MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2013) **13**, 696–705 doi: 10.1111/1755-0998.12096

The plover neurotranscriptome assembly: transcriptomic analysis in an ecological model species without a reference genome

HOOMAN K. MOGHADAM,* PETER W. HARRISON,† GERGELY ZACHAR,‡ TAMÁS SZÉKELY§ and JUDITH E. MANK†

**Institute of Marine Biology, Biotechnology & Aquaculture (IMBBC), Hellenic Centre for Marine Research (HCMR), PO Box 2214, 71500 Heraklion, Crete, Greece, †Department of Genetics, Evolution and Environment, University College London, The Darwin Building, Gower Street, London WC1E 6BT, UK, ‡Department of Anatomy, Histology and Embryology, Semmelweis University, Budapest H-1094, Hungary, §Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK*

Abstract

We assembled a de novo transcriptome of short-read Illumina RNA-Seq data generated from telencephalon and diencephalon tissue samples from the Kentish plover, *Charadrius alexandrinus*. This is a species of considerable interest in behavioural ecology for its highly variable mating system and parental behaviour, but it lacks genomic resources and is evolutionarily distant from the few available avian draft genome sequences. We assembled and identified over 21 000 transcript contigs with significant expression in our samples, showing high homology to exonic sequences in avian draft genomes. From these, we identified >31 000 high-quality SNPs and > 2500 simple sequence repeats (SSRs). We also analysed expression patterns in our data to identify potential candidate genes related to differences in male and female behaviour, identifying over 200 nonoverlapping putative autosomal transcripts that show significant expression differences between males and females. Gene ontology analysis revealed that female-biased transcripts were significantly enriched for cerebral functions related to learning, cognition and memory, and male-biased transcripts were mostly enriched for terms related to neural function such as neuron projection and synapses. This data set provides one of the first de novo transcriptome assemblies from non-normalized short-read next-generation data and outlines an effective strategy for measuring sequence and expression variability simultaneously without the aid of a reference genome.

Keywords: behavioural transcriptomics, de novo transcriptome assembly, neurogenomics, sex-biased gene expression, SNP discovery

Received 13 November 2012; revision received 24 February 2013; accepted 25 February 2013

Introduction

The development of next-generation sequencing methods, which make possible the generation of large quantities of relatively inexpensive genetic data, has the potential to bring genome-level analysis to nonmodel organisms of ecological and evolutionary interest (Rokas & Abbot 2009; Stapley *et al.* 2010; Ekblom & Galindo 2011). Despite these advances, there is still a significant barrier to implementing genome-level analyses to organisms that lack annotated draft genomes. With current technologies, researchers face a choice between three approaches: (i) investing considerable time and effort in assembling and annotating a draft genome of their organism of interest; (ii) using long-read next-generation approaches to sequence normalized transcriptomes to build a library of coding sequence and nucleotide variation (Kunstner *et al.* 2011; Santure *et al.* 2011); (iii) sequencing unmapped nucleotide variation on a large-scale (Miller *et al.* 2007; Hohenlohe *et al.* 2010). The first two approaches require extensive investment before the ecological or evolutionary questions can even be addressed. Although the third approach is more direct, it is anonymous with regard to genomic region without an available reference genome, making it difficult to identify the functional nature of single nucleotide polymorphic (SNP) variation, as well as to pinpoint the physical position of variable loci for further fine-scale analysis.

Correspondence: Hooman K. Moghadam, Fax: + 30-2810-337870; E-mail: hkm@hcmr.gr

A fourth option has recently become available and involves assembling short-read data generated from mRNA (RNA-Seq data) into a de novo transcriptome (Grabherr *et al.* 2011; Zhao *et al.* 2011). Previously, RNA-Seq data had to be mapped to a reference genome; however, advances in both molecular and bioinformatics technologies have now made it possible to bypass this step and assemble RNA-Seq data into transcript contigs de novo (Martin & Wang 2011; Vijay *et al.* 2012). Although there are drawbacks to this approach, de novo transcriptomes serve several simultaneous purposes: they reveal coding sequence for all expressed genes, measure gene expression level if a non-normalized mRNA library is used and identify nucleotide and protein isoform variation. This means that a single data set can be used for many different types of ecological and evolutionary analysis without initial investment in reference genome construction. This approach has an added advantage that it captures all genes expressed within the tissue sampled, some of which can be missed due to incomplete annotation or assembly in traditional genome-mapping approaches.

The Kentish plover (*Charadrius alexandrinus*) is a shorebird of ecological interest due to the sexual conflict between males and females over parental care that can lead to brood desertion by the male or the female parent (Székely *et al.* 2007). The mating system of this species is highly variable: polygyny, polyandry and monogamy often coexist in a single population (Székely & Lessells 1993; Fraga & Amat 1996). The effects of various ecological, life history and population demography variables have been investigated in a series of experiments and observational studies (Fraga & Amat 1996; Székely & Cuthill 2000; AlRashidi *et al.* 2011; Kosztolányi *et al.* 2011). The species is therefore of evolutionary, ecological and conservation interest. However, plovers and other shorebirds (e.g. sandpipers, jacanas, and oystercatchers) are evolutionarily equidistant from the main avian reference genomes, zebra finch and chicken, in the passeriformes and galliformes (Hillier *et al.* 2004; Dalloul *et al.* 2010; Warren *et al.* 2010), limiting the utility of these available avian genomic resources for the study of these ecological model species.

