euthanasia and tissue preparation. RNA isolation was performed using TRIzol (Life Technologies) followed by cleaning with the RNeasy kit (Qiagen). RNA integrity and quantity were checked with the Bioanalyser RNA nano 6000 assay (Agilent Technologies).

cDNA synthesis

The MINT kit (Evrogen) was used for cDNA synthesis. A total of 1–2 μg RNA for each cDNA preparation and 18 cycles in the amplification step were found to work optimally for all samples. cDNA was normalised using the Trimmer kit (Evrogen) to remove excess quantities of very highly expressed genes and to maximise the likelihood of detecting genes with a low degree of expression. We have previously shown that cDNA samples can be highly informative for the analysis of differential gene expression, despite such normalisation, since much of the original variation in cDNA quantity still remains (Ekblom *et al.* 2012b).

Library preparation and sequencing

Five μg of each double-stranded cDNA sample was submitted to the Natural Environment Research Council (NERC) Biomolecular Analysis Facility at the University of Liverpool for sequencing. Library preparation and Roche 454 sequencing were performed according to the manufacturer's recommendations. Libraries for each tissue were individually tagged and sequenced together on half a plate of the 454 FLX system (Roche).

Transcriptome assembly and annotation

All raw sequencing reads were quality trimmed and filtered from adapter sequences, SMART primers and poly-A-tails using the commercially available SeqMan software (Lasergene). All reads with a trimmed length of less than 41 base pairs were discarded. The remaining reads from all three tissues together were assembled *de-novo* using SeqMan NGen 2.0 (Lasergene). Contigs were created if they included at least four reads. Reads remaining after contig assembly were kept as "singletons". In order to obtain gene expression levels for the different tissues individually, the template assembly function in NGen 2.0 was then used to map all reads from each tissue separately to the full transcriptome contig consensus sequence.

All contigs and singletons were annotated using blastn (Altschul *et al.* 1997), by matching the sequences to the Ensembl (Flicek *et al.* 2013) zebra finch (*Taeniopygia guttata*) transcript database (taeGut3.2.4, Ensembl database version 71.1) (Warren *et al.* 2010). An e-value cut off of 1e-10 was applied and only the best hit of each query sequence was considered. The total expression level for each gene and tissue was then calculated by counting all contig reads and singletons from the given gene and tissue (often more than one contig and/or singleton was annotated to the same gene).

We also specifically mined the house sparrow transcriptome data for 38 MHC genes that had previously been characterised in zebra finch (Balakrishnan *et al.* 2010). We used tblastx (Altschul *et al.* 1997), with an e-value cut-off of 1e-5, to identify MHC candidates in all house sparrow contigs and singletons. These candidate sequences were then reciprocally matched back against the zebra finch genome (using blastn) and cDNA database (using tblastx), and only candidates with a best reciprocal blast against the target gene were

annotated. The resulting sequences were then aligned against zebra finch and chicken homologues using Clustal W (Thompson *et al.* 1994) and manually inspected in BioEdit 7.2.0 (Hall 1999).

Differential gene expression and gene ontology analysis

Differential gene expression analyses were performed using the edgeR, Bioconductor R package (Gentleman *et al.* 2004; Robinson *et al.* 2010), applying the "TMM" normalisation procedure to account for sequencing depth and between-sample differences in RNA composition. Since we did not have any biological replicates, the common dispersion parameter could not be reliably estimated from our data and significance testing could not be performed. Instead we took the 5^{th} upper percentile of most differentially expressed genes for each pairwise comparison as the basis for the functional analyses (see below).

Gene ontology (GO) enrichment analyses were performed for all differentially expressed genes using the CORNA algorithm (Wu & Watson 2009), applied using the web interface provided by Michael Watson at the Institute for Animal Health (http://www.ark-genomics.org/tools/GOfinch). For each tissue, all genes found to be overexpressed in comparison with either of the other two tissues were considered in these analyses. An adjusted p-value of < 0.05 from the Fisher exact test was considered as significant and only GO terms occurring more often than expected were extracted.

Identification of microsatellites and SNPs

All contigs and singletons (246,761 sequences in total) were searched for microsatellite repeats (di- to hexa nucleotide repeats) using the software MsatCommander (Faircloth 2008).

Primers were designed from flanking sequences using the Primer3 plugin (Rozen & Skaletsky 2000).

We identified SNPs in the transcriptome data using the software PanGEA (Kofler *et al.* 2009). Alignments between all contigs and trimmed reads were performed using default settings with the "homopolymer Smith–Waterman" algorithm. We searched for non-indel SNPs with exactly two alleles, a minor allele count of at least two and a minimum alignment depth at the SNP site of four reads. Only SNPs with a minimum distance of 50 bp from the alignment end were extracted. Flanking sequences were extracted from the contig file using custom Perl and R scripting.

Results

Sequencing results

In total, 447,967 transcriptome sequencing reads with a mean read length of 302.6 base pairs were obtained for the three tissues (Table 1). Raw reads are available from NCBI-SRA (project accession number SRP012188). After removal of adaptor-, primer- and poly-A-sequences and quality trimming, 381,842 reads with a mean read length of 302.5 base pairs remained. In this way, fifteen per cent of the raw data were trimmed away before assembly.

Transcriptome characterisation

The trimmed reads from all tissues combined were assembled *de-novo* into 15,250 contigs. The complete sequences of all contigs are available in the supplementary material (in fasta format). The length of the contigs varied between 41 and 2,462 base pairs (Figure 1), with a total contig length of 8.4 Mbp. In total, 150,331 of the reads were included in the assembly

while the rest were kept as singletons. The mean number of reads per contig was 9.9 (range 4–3,302).

Annotation using information from zebra finch (taeGut3.2.1) gave significant hits on 8,756 known genes. This represents half of all the transcripts included in the zebra finch gene build used in the annotation process (Warren *et al.* 2010). Of these, 844 genes were covered by more than one contig, indicating some degree of fragmentation of the transcript assembly. The coverage was in the same order of magnitude for all tissues, but in the spleen, sampled from the adult bird, about 35% more genes were found than in the other two tissues (both sampled from a nestling). In total, 5,932 genes were expressed in spleen, 4,425 in blood and 4,233 in bursa (Figure 2a). When lowly expressed genes (fewer than five reads) were discarded, the proportions stayed the same, with 3,115 genes for spleen, 2,390 for blood and 2,390 for bursa (Figure 2b).

Annotation of MHC genes

We manually annotated transcriptome sequence data from nine MHC genes (*B2M*, *CD1A*, *CIITA*, *MHC Class I*, *MHC class IIB*, *Ii* (*CD74*), *Trim7.2*, *TAP1* and *B.NK*). The sequences of these are available in the Supplementary material. Of these, *CD1A* and *CD74* showed evidence for the presence of more than one splice variant. Coverage varied from only a few reads aligning to a small part of the coding sequence (for example *Class IIB*, *B.NK* and *Trim7.2*), to many contigs and singletons covering the complete coding sequence and untranslated regions of the transcript (for example *CD1A* and *CD74*).

Tissue specific expression patterns

For all three pairwise analyses of differential expression, differentially expressed genes were found in both directions. There were 254 genes overexpressed in the spleen tissue, versus 140 in blood (figure 3a). Three hundred and twenty-five genes were overexpressed in the spleen tissue, versus 34 in bursa (figure 3b), and 287 genes were overexpressed in the blood tissue, versus 50 in bursa (figure 3c). For spleen, in total 407 genes were overexpressed if the two analyses were combined. One hundred and seventy-two of those were overexpressed in both analyses. For blood, 303 genes were overexpressed, of which 124 were overexpressed in both analyses. For bursa, 68 overexpressed genes were found, and 16 of those were overexpressed against both other tissues.

Gene ontology terms that were significantly overrepresented were found for genes overexpressed in all three tissues (27 terms for spleen, 12 terms for blood and 29 for bursa; Table 2). Many of the gene ontology terms represented common cell maintenance functions, such as protein binding or biosynthesis. For bursa, most of the significant GO terms were found only once in the overexpressed gene list. For blood, the most highly significant terms were related to haemoglobin and oxygen binding. For spleen, the most highly significant terms were related to translation and ribosomal functions. Different immune related gene ontology term were also found to be enriched in each of the tissues.

Identification of microsatellites and SNPs

We identified a total of 968 microsatellite repeat sequences in the house sparrow transcriptome (Table 3). As tri- and hexa-nucleotide repeats (non-reading frame interrupting) are the least common types, a majority of these are expected to be situated in untranslated regions of the expressed genes (Primmer 2009). Of the identified repeats, 327 (34%) had

sufficient flanking sequence information to allow for PCR primer design (Table 3). Primer sequences for these are available in the Supplementary Material.

We also identified 3,177 SNPs situated in 1,702 of the transcriptome contig sequences. The numbers of transitions and transversions at the SNP sites were 2,099 and 1,078, respectively, giving a Ts/Tv ratio of 1.95. The GC ratio at the SNP sites was 49.6 %. Minor allele frequencies at SNPs could not be reliably calculated, as sequence data were only available for two individuals. SNP information, including flanking sequences, is available in the Supplementary Material. Note that there is a risk that the described flanking sequences of some of these SNPs span an exon/intron boundary. Consequently, SNP genotyping efficiency using gDNA might be reduced.

Discussion

Next-generation cDNA sequencing has proved to be a rapid and effective way to develop molecular tools in species without an existing genome sequence (Ekblom & Galindo 2011). Since the first proofs of concept in the late 00s (Toth *et al.* 2007; Vera *et al.* 2008), many species have now had their transcriptomes characterised in this way. Such descriptive studies have proven to lay important foundations for investigating genomic aspects of fields such as quantitative trait variation (Robinson *et al.* 2013), life history evolution (Santure *et al.* 2013), speciation (Jeukens *et al.* 2010), local adaptation (De Wit & Palumbi 2013) and sexual selection (Ekblom *et al.* 2012a). Transcriptome information is also of great help in the assembly and annotation of genomic sequences. Here, we provide transcriptome sequences for several important immune tissues of the house sparrow. Given the importance of this

280 species as an ecological model (Anderson 2006; Jensen *et al.* 2004), we expect the
281 transcriptome data to aid future studies at the intersection of ecology and genomics.
282
283 Our house sparrow transcriptome covers more than eight thousand annotated avian genes,
284 representing almost half of all genes identified in the zebra finch (Ekblom *et al.* 2010a;
285 Warren *et al.* 2010). Many of these are, however, expressed at a very low level and only a
286 fraction of the coding sequence is present in the assembly. Our mean (550bp) and maximum
287 (2,462bp) contig lengths also suggest that most transcripts were not covered by one
288 contiguous sequence. These figures are, however, in line with expectations from previous
289 454-sequenced transcriptomes (e.g. Ekblom *et al.* 2010a; Wang *et al.* 2012; Vera *et al.* 2008),
290 although considerably higher coverage have also recently been reported (Balakrishnan *et al.*
291 2013). Spleen was the tissue with most annotated genes, and also the tissue with the highest
292 proportion of uniquely expressed genes. Bursa had the lowest overall expression levels and
293 also relatively few sequencing reads in the raw data. Given that the libraries were constructed
294 from equal amounts of cDNA, these differences are likely to be a result of technical artefacts
295 rather than biological variation.
296
297 Unsurprisingly, there was an overrepresentation of oxygen-binding and transport functions in
298 genes differentially expressed in blood. Perhaps somewhat more unexpected was the strong
299 association between overexpression in spleen and GO terms related to translation and
300 ribosomal function. A very similar result was previously reported in the characterisation of
301 the spleen transcriptome in zebra finch (Ekblom *et al.* 2010a). It thus seems to be a general
302 trend that ribosomal genes are highly expressed in this tissue. This could be an indication of
303 unusually high rates of translation in the spleen. A recent study in humans also established a