Here, we assembled the de novo neurotranscriptome from RNA-Seq data derived from the telencephalon and diencephalon of breeding adult male and female plovers to test the efficacy of this approach in nonmodel organisms, to build a database of plover exonic sequences and nucleotide variations, as well as to determine expression differences in the brain that might explain the neurogenetic basis of sex-specific social traits (Donaldson & Young 2008; Robinson *et al.* 2008).

Materials and methods

Sample collection and preparation, data quality filtering

Samples from six female and six male breeding adults were collected on 15–17 May, 2010, from the Tachart Estuary and Oued Gharifa, Morocco (35°34′25″N, 5°59′31″W). Prior to sample collection, all birds were incubating eggs and were caught on the nest with funnel traps. After the biometric measurements were taken (see http://www.bath.ac.uk/bio-sci/biodiversity-lab/pdfs/KP_Field_Guide_v3.pdf), telencephalon and diencephalon brain tissue were removed and stored in RNAlater. Time between trapping and sample collection was approximately 5–10 min, and brain removal and tissue homogenisation took up to an additional 5 min. Although we attempted to keep samples as cold as possible during field collections, we were unable to maintain temperatures <4 °C for most of the time in the field, and we therefore changed RNAlater fluid for each sample after 3 days. This allowed sufficient time for samples to dehydrate, but RNAlater preservation had not yet started to break down. Samples were stored at −80 °C on returning to the lab, and total RNA was later extracted using the Qiagen RNAeasy Lipid Kit. Despite several days of storage at >4 °C in the field, all RNA samples passed strict quality control criteria.

mRNA libraries and RNA-Seq samples were prepared by The Genome Analysis Centre (TGAC) in Norwich, U.K. using standard methods. Each of six samples for each sex was run on independent Illumina Genome Analyser II or Illumina HiSeq lanes as paired-end 80 bp reads with an approximate insert size of 230 bp, resulting in on average 65 million paired-end (PE) reads per individual. The quality of the sequence data was first assessed with FastQC (version 0.10.0; http://www.bioinformatics.babraham.ac.uk/projects/fastqc), and the reads were trimmed to 55 bp using PRINSEQ (Schmieder & Edwards 2011) to retain only the highest quality nucleotides for the de novo assembly. Trimmomatic (Lohse *et al.* 2012) along with the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit) were used for filtering putative rRNA sequences and reads contaminated with Illumina adapters, resulting in approximately 700 million clean reads.

Transcriptome assembly

The transcriptome was assembled using Trinity, version 2012-01-25 (Grabherr *et al.* 2011) with the default parameter settings. Reads from all 12 individuals were combined into a single transcriptome assembly, which allows for orthology determination. To overcome the challenges associated with high computing requirements necessary for de novo assembly of large numbers of sequence reads, we performed a two-stage assembly process. First, we assembled a subset of 50 million randomly selected paired-end reads and mapped all the sequences from the 12 individuals against these contigs. We then selected only those reads that had failed to align to any transcript and pooled these unmapped reads with the original 50 million sequences and performed a second transcriptome assembly. This two-step approach substantially reduced the computational memory usage, as many redundant reads were removed from the assembly process (Fig. S1).

In the final Trinity assembly, we recovered 569 987 transcripts. However, considering the extreme read depth of our dataset, we expected that many of these reads to represent transcriptional mistakes or rare variants (e.g. sequencing errors, chimeras, introns, noncoding sequences). To remove these from our transcriptome, we filtered out transcripts with expression support of less than four reads per million mappable reads in at least three males or three females, an approach that we have previously implemented to eliminate low-level expression noise (Table S1) (Harrison *et al.* 2012).

To differentiate endogenous plover transcripts from exogenous contigs that could represent bacterial, viral or environmental contamination, we BLASTed those transcripts that passed our expression profile filtering against the NCBI nonredundant protein and nucleotide databases, using the standalone BLAST tools (version 2.2.25) (Altschul *et al.* 1990) with a cut-off *e*-value of 10^{-6} and 10^{-10}, respectively. We also used a cut-off *e*-value of 10^{-10} and searched for homologous sequences in the chicken (release 13) and zebra finch (release 1) expressed sequence transcripts (EST) (http://compbio.dfci.harvard.edu/tgi/).

Next, through reciprocal best blast hits with a cut-off of *e*-value of 10^{-10}, we identified putative 1:1 orthologous sequences with the annotated genomes of chicken (version WUGSC2.1) and zebra finch (WUSTL3.2.4) to infer their likely chromosomal locations on the shorebird genome. For all the downstream analyses, we only utilized the information from these latter putative 1:1 orthologous transcripts.