clear link between mutations in a ribosomal protein and aberrant spleen development (Bolze *et al.* 2013). The results from the differential expression analyses need to be treated with caution for several reasons. First, we used a cDNA normalisation step in the library preparation in order to produce a more even coverage of the transcribed genes. Normalisation reduces extreme expression differences but there is thus still an expression signal left also after normalisation. We have previously found that there is a rather strong correlation between expression differentiation in normalised and un-normalised samples from the same tissues (Ekblom *et al.* 2012b). Second, we only included one sample for each tissue type. This lack of replication may have severe effects on the inference of differential expression, as it prevents reliable significance testing of the results and adds noise to the down-stream analyses. Third, spleen RNA was extracted from an adult individual, while whole blood and bursa came from a nestling, so expression differences between spleen and other tissues may be related to age specific and individual variation as well as tissue specific expression levels. Fourth, the large differences in sequencing output and coverage between the different tissues could influence these analyses (although the software used is designed to correct for such differences).

Immune genes have been of special interest in many investigations of adaptive genetic variation that have used a candidate gene approach. Genes involved in the avian immune system have previously been found to be evolving at a higher rate than other parts of the genome (Axelsson *et al.* 2008; Lindblad-Toh *et al.* 2011). Such genes also have a more tissue-specific expression pattern than other genes, and primary expression of most immune genes was found in the three specific tissues sampled in this study (Ekblom *et al.* 2010b). We manually annotated nine MHC genes from the house sparrow transcriptome sequence. Among these were a nearly complete coding sequence of an *MHC class I* locus. The third exon of this

sequence is almost identical to the previously described house sparrow allele Pado*109 (Loiseau *et al.* 2011). The *MHC class IIB* sequence found here is not identical to any of the previously sequenced house sparrow class II alleles (Bonneaud *et al.* 2004b), but the sequenced part of the second exon is strikingly similar to the annotated zebra finch *MHC class IIB* locus 3 (Balakrishnan *et al.* 2010). Our manually annotated MHC sequences also include the complete coding sequence of two isoforms of the *CD74* gene (also known as the *MHC invariant chain* or *Ii*). Expression of this gene has previously been found to be of importance for the response to *Mycoplasma* infection in the phylogenetically related house finch (Backström *et al.* 2013; Bonneaud *et al.* 2011). For 29 of the MHC genes that we tried to annotate manually we could not find a homologous sequence in the house sparrow transcriptome data. This is likely a result of very low expression of these genes in the absence of an infection. An alternative approach that might be more successful in this respect would be to sample immune tissues from individuals with a triggered immune response, either from infection or experimental manipulation. It is also possible that the nucleotide sequences of these genes have diverged sufficiently from the zebra finch orthologs (as a result of strong positive selection) to hinder proper sequence alignment.

High-throughput genomic or transcriptomic sequences can be readily mined for molecular genetic markers such as microsatellite repeat sequences or simple sequence repeats (SSRs). If more than one haploid genome is sequenced and the depth of sequencing is sufficient, the data can also be used to identify polymorphic nucleotide positions (for example SNPs and indels). We have utilised the house sparrow transcriptome for both these applications. The novel microsatellite markers developed here will add to the previously identified species-specific markers (Dawson *et al.* 2012), as well as recently-developed microsatellite markers with

general utility across the birds (Dawson *et al.* 2013; Dawson *et al.* 2010). Microsatellites developed from transcriptomes may not be selectively neutral, since they are linked to expressed genes, and should thus be used with caution in population genetic analyses. On the other hand, such markers can be highly informative in genomic outlier analyses and other kinds of investigations of functional genetic variation (Narum & Hess 2011; Willing *et al.* 2010). As microsatellite markers can often be successfully applied to species other than those in which they were developed, it will be possible to utilise this resource in studies of other bird species as well.

SNP markers have recently become widely used in population genetics and molecular ecology (Morin *et al.* 2004). They are cheap and easy to score, meaning that information on genetic variation across the whole genome can be easily obtained. In addition, these markers do not do not suffer from the high levels of homoplasy commonly found in microsatellites (Payseur & Cutter 2006), and have a well-defined and simple mutation model (Garvin *et al.* 2010). In an important recent effort, Hagen and co-workers (2013) developed a 10k SNP chip for the house sparrow based on genome and transcriptome sequences from a variety of tissues (including blood but not bursa or spleen). SNPs will thus be the likely marker of choice for many forthcoming studies in this species. The SNPs we have identified from the transcriptome of immunologically important tissues, will complement those already developed from the Norwegian island populations. As our samples come from a different population, combining these two sets of SNPs may decrease the potential problem of ascertainment bias in future applications. We used a blast approach to check for overlap between the 3,177 SNPs reported here and the verified 8,491 SNP chip. We found exact matches for only ten SNPs between these and an additional 42 SNPs were situated in the close

vicinity (in the flanking sequence of) to the SNPs on the chip. This rather small overlap between the two datasets may suggest some degree of genetic differentiation between the two sampled populations. Compared to microsatellite markers, SNPs are often much more species specific. It has, however, recently been found that many SNP markers can also be utilised in closely related species of interest (Miller *et al.* 2011; Pertoldi *et al.* 2010), making this resource more generally applicable than just in the focal organism.

The molecular markers (candidate gene sequences, SNPs and microsatellites) identified here, together with the transcriptome sequence and expression profiles provided, will provide important genomic tools for future studies of the house sparrow. We hope that the ongoing long-term projects of this fascinating bird will benefit from this resource and that they will be able to address new and exciting questions in molecular ecology. The description of MHC gene sequences and expression in whole blood will clearly be of use in ongoing efforts to understand the role of this important multigene family in ecology and evolution. The genetic markers developed here are also likely to be applicable to other closely related species of interest.

Acknowledgements

We thank Amy Llewellyn for assistance in lab and Tim Birkhead for preparing tissue samples. Margaret Hughes at the NBAF–Liverpool assisted with library preparation and sequencing. Ingerid J. Hagen kindly provided flanking sequences for the 10k house sparrow SNP chip. Jelmer Poelstra, and Özden Baltekin discussed analyses and results, and three anonymous reviewers kindly provided comments on previous versions of this manuscript. Ethical permission to collect wild house sparrows was provided by Natural England. R.E. was

funded by the Carl Trygger Foundation (grant code CTS 09:87). This study was supported by the UK Natural Environment Research Council.

References

Altschul SF, Madden TL, Schaffer AA*, et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acids Res* **25**, 3389-3402.

Anderson TR (2006) *Biology of the ubiquitous house sparrow: from genes to populations* Oxford University Press New York.

Axelsson E, Hultin-Rosenberg L, Brandstrom M*, et al.* (2008) Natural selection in avian protein-coding genes expressed in brain. *Molecular Ecology* **17**, 3008-3017.

Backström N, Shipilina D, Blom MPK, Edwards SV (2013) Cis-regulatory sequence variation and association with Mycoplasma load in natural populations of the house finch (Carpodacus mexicanus). *Ecology and Evolution* **3**, 655-666.

Balakrishnan C, Ekblom R, Volker M*, et al.* (2010) Gene duplication and fragmentation in the zebra finch major histocompatibility complex. *BMC Biology* **8**, 29.

Balakrishnan CN, Chapus C, Brewer MS, Clayton DF (2013) Brain transcriptome of the violet-eared waxbill Uraeginthus granatina and recent evolution in the songbird genome. *Open Biology* **3**.

Barnard C (1980) Flock feeding and time budgets in the house sparrow (*Passer domesticus* L.). *Animal Behaviour* **28**, 295-309.

Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. B* **57**, 289-300.

Bolze A, Mahlaoui N, Byun M*, et al.* (2013) Ribosomal Protein SA Haploinsufficiency in Humans with Isolated Congenital Asplenia. *Science* **340**, 976-978.

Bonneaud C, Balenger SL, Russell AF*, et al.* (2011) Rapid evolution of disease resistance is accompanied by functional changes in gene expression in a wild bird. *Proceedings of the National Academy of Sciences of the USA* **108**, 7866-7871.

Bonneaud C, Mazuc J, Chastel O, Westerdahl H, Sorci G (2004a) Terminal investment induced by immune challenge and fitness traits associated with major histocompatibility complex in the house sparrow. *Evolution* **58**, 2823-2830.

Bonneaud C, Richard M, Faivre B, Westerdahl H, Sorci G (2005) An Mhc class I allele associated to the expression of T-dependent immune response in the house sparrow. *Immunogenetics* **57**, 782-789.

Bonneaud C, Sorci G, Morin V*, et al.* (2004b) Diversity of Mhc class I and IIB genes in house sparrows (Passer domesticus). *Immunogenetics* **55**, 855-865.

Borg ÅA, Pedersen SA, Jensen H, Westerdahl H (2011) Variation in MHC genotypes in two populations of house sparrow (*Passer domesticus*) with different population histories. *Ecology and Evolution* **1**, 145-159.

Click B (1983) Bursa of Fabricius. *Avian biology* **7**, 443-500.

Cole SR, Parkin DT (1981) Enzyme polymorphisms in the House Sparrow, Passer domesticus. *Biological Journal of the Linnean Society* **15**, 13-22.

Dawson D, Ball A, Spurgin L*, et al.* (2013) High-utility conserved avian microsatellite markers enable parentage and population studies across a wide range of species. *BMC Genomics* **14**, 176.

Dawson DA, Horsburgh GJ, Krupa AP*, et al.* (2012) Microsatellite resources for Passeridae species: a predicted microsatellite map of the house sparrow *Passer domesticus*. *Molecular Ecology Resources* **12**, 501-523.
Dawson DA, Horsburgh GJ, Küpper C*, et al.* (2010) New methods to identify conserved microsatellite loci and develop primer sets of high cross-species utility – as demonstrated for birds. *Molecular Ecology Resources* **10**, 475-494.
De Wit P, Palumbi SR (2013) Transcriptome-wide polymorphisms of red abalone (Haliotis rufescens) reveal patterns of gene flow and local adaptation. *Molecular Ecology* **22**, 2884-2897.
Edwards SV, Hedrick PW (1998) Evolution and ecology of MHC molecules: from genomics to sexual selection. *Trends Ecol Evol* **13**, 305-311.
Ekblom R, Balakrishnan CN, Burke T, Slate J (2010a) Digital gene expression analysis of the zebra finch genome. *BMC Genomics* **11**, 219.
Ekblom R, Farrell LL, Lank DB, Burke T (2012a) Gene expression divergence and nucleotide differentiation between males of different color morphs and mating strategies in the ruff. *Ecology and Evolution* **2**, 2485-2505.
Ekblom R, French L, Slate J, Burke T (2010b) Evolutionary Analysis and Expression Profiling of Zebra Finch Immune Genes *Genome Biology and Evolution* **2**, 781-790.
Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity* **107**, 1-15.
Ekblom R, Slate J, Horsburgh GJ, Birkhead T, Burke T (2012b) Comparison between Normalised and Unnormalised 454-Sequencing Libraries for Small-Scale RNA-Seq Studies. *Comparative and Functional Genomics* **2012**, 8.
Ericson PGP, Tyrberg T, Kjellberg AS, Jonsson L, Ullén I (1997) The Earliest Record of House Sparrows (Passer domesticus) in Northern Europe. *Journal of Archaeological Science* **24**, 183-190.
Evans MR, Goldsmith AR, Norris SR (2000) The effects of testosterone on antibody production and plumage coloration in male house sparrows (Passer domesticus). *Behavioral Ecology and Sociobiology* **47**, 156-163.
Faircloth BC (2008) msatcommander: detection of microsatellite repeat arrays and automated, locus-specific primer design. *Molecular Ecology Resources* **8**, 92-94.
Flicek P, Ahmed I, Amode MR*, et al.* (2013) Ensembl 2013. *Nucleic Acids Research* **41**, D48-D55.
Garnier S, Durand P, Arnathau C*, et al.* (2009) New polymorphic microsatellite loci in the house sparrow, Passer domesticus. *Molecular Ecology Resources* **9**, 1063-1065.
Garvin MR, Saitoh K, Gharrett AJ (2010) Application of single nucleotide polymorphisms to non-model species: a technical review. *Molecular Ecology Resources* **10**, 915-934.
Gentleman R, Carey V, Bates D*, et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* **5**, R80.
Glick B, Chang TS, Jaap RG (1956) The Bursa of Fabricius and Antibody Production. *Poultry Science* **35**, 224-225.
Gonzalez G, Sorci G, Moller AP*, et al.* (1999) Immunocompetence and condition-dependent sexual advertisment in male house sparrows (Passer domesticus). *J. Anim. Ecol.* **68**, 1225-1234.
Griffith SC, Dawson DA, Jensen H*, et al.* (2007) Fourteen polymorphic microsatellite loci characterized in the house sparrow Passer domesticus (Passeridae, Aves). *Molecular Ecology Notes* **7**, 333-336.
Griffith SC, Stewart IRK, Dawson DA, Owens IPF, Burke T (1999) Contrasting levels of extra-pair paternity in mainland and island populations of the house sparrow (*Passer domesticus*): is there an 'island effect'? *Biological Journal of the Linnean Society* **68**, 303-316.
Hagen IJ, Billing AM, Rønning B*, et al.* (2013) The easy road to genome-wide medium density SNP screening in a non-model species: development and application of a 10 K SNP-chip for the house sparrow (Passer domesticus). *Molecular Ecology Resources*, n/a-n/a.