Gene expression analysis

To assess the gene expression profiles, reads from each individual were mapped against the contigs of putative avian origin and expression abundances were quantified using the default settings of RSEM version 1.1.13 (Li & Dewey 2011). The global patterns of gene expression among individuals were investigated by hierarchical clustering of the data using Euclidean distance with complete linkage, as implemented in Cluster 3.0 (de Hoon *et al.* 2004) and visualized by TreeView (v.1.1.6; Saldanha 2004).

We inferred putative synteny to differentiate Z versus autosomal genes based on 1:1 orthology with the chicken and zebra finch draft genomes, with genes that were Z-linked in both reference genomes assumed to be Z-linked in our species. To validate this approach, we used the fact that all birds so far assessed show incomplete Z chromosome dosage compensation (Itoh *et al.* 2007; Mank & Ellegren 2009; Wolf & Bryk 2011), resulting in an overall male bias for Z-linked genes. We therefore tested the overall veracity of our chromosomal predictions by calculating average $\log_2$ male/female expression for all inferred Z-linked and autosomal contigs. We also tested the assumption of Z linkage by identifying male and female SNPs on putative Z-linked genes, as female heterozygosity is not expected, given their single copy of the Z chromosome.

Putative autosomal and Z-linked genes were then clustered separately due to the unique sex-biased profile of Z-linked genes. The reliability of the inferred trees were tested by bootstrap resampling (1000 replicates) of the expression values using Pvclust (Suzuki & Shimodaira 2006). The abundance estimates were further normalized and investigated for differential patterns of gene expression between sexes using the R Bioconductor package, DESeq (Anders & Huber 2010). A gene was assigned to have a differential pattern of expression if we observed >2-fold difference in expression between the sexes with a $P < 0.05$ after correcting for multiple testing (Benjamini & Hochberg 1995) (Fig. S1).

Functional annotations Gene Ontology (GO) terms were assigned by BLAST2GO software (Conesa & Gotz 2008) and further mapped to more generic terms using GO-Slim (http://www.geneontology.org/GO.slims.shtml). Enrichment in GO terms for genes with sex-biased patterns of expression was examined by Ontologizer (Bauer *et al.* 2008) and using the orthologous mouse GO database (Ashburner *et al.* 2000). Significance was assessed using Fisher's exact test and adjusted for multiple comparisons using the Benjamini–Hochberg method (Benjamini & Hochberg 1995).

Single Nucleotide Polymorphism (SNP) and Simple Sequence Repeat (SSR) identification

We first used CD-HIT-EST (Li & Godzik 2006) to remove redundant sequences and retained only the longest transcripts. The aligned sequence data were screened for single nucleotide polymorphic sites using freebayes (version 0.8.7; http://bioinformatics.bc.edu/marthlab/FreeBayes) for the putative autosomal genes as well as the Z-linked genes in males. For a site to be regarded as polymorphic, we adjusted for the read depth and set the threshold of minimum coverage of 10 reads where at least 35% of the unique sequences, with quality scores greater than 30 and a minimum distance of 25 bp between SNPs, had to support the alternative allele. We also assessed all the putative avian contigs for their simple sequence repeat content using Tandem Repeat Finder (Benson 1999) and MIcroSAtellite identification tool (MISA; Thiel *et al.* 2003). In particular, the contigs were screened for motifs of di-, tri-, tetra-, penta- and hexa-nucleotides for a minimum of 6, 5, 5, 5 and 5 repeats, respectively.

Results

Sample collection, Illumina sequencing and de novo assembly

Telencephalon and diencephalon tissue samples from six breeding pairs of males and females, each sequenced on a single lane of Illumina, resulted in more than 770 million 80 bp paired-end (PE) reads in total. These were assembled using Trinity, and our de novo assembly of the transcriptome yielded 569 987 contigs. After filtering out lowly expressed transcripts by setting a threshold of at least four reads per million mappable reads in at least three individuals of either sex (Table S1), we retained 25 595 transcripts, representing endogenous (plover) and exogenous (infectious agents and environmental contamination) expressed genes and their isoforms. We expect that the vast number of minimally expressed contigs that fall below our threshold to largely represent transcriptional errors such as intron expression, intraspecific variation between individuals, exonic chimeras (hybrids of two or more sequences), sequencing errors or other forms of mis-expression, sequencing or current assembly shortcomings. However, the detection and assembly of these lowly expressed contigs also indicate that our sequencing depth approached, if not exceeded, saturation and that additional read depth would not provide additional information.

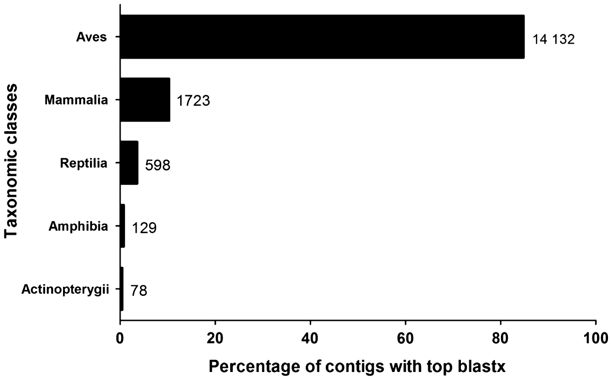

Fig. 1 Distribution of species taxonomic classes based on BLASTX top hits. The values in front of each column represent the number of contigs with the top BLASTX hit identified for each class.