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* **41**, 95-98.
Hegner RE, Wingfield JC (1987) Social status and circulating levels of hormones in flocks of house sparrows, *Passer domesticus*. *Ethology* **76**, 1-14.
Hermansen JS, Sæther SA, Elgvin TO*, et al.* (2011) Hybrid speciation in sparrows I: phenotypic intermediacy, genetic admixture and barriers to gene flow. *Molecular Ecology* **20**, 3812-3822.
Jensen H, Moe R, Hagen IJ*, et al.* (2013) Genetic variation and structure of house sparrow populations: is there an island effect? *Molecular Ecology* **22**, 1792-1805.
Jensen H, Sæther B-E, Ringsby TH*, et al.* (2004) Lifetime reproductive success in relation to morphology in the house sparrow Passer domesticus. *Journal of Animal Ecology* **73**, 599-611.
Jeukens J, Renaut S, St-Cyr J, Nolte AW, Bernatchez L (2010) The transcriptomics of sympatric dwarf and normal lake whitefish (*Coregonus clupeaformis* spp., Salmonidae) divergence as revealed by next-generation sequencing. *Molecular Ecology* **19**, 5389-5403.
Karlsson M, Westerdahl H (2013) Characteristics of MHC Class I Genes in House Sparrows Passer domesticus as Revealed by Long cDNA Transcripts and Amplicon Sequencing. *Journal of Molecular Evolution*, 1-14.
Kofler R, Teixeira Torres T, Lelley T, Schlötterer C (2009) PanGEA: identification of allele specific gene expression using the 454 technology. *BMC Bioinformatics* **10**, 143.
Liker A, Bókony V (2009) Larger groups are more successful in innovative problem solving in house sparrows. *Proceedings of the National Academy of Sciences* **106**, 7893-7898.
Lindblad-Toh K, Garber M, Zuk O*, et al.* (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482.
Loiseau C, Richard M, Garnier S*, et al.* (2009) Diversifying selection on MHC class I in the house sparrow (Passer domesticus). *Molecular Ecology* **18**, 1331-1340.
Loiseau C, Zoorob R, Robert A*, et al.* (2011) Plasmodium relictum infection and MHC diversity in the house sparrow (Passer domesticus). *Proceedings of the Royal Society B: Biological Sciences* **278**, 1264-1272.
Martin LB, Kidd L, Liebl AL, Coon CAC (2011) Captivity induces hyper-inflammation in the house sparrow *(Passer domesticus*). *The Journal of Experimental Biology* **214**, 2579-2585.
McDonald PG, Griffith SC (2011) To pluck or not to pluck: the hidden ethical and scientific costs of relying on feathers as a primary source of DNA. *Journal of Avian Biology* **42**, 197-203.
Miller JM, Poissant J, Kijas JW, Coltman DW, the International Sheep Genomics C (2011) A genome-wide set of SNPs detects population substructure and long range linkage disequilibrium in wild sheep. *Molecular Ecology Resources* **11**, 314-322.
Mock DW, Schwagmeyer PL, Dugas MB (2009) Parental provisioning and nestling mortality in house sparrows. *Animal Behaviour* **78**, 677-684.
Morin PA, Luikart G, Wayne RK, the SNPwg (2004) SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution* **19**, 208-216.
Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Meth* **5**, 621-628.
Møller AP (1987) Variation in badge size in male house sparrows Passer domesticus: evidence for status signalling. *Animal Behaviour* **35**, 1637-1644.
Narum SR, Hess JE (2011) Comparison of FST outlier tests for SNP loci under selection. *Molecular Ecology Resources* **11**, 184-194.
Nielsen R, Bustamante C, Clark AG*, et al.* (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biology* **3**, e170.
Payseur BA, Cutter AD (2006) Integrating patterns of polymorphism at SNPs and STRs. *Trends in Genetics* **22**, 424-429.
Pertoldi C, Wójcik J, Tokarska M*, et al.* (2010) Genome variability in European and American bison detected using the BovineSNP50 BeadChip. *Conservation Genetics* **11**, 627-634.

Peterson M, Whittaker D, Ambreth S*, et al.* (2012) De novo transcriptome sequencing in a songbird, the dark-eyed junco (Junco hyemalis): genomic tools for an ecological model system. *BMC Genomics* **13**, 305.
Piertney SB, Oliver MK (2006) The evolutionary ecology of the major histocompatibility complex. *Heredity* **96**, 7-21.
Primmer CR (2009) From conservation genetics to conservation genomics. *Annals of the New York Academy of Sciences* **1162**, 357-368.
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 MR, Santure AW, DeCauwer I, Sheldon BC, Slate J (2013) Partitioning of genetic variation across the genome using multimarker methods in a wild bird population. *Molecular Ecology* **22**, 3963-3980.
Roitt IM (1997) *Essential immunology*, 9th ed. edn. Blackwell Science Ltd, Oxford.
Rozen S, Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. In: *Bioinformatics Methods and Protocols: Methods in Molecular Biology* (eds. Krawetz S, Misener S). Humana Press, Totowa, NJ.
Santure AW, De Cauwer I, Robinson MR*, et al.* (2013) Genomic dissection of variation in clutch size and egg mass in a wild great tit (Parus major) population. *Molecular Ecology* **22**, 3949-3962.
Schroeder J, Cleasby I, Dugdale HL, Nakagawa S, Burke T (2013) Social and genetic benefits of parental investment suggest sex differences in selection pressures. *Journal of Avian Biology* **44**, 133-140.
Schwagmeyer PL, Mock DW, Parker GA (2002) Biparental care in house sparrows: negotiation or sealed bid? *Behavioral Ecology* **13**, 713-721.
Smith K, Hunt J (2004) On the use of spleen mass as a measure of avian immune system strength. *Oecologia* **138**, 28-31.
Spurgin LG, Richardson DS (2010) How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proceedings of the Royal Society B: Biological Sciences* **277**, 979-988.
Summers-Smith JD (1988) *The Sparrows: A study of the genus Passer* T & AD Poyser, Calton, Staffordshire, England.
Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673-4680.
Toth AL, Varala K, Newman TC*, et al.* (2007) Wasp gene expression supports an evolutionary link between maternal behavior and eusociality. *Science* **318**, 441-444.
Wang B, Ekblom R, Castoe TA*, et al.* (2012) Transcriptome sequencing of black grouse (*Tetrao tetrix*) for immune gene discovery and microsatellite development. *Open Biology* **2**, 120054.
Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57-63.
Warren WC, Clayton DF, Ellegren H*, et al.* (2010) The genome of a songbird. *Nature* **464**, 757-762.
Vera JC, Wheat CW, Fescemyer HW*, et al.* (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology* **17**, 1636-1647.
Westneat D, Weiskittle J, Edenfield R, Kinnard T, Poston J (2004) Correlates of cell-mediated immunity in nestling house sparrows. *Oecologia* **141**, 17-23.
Westneat DF, Hatch MI, Wetzel DP, Ensminger AL (2011) Individual Variation in Parental Care Reaction Norms: Integration of Personality and Plasticity. *The American Naturalist* **178**, 652-667.
Wetton JH, Carter RE, Parkin DT, Walters D (1987) Demographic study of a wild house sparrow population by DNA fingerprinting. *Nature* **327**, 147 - 149.

Vijay N, Poelstra JW, Künstner A, Wolf JBW (2013) Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. *Molecular Ecology* **22**, 620-634.

Willing E-M, Bentzen P, OOSTERHOUT Cv*, et al.* (2010) Genome-wide single nucleotide polymorphisms reveal population history and adaptive divergence in wild guppies. *Molecular Ecology* **19**, 968-984.

Wu X, Watson M (2009) CORNA: testing gene lists for regulation by microRNAs. *Bioinformatics* **25**, 832-833.

Data Accessibility

Raw sequence reads are available from NCBI-SRA (project accession number SRP012188).

The following data is available as Supplementary Material:

- Contig sequences for the complete house sparrow transcriptome (as fasta file).

- Transcriptome sequences of 9 house sparrow MHC genes (as fasta file).

- PCR primers and information about 327 microsatellite markers identified from the house sparrow transcriptome (as .csv).

- Information (including flanking sequence) about 3177 SNP markers identified in the house sparrow transcriptome (as .csv).

Author Contributions

RE and TB designed the study and conducted field sampling. GJH conducted the laboratory work. RE and PW performed the analyses and drafted the manuscript. All authors contributed significantly to the writing and commenting on the manuscript.

Figure Legends

Fig. 1 Distribution of contig lengths of the *de-novo* assembly of the house sparrow transcriptome (all tissues combined).

Fig. 2 Venn diagrams showing the total number genes (left), and number of genes represented by at least five sequence reads (right), expressed in spleen, blood or bursa. Overlapping areas represent genes that were expressed in more than one tissue.

Fig. 3 Differential expression of genes in the three pairwise comparisons between the sampled tissues: (a) spleen vs. blood, (b) bursa vs. blood and (c) bursa vs. spleen. The right part of the graphs (sideways V-shaped parts) represent genes that were expressed in both tissues, x-axis represents total expression in both tissues together, and the y-axis represents log-fold difference in expression between the tissues. The blue lines represent a four-fold expression difference; the red dots represent genes that were overexpressed in blood (low on the y axis of a and b), bursa (high on the y axis of b and c) or spleen (high on the y-axis of a and low on the y axis of c). In the left part of the graphs all genes that were only expressed in one of the two tissues are plotted separately. The expression levels are shown as if there was one case of expression in the tissue not expressed, to prevent dividing by zero; overexpressed genes are similarly shown in red.