Annotation of the assembly

We next focused on characterizing the genomic properties of the expressed contigs. The length of the expressed assembled sequences ranged from 100–14 356 bp with an average size of 1911 bp and a median size of 1507 bp (N50: 2509 bp, N90: 1012 bp) (Fig. S2). To annotate the assembly, we first performed a BLASTX search against the nonredundant NCBI protein database (nr) using a minimum threshold *e*-value of 10^{-6}. Approximately 16 660 transcripts had at least one significant hit, and 85% of these (14 132 contigs) showed a top BLAST hit against sequences from chicken, zebra finch or turkey (Fig. 1). On average, contigs without significant expression tended to have a smaller length (N50: 1633 bp, N90: 714 bp, median: 1077 bp), and this may be because smaller contigs were more probably to represent chimeras or assembly errors. Those contigs that returned a significant BLAST hit, but did not have a top hit in Aves showed similarity to a wide variety of taxa. Interestingly, many top hits were to other vertebrates, suggesting that these contigs represent endogenous plover transcripts rather than environmental or infectious contamination; however, because we could not be sure of the putative genomic location of these contigs, and specifically whether they were probably Z-linked, they were discarded from further analysis.

We also performed a BLAST homology search against the NCBI nucleotide database (nt) as well as the chicken and zebra finch TIGR expressed sequences (http://compbio.dfci.harvard.edu/tgi/) to identify orthologous regions to putative noncoding in addition to the coding transcripts. Setting a cut-off e-value of 10^{-10}, 21 063 contigs (including about 19 271 1:1 orthologs with chicken and zebra finch) showed significant BLASTn hits against the reference avian genomes and ESTs. These set of contigs, which also comprised of all previously identified transcripts through BLASTX search, constitute sequences with putative avian origins. The average GC content of these sequences is 47%, which is comparable with the GC content observed in the cDNA databases for zebra finch (50%, taeGut3.2.4.69) and chicken (49%, WASHUC2.69). These contigs mapped to 8963 and 9247 1:1 Ensembl annotated orthologous genes in the chicken and the zebra finch reference genomes, respectively, providing us with a conservative estimate of the number of genes that are expressed in our data set.

Patterns of gene expression between sexes

We tested our inference of chromosomal locations based on the chicken and zebra finch genome syntenies, considering the fact that birds lack complete Z chromosome dosage compensation (Itoh *et al.* 2007; Mank & Ellegren 2009; Wolf & Bryk 2011). Because of this, we expect the average expression of the putative Z-linked genes to show an overall male-biased expression pattern. Accordingly, the $\log_2$ male/female expression ratio was estimated to be 0.47 for the inferred Z-linked transcripts and equally expressed in both sexes (statistically identical to 0) for the putative autosomal contigs. The reduced SNP heterozygosity for females compared with males on putative Z-linked genes (explained below) also supports our use of synteny to infer Z-linked genes in plover.

To investigate the global gene expression profiles between males and females, we performed cluster analysis on the abundance data estimated for contigs with putative avian origin. We found an overall higher correlation between individuals within each sex in their gene expression landscape for both autosomal as well as the sex-linked genes. These correlations are generally supported by high bootstrap values (Fig. S3).

Also searching the data for signatures of differentially expressed genes between males and females, we identified a total of 176 female-biased and 255 male-biased transcripts that exhibit significant patterns of expression between the two sexes (>2-fold in expression and an adjusted *P* value <0.05; Fig. 2). Of these, 59 contigs showed expression only in females (female limited), and no contigs showed male-limited expression. Sex-biased and sex-limited contigs showed different patterns of genome distribution between males and females on the basis of their chromosomal mapping to the chicken and zebra finch genomes. In particular, 7% of the female-limited contigs showed highest affinity with the sequences on the W chromosome in chicken, while no male-biased contig had a significant BLAST hit against this chromosome (Fig. 3). Interestingly, although about 50% of the male- and female-biased contigs mapped to the Z chromosome or to the unmapped contig scaffolds (designated 'Un' in Ensembl), many of these sequences have expression that is only limited to females, suggesting their possible W-linkage, while no transcript showed male-limited expression. Of the 59 female-limited contigs, 41 mapped to the Z chromosome, nine to the autosomes and the other nine transcripts to either the W chromosome or the unmapped contigs of the Ensembl chicken assembly (WUGSC2.1). Such findings suggest that either the true chicken orthologs might be located on the W chromosome, but this information is currently missing from the chicken reference assembly that the W copy has been lost in the chicken lineage that genes might have relocated to the W chromosome in shorebirds, or some of these genes are present in both sexes but have expression which is only limited to females. Of the 176 female-biased transcripts (165 nonoverlapping), 12 mapped to the chicken W chromosome, 77 to the Z chromosome, 11 to unmapped contigs and 76 to autosomes. Of the 255 male-biased contigs (251 nonoverlapping), none mapped to the W chromosome, 123 mapped to the Z chromosome, 4 to unmapped contigs and 128 to chicken autosomes.

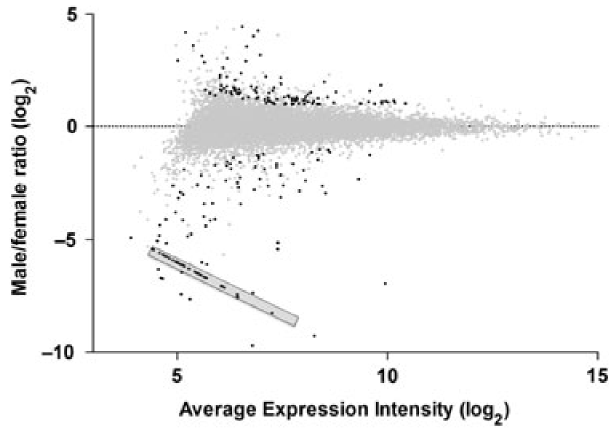

Fig. 2 $\log_2$ male/female gene expression ratios plotted against the average gene expression intensity ($\log_2$). Biased genes are shown in black, and the un-biased genes are in grey. The grey box highlights contigs with expression limited to only females. It should be noted that for these transcripts, only the $\log_2$ expression in the females is shown.

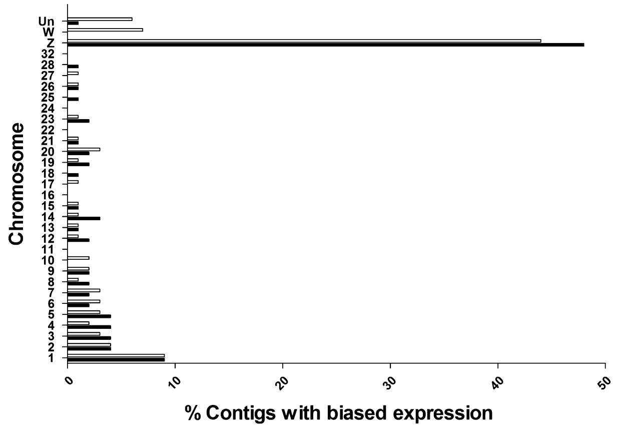

Fig. 3 Chromosomal locations of contigs with biased patterns of gene expression mapped to the chicken reference genome. Black bars represent male-biased contigs and white bars show the frequency distribution of the female-biased sequences.

Our deep sequencing strategy also allowed us to test the possibility of the Z chromosome in *C. alexandrinus* to contain the male hypermethylated (MHM) ortholog. We first excluded all the putative W-linked transcripts, particularly those with female-limited expression that had been mapped to the Z chromosome and then investigated the male/female expression ratios for the remainder of loci along the length of the chicken Z chromosome. We also compared the expression of the putative plover Z-linked genes with the pattern observed from available RNA expression data obtained from the embryonic stage day 19 in male and female replicates of red jungle fowl (Moghadam *et al*. 2012). The MHM locus is located at approximately 26 Mb on the chicken Z chromosome (Teranishi *et al*. 2001; Melamed & Arnold 2007), where we observed reduced male/female expression ratios for genes which are in close proximity to the MHM region in both plover and chicken (Fig. S4).

We next assessed the expression characteristics of the sex-biased autosomal genes where we first excluded all the putative sex-linked loci. We found the average magnitude of expression in the autosomal female-biased contigs to be significantly greater than the male-biased genes (average male/female ratio for female biased: −2.56 and male biased: 1.69; *t*-test $P < 0.0005$). Females also showed more variation in their expression relative to the male-biased expression (*F*-test, $P < 0.0001$). This pattern remained consistent even after excluding all the putative autosomal female-limited genes.

Gene Ontology (GO) enrichment analysis indicates that genes with female-biased expression are significantly enriched for genes associated with learning and memory (GO:0007611, GO:0007612), cognition (GO:0050890), circadian rhythm (GO:0007623) and behaviour and response to fear (GO:0001662) (Table S2). On the other hand, for the male-biased genes, we identified a wider range of gene ontology terms, with the greatest enrichment occurring within neuron projection (GO:0043005), regulation of multicellular organismal process (GO:0051239) and synapse part (GO:0044456) (Table S3).

Single nucleotide polymorphism and simple sequence repeat identification

Screening the sequence alignments for each individual, we identified a large number of single nucleotide polymorphic markers. Following our conservative approach where we excluded the polymorphic sites with low allele frequency (<17%) (Gao *et al*. 2012), we obtained a total of 31 729 high-quality autosomal and 1445 Z-linked common SNPs that are distributed in reads with putative avian origins. We also assessed the level of heterozygosity for the putative Z-linked transcripts of the pooled replicates between sexes and found that females have approximately a three-fold lower number of polymorphic sites among their replicates compared with the males, lending further weight to our synteny-based inference of Z-linked genes.