Tables

Table 1 Number of sequencing reads (and mean length) obtained for each sequenced tissue type

Tissue	Raw data	After primer and poly-A trimming	After quality filtering
Spleen	178,389 (285 bp)	154,742 (280 bp)	152,804 (280 bp)
Blood	198,411 (309 bp)	168,760 (311 bp)	165,747 (314 bp)
Bursa	71,167 (329 bp)	64,186 (324 bp)	63,291 (327 bp)

641 **Table 2** Gene ontology term enrichment analysis of over-expressed genes for each of the three sampled tissues

642 (spleen, blood and bursa). Immune related GO terms are highlighted in bold.

go description	total	expectation	observation	adjusted fisher p-value
Spleen				
ribosome	125	3	43	1.00E-32
structural constituent of ribosome	123	3	41	9.80E-31
translation	154	4	44	4.80E-30
ribonucleoprotein complex	106	3	32	2.50E-22
cytosolic large ribosomal subunit	34	1	17	8.30E-16
cytosolic small ribosomal subunit	18	1	11	3.30E-11
respiratory chain	11	0	7	5.70E-07
electron transport chain	14	0	7	4.80E-06
NADH dehydrogenase (ubiquinone) activity	24	1	7	0.00029
cytochrome-c oxidase activity	16	0	6	0.00029
ATP synthesis coupled electron transport	5	0	4	0.00029
translational elongation	11	0	5	0.00059
mitochondrial inner membrane	126	4	14	9.00E-04
translation elongation factor activity	8	0	4	0.0029
respiratory electron transport chain	9	0	4	0.0048
intracellular	1674	47	73	0.0059
protein folding	100	3	11	0.0066
oxidoreductase activity	361	10	23	0.014
small ribosomal subunit	12	0	4	0.014
catalytic step 2 spliceosome	49	1	7	0.021
external side of plasma membrane	83	2	9	0.025

aconitate hydratase activity	2	0	2	0.03
positive regulation of tyrosine phosphorylation of STAT protein	2	0	2	0.03
activation of signaling protein activity involved in unfolded protein response	2	0	2	0.03
DNA double-strand break processing	2	0	2	0.03
UBC13-MMS2 complex	2	0	2	0.03
wound healing involved in inflammatory response	2	0	2	0.03
Blood				
oxygen transporter activity	8	0	4	0.0024
oxygen binding	8	0	4	0.0024
oxygen transport	8	0	4	0.0024
hemoglobin complex	4	0	3	0.0057
antigen processing and presentation, exogenous lipid antigen via MHC class Ib	2	0	2	0.042
endosome membrane	19	0	4	0.042
cyclic nucleotide biosynthetic process	20	0	4	0.042
heme binding	97	2	8	0.042
guanylate cyclase activity	9	0	3	0.042
chaperone binding	21	0	4	0.042
phosphorus-oxygen lyase activity	21	0	4	0.042
cyclin binding	10	0	3	0.049
Bursa				
DNA binding	907	4	11	0.043
T cell receptor signaling pathway	19	0	2	0.043
protein autophosphorylation	68	0	3	0.043

nuclear speck organization	1	0	1	0.043
centrosome separation	1	0	1	0.043
orotate phosphoribosyltransferase activity	1	0	1	0.043
'de novo' UMP biosynthetic process	1	0	1	0.043
orotidine-5'-phosphate decarboxylase activity	1	0	1	0.043
pilomotor reflex	1	0	1	0.043
RNA polymerase II regulatory region DNA binding	1	0	1	0.043
integrase activity	1	0	1	0.043
blastocyst formation	1	0	1	0.043
annulate lamellae	1	0	1	0.043
positive regulation of neutrophil differentiation	1	0	1	0.043
positive regulation of histone modification	1	0	1	0.043
UBC13-UEV1A complex	1	0	1	0.043
regulation of histone ubiquitination	1	0	1	0.043
receptor signaling protein tyrosine phosphatase activity	1	0	1	0.043
pre-B cell differentiation	1	0	1	0.043
ventral spinal cord development	1	0	1	0.043
cardiac muscle cell differentiation	1	0	1	0.043
sarcomere organization	1	0	1	0.043
immunoglobulin V(D)J recombination	1	0	1	0.043
motor axon guidance	1	0	1	0.043
cardiovascular system development	1	0	1	0.043
RNA polymerase II distal enhancer sequence-specific DNA binding	1	0	1	0.043

transcription factor activity involved in positive regulation of transcription				
1,4-alpha-glucan branching enzyme activity	1	0	1	0.043
microglial cell activation involved in immune response	1	0	1	0.043
nucleus	2257	10	19	0.044

643

644

645

646 **Table 3** Number and type of microsatellite repeats identified from the house sparrow transcriptome

Type	Min. repeat number	Total number of repeats	Number of repeats with primer sequence
Di-	10	151	19
Tri-	8	90	18
Tetra-	4	464	226
Penta-	4	197	52
Hexa-	4	66	12
All		968	327