Considering the total length of the contig assemblies, we expect an average density of one nucleotide variation per thousand base pairs of the coding sequence data for either the autosomal or the Z-linked genes. This estimate is lower than the frequency of sequence polymorphism reported in the comparison between the red jungle fowl and the broiler chicken breeds within their coding regions (2.1 SNP/kb) (Wong *et al*. 2004), mainly due to our more stringent flitering of the minor alleles and much higher threshold for the minimum read depth. We also identified 2532 and 97 simple sequence repeats that are distributed across 2222 and 84 autosomal and Z-linked transcripts, respectively. As expected, the tri-nucleotides (63%, with the highest occurrence of the following repeats: GGC, TCC, CCG) followed by di-nucleotides (34%) had the largest number of microsatellite repeats (Fig. S5).

Discussion

Here, we report the first transcriptome assembly for the Kentish plover, a species of considerable interest for conservation and behavioural ecology. Unlike many previous transcriptome assemblies on nonmodel species (e.g. Fraser *et al*. 2011; Santure *et al*. 2011), we used non-normalized mRNA libraries which also make it possible to assess expression differences along with the EST sequence. Our data yielded >550 000 contigs however, the vast majority of these transcripts showed very low levels of expression and only slightly more than 25 000 contigs showed significant expression in more than half the individuals of either sex. Due to the nature of our downstream analysis, which heavily relies on accurate assessment of the expression patterns of true biological entities, we applied a stringent filtering scheme. However, one might expect that the contigs below the defined threshold to more often represent transcriptional errors rather than true contigs. This suggests that our read depth is well into the data saturation curve and therefore covers nearly all expressed sequences and that our contigs correspond to full-length or nearly full-length coding regions in the majority of cases.

Characterization of the transcriptome data involved high throughput sequencing of the mRNA and was followed by careful assessment of the sequence reads and systematic analyses and filtering of the assembled transcripts. Such a stepwise process resulted in the identification of 21 063 high-quality coding, as well as some noncoding transcripts that showed significant levels of expression in at least half of the individuals in either sex and had a significant BLAST hit against one of the avian reference genomes or expressed transcript databases. This transcriptome assemblage represents a substantial amount of sequence data in a species for which there were relatively few pre-existing genomic resources and can offer immediate applications in conservation, population genetics and functional studies. On the other hand, approximately one-fifth of the expressed contigs did not show any significant similarity match against chicken, zebra finch or turkey genomes, were more closely associated with other nonvertebrate sequences or did not return any significant hit and were classified as unknown. These transcripts may contain some plover genes that have not yet been annotated in zebra finch or chicken, but may also represent exogenous contamination from the collection point, or misassembled contigs.

Various studies have suggested that relocation of genes between the avian Z chromosome and the autosomal chromosomes is very rare across the avian phylogeny, and because of this, karyotypes and syntenies are highly conserved during the 100 million years of avian evolution (Nanda *et al.* 2008; Ellegren 2010; Wolf & Bryk 2011). This genomic characteristic is of great importance particularly in comparative studies, as it allows inference of the chromosomal locations of the orthologous loci on the basis of other phylogenetically distinct avian species. Here, we inferred the chromosomal locations of the expressed contigs in the *C. alexandrinus* genome on the basis of orthology information in chicken and zebra finch. We verified this based on both expression and sequence variation. Studies on a diverse range of avian species show incomplete Z dosage compensation (Itoh *et al.* 2007; Mank & Ellegren 2009; Warren *et al.* 2010; Wolf & Bryk 2011), as this means that our inferred Z-linked genes should show on average male-biased expression. Therefore, it is not surprising that about 50% of all the male-biased transcripts identified in this study possess an orthologous counterpart on the chicken and zebra finch Z chromosomes. In males, these putative Z-linked transcripts have on average 1.38 times higher expression than their counterparts in females. These values are comparable with the estimates reported in other bird species and are consistent with the idea of incomplete dosage compensation, as the Z-linked expression ratios between males and females are generally below 1.5 (e.g. Ellegren *et al.* 2007; Itoh *et al.* 2007; Melamed & Arnold 2007; Wolf & Bryk 2011). This combined with the low level of female heterozygosity for the putative Z-linked genes, as expected for female heterogametic sex chromosomes, indicates that synteny-based inference of Z chromosome linkage is effective in plovers and other avian species.

On the other hand, although about 50% of the female-biased transcripts also mapped to the Z chromosome in chicken or zebra finch, many of these contigs have female-limited expression (as opposed genes expressed in both sexes but to a greater degree in females). Because relatively few loci present in both sexes show completely sex-limited expression (Moghadam *et al.* 2012), it may be that a large proportion of the female-limited loci are located on the W chromosome in the Kentish plover. Because the avian Z and W chromosomes originally evolved from the same pair of autosomes, the majority of W genes in birds have a Z chromosome gametolog (Fridolfsson *et al.* 1998; Wright *et al.* 2012). The zebra finch genome sequence was performed on a male and therefore the assembly completely lacks any W genes (Warren *et al.* 2010), and the chicken W chromosome assembly is in large part incomplete (Moghadam *et al.* 2012), and in these cases, we would expect plover W genes to show the greatest similarity to the paralogous Z locus. However, we cannot exclude the possibility that some of these genes that have lost their W ortholog in the chicken lineage, have been relocated to the W chromosome in shorebirds or simply have a female-limited expression despite being located on chromosomes present in both sexes. Nonetheless, collectively all these loci constitute probably candidates that contribute to certain aspects of sexual dimorphism such as fitness and reproduction given their expression in females.