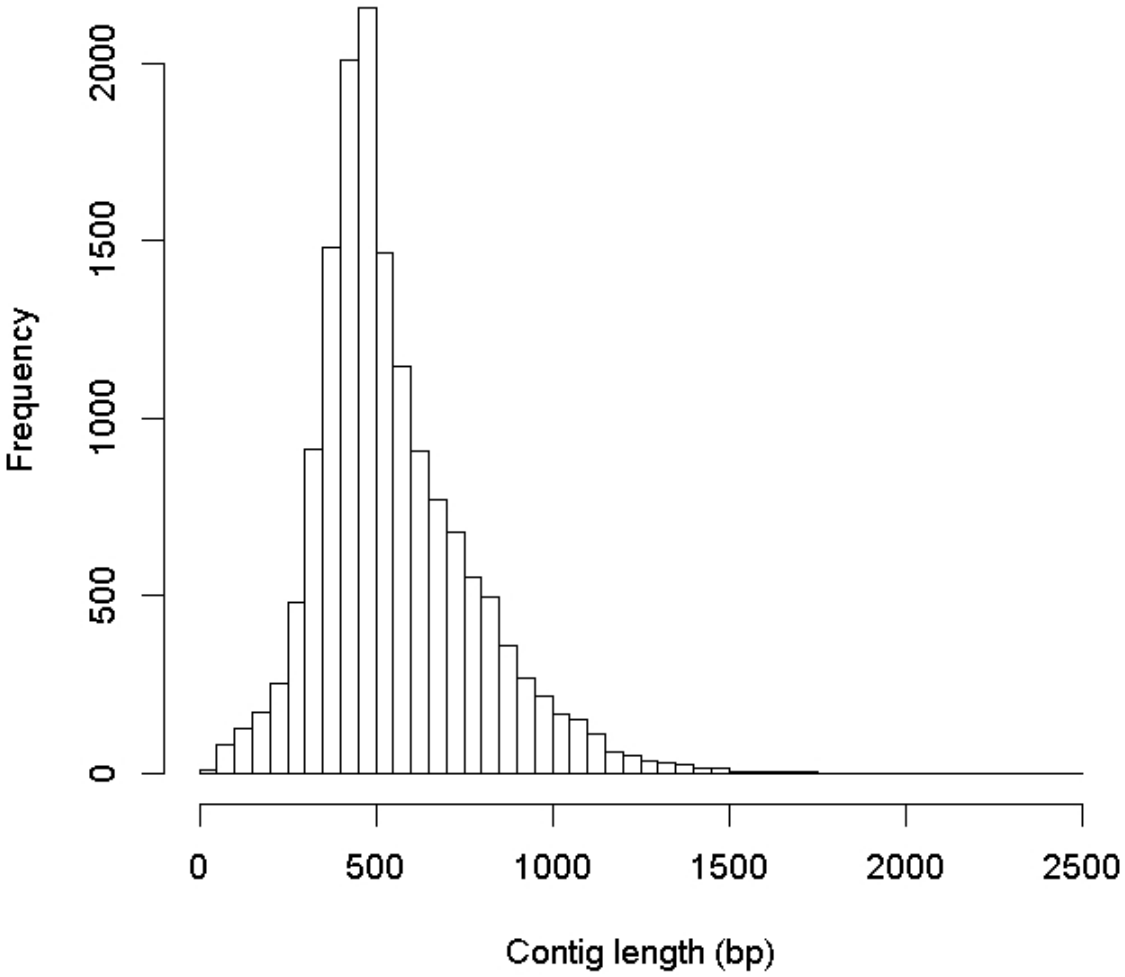
647

648

649

650 **Figures**

651 Figure 1.



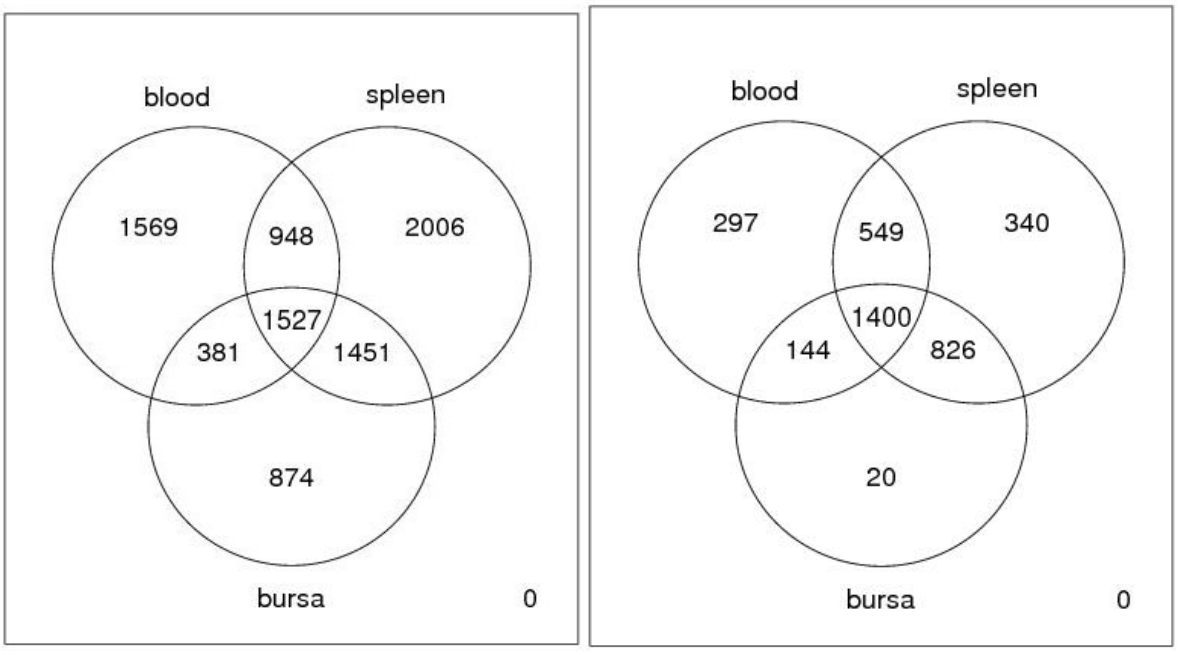
652

653

654

655 Figure 2.

656

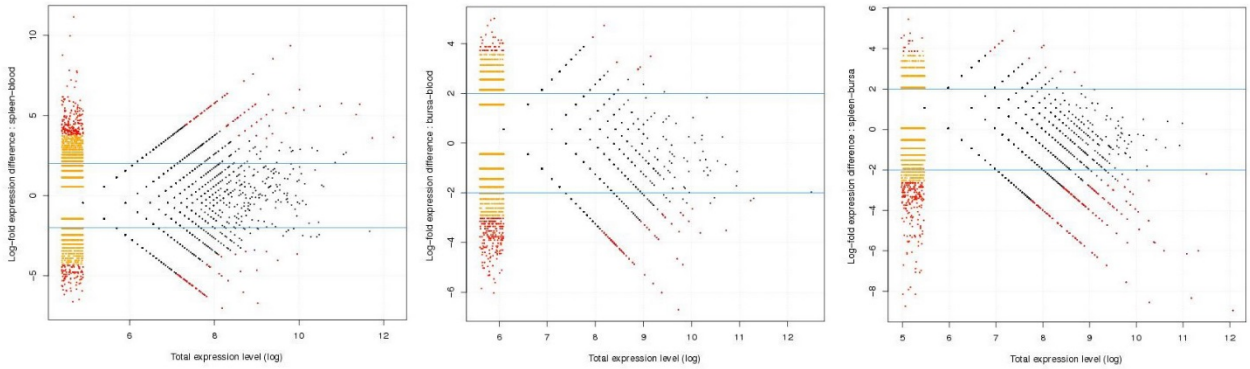


657

658

659 Figure 3.

660



661