Gene function analysis

One of the main objectives of the current study was to identify candidate genes that according to their expression profiles and in association with their chromosomal locations and their predicted genomic functions might play a key role in the observed behavioural variation between male and female Kentish plovers. To achieve this goal, we obtained gene expression data from the telencephalon and diencephalon tissues of the replicate male and female individuals, as it has been shown that in vertebrate animal models, these regions of the brain are of primary importance in regulation of cognitive tasks such as attention, communication, working memory and learning (e.g. Glickstein 2007; Strick *et al.* 2009).

We found a total of 204 transcripts, mapping to approximately 165 Ensembl annotated autosomal genes in chicken and zebra finch that exhibit significant expression differences between females and males. Analysis of the overrepresented GO terms in the female-biased genes revealed enrichment for genes associated with terms such as learning and memory, cognition and behaviour. Our list of female-biased genes includes candidates such as dopamine receptor D1A (*Drd1a*), voltage-dependent anion channel 1 (*Vdac1*), McKusick–Kaufman syndrome (*Mkks*), glutamate receptor-metabotropic 5 (*Grm5*), activity-dependent neuroprotective protein (*Adnp*) and period homologs 2 and 3 (*Per2* and *Per3*) (Table S2). Previous studies, mainly on the basis of human or mouse mutant models, have proposed that these genes are expressed in the nervous system and play an important role in cognition, learning and behavioural responses (e.g. Weeber *et al.* 2002; Pinhasov *et al.* 2003; Tran *et al.* 2008; Cozzoli *et al.* 2009). Therefore, these genes are probably candidates which may contribute to the underlying molecular architecture of some of the observed behavioural characteristics of the female Kentish plovers and as such are an interesting target for future functional studies.

On the other hand, the male-biased transcripts were enriched for a wider range of term ontologies (Table S3) but mainly for genes that function in various neurological processes. In particular, genes such as discs, large homolog 3 (*Dlg3*), which is required for normal spatial learning in mouse (Cuthbert *et al*. 2007), oxytocin (*Oxt*), where the mutant male mice fails to develop social memory and are less aggressive (Young *et al*. 1997; Lim & Young 2006) or adenylate cyclase activating polypeptide 1 (*Adcyap1*), a gene that encodes a neuropeptide with neurotransmission modulating activity known as PACAP are only a few examples. In the rodent models, knock-out mutants for the latter gene show schizophrenia-like symptoms in addition to other abnormal behaviours such as elevated locomotor activity and abnormal social development (Hashimoto *et al*. 2007).

Concluding remarks

In conclusion, this study presents the first major expressed sequence resources for Kentish plover and provides a glimpse into the gene expression landscape from the brains of an important ecological model species. Our approach in comparing the gene expression data between males and females proved to be sensitive enough to identify subtle expression differences between males and females, identifying many genes that have been previously implicated as cognitive regulators. Our deep sequencing strategy, from different individuals and both sexes, provided us with rich genomic resources to discover a large number of common SNPs and SSR genetic markers. We expect that the high depth of coverage along with our stringent filtering criteria to have significantly reduced the potential impacts of sequencing errors on the assembled transcripts, making the polymorphic data highly reliable.

Acknowledgements

Permit for fieldwork and tissue collection was provided by the Moroccan Ministry of Forestry and Environment (Decision No. 13 HCEFLCD/DLCDPN/DPRN/CFF), and we thank Mohammed Ammezian and his assistant in fieldwork. Sequencing was conducted at The Genome Analysis Centre (TGAC), and we are grateful to David Swarbreck and Nizar Drou for the initial data analysis. This project was funded by The Genome Analysis Centre Capacity and Capability Challenge Program and the Biotechnology and Biological Sciences Research Council and European Research Council grant agreement 260233 (to JEM). The authors acknowledge the use of the UCL Unity SMP Facility, and associated support services, in the completion of this work.

References

AlRashidi M, Kosztolányi A, Shobrak M, Kupper C, Székely T (2011) Parental cooperation in an extreme hot environment: natural behaviour and experimental evidence. *Animal Behaviour*, **82**, 235–243.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

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

Ashburner M, Ball CA, Blake JA *et al*. (2000) Gene ontology: tool for the unification of biology The Gene Ontology Consortium. *Nature Genetics*, **25**, 25–29.

Bauer S, Grossmann S, Vingron M, Robinson PN (2008) Ontologizer 2.0-a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics*, **24**, 1650–1651.

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 (Statistical Methodology)*, **57**, 289–300.

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, **27**, 573–580.

Conesa A, Gotz S (2008) Blast2GO: a comprehensive suite for functional analysis in plant genomics. *International Journal of Plant Genomics*, **2008**, 619832.

Cozzoli DK, Goulding SP, Zhang PW *et al*. (2009) Binge drinking upregulates accumbens mGluR5-Homer2-PI3K signaling: functional implications for alcoholism. *Journal of Neuroscience*, **29**, 8655–8668.

Cuthbert PC, Stanford LE, Coba MP *et al*. (2007) Synapse-associated protein 102/dlgh3 couples the NMDA receptor to specific plasticity pathways and learning strategies. *Journal of Neuroscience*, **27**, 2673–2682.

Dalloul RA, Long JA, Zimin AV *et al*. (2010) Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): genome assembly and analysis. *PLoS Biology*, **8**, e1000475.

Donaldson ZR, Young LJ (2008) Oxytocin, vasopressin, and the neurogenetics of sociality. *Science*, **322**, 900–904.

Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, **107**, 1–15.

Ellegren H (2010) Evolutionary stasis: the stable chromosomes of birds. *Trends in Ecology and Evolution*, **25**, 283–291.

Ellegren H, Hultin-Rosenberg L, Brunstrom B *et al*. (2007) Faced with inequality: chicken do not have a general dosage compensation of sex-linked genes. *BMC Biology*, **5**, 40.

Fraga RM, Amat JA (1996) Breeding biology of a Kentish Plover (*Charadrius alexandrinus*) population in an inland saline lake. *Ardeola*, **43**, 69–85.

Fraser BA, Weadick CJ, Janowitz I, Rodd FH, Hughes KA (2011) Sequencing and characterization of the guppy (*Poecilia reticulata*) transcriptome. *BMC Genomics*, **12**, 202.

Fridolfsson AK, Cheng H, Copeland NG *et al*. (1998) Evolution of the avian sex chromosomes from an ancestral pair of autosomes. *Proceedings of the National Academy of Sciences USA*, **95**, 8147–8152.

Gao Z, Luo W, Liu H *et al*. (2012) Transcriptome analysis and SSR/SNP markers information of the blunt snout bream (Megalobrama amblycephala). *PLoS ONE*, **7**, e42637.

Glickstein M (2007) What does the cerebellum really do? *Current Biology*, **17**, R824–R827.

Grabherr MG, Haas BJ, Yassour M *et al*. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **29**, 644–652.

Harrison PW, Mank JE, Wedell N (2012) Incomplete sex chromosome dosage compensation in the Indian meal moth, *Plodia interpunctella*, based on de novo transcriptome assembly. *Genome Biology and Evolution*, **4**, 1118–1126.

Hashimoto R, Hashimoto H, Shintani N *et al*. (2007) Pituitary adenylate cyclase-activating polypeptide is associated with schizophrenia. *Molecular Psychiatry*, **12**, 1026–1032.

Hillier LW, Miller W, Birney E *et al*. (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, **432**, 695–716.

Hohenlohe PA, Bassham S, Etter PD, *et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862.

de Hoon MJ, Imoto S, Nolan J, Miyano S (2004) Open source clustering software. *Bioinformatics*, **20**, 1453–1454.

Itoh Y, Melamed E, Yang X *et al.* (2007) Dosage compensation is less effective in birds than in mammals. *Journal of Biology*, **6**, 2.

Kosztolányi A, Barta Z, Kupper C, Székely T (2011) Persistence of an extreme male-biased adult sex ratio in a natural population of polyandrous bird. *Journal of Evolutionary Biology*, **24**, 1842–1846.

Kunstner A, Wolf JBW, Backstrom N *et al.* (2011) Comparative genomics based on massive parallel transcriptome sequencing reveals patterns of substitution and selection across 10 bird species. *Molecular Ecology*, **20**, 2871.

Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Lim MM, Young LJ (2006) Neuropeptidergic regulation of affiliative behavior and social bonding in animals. *Hormones and Behavior*, **50**, 506–517.

Lohse M, Bolger AM, Nagel A *et al.* (2012) RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research*, **40**, W622–W627.

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.

Martin JA, Wang Z (2011) Next-generation transcriptome assembly. *Nature Review*. *Genetics*, **12**, 671–682.

Melamed E, Arnold AP (2007) Regional differences in dosage compensation on the chicken Z chromosome. *Genome Biology*, **8**, R202.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, **17**, 240–248.

Moghadam HK, Pointer MA, Wright AE, Berlin S, Mank JE (2012) W chromosome expression responds to female-specific selection. *Proceedings of the National Academy of Sciences USA*, **190**, 8207–8211.

Nanda I, Schlegelmilch K, Haaf T, Schartl M, Schmid M (2008) Synteny conservation of the Z chromosome in 14 avian species (11 families) supports a role for Z dosage in avian sex determination. *Cytogenetic and Genome Research*, **122**, 150–156.

Pinhasov A, Mandel S, Torchinsky A *et al.* (2003) Activity-dependent neuroprotective protein: a novel gene essential for brain formation. *Brain Research Developmental Brain Research*, **144**, 83–90.

Robinson GE, Fernald RD, Clayton DF (2008) Genes and social behavior. *Science*, **322**, 896–900.

Rokas A, Abbot P (2009) Harnessing genomics for evolutionary insights. *Trends in Ecology and Evolution*, **24**, 192–200.

Saldanha AJ (2004) Java Treeview-extensible visualization of microarray data. *Bioinformatics*, **20**, 3246–3248.

Santure AW, Gratten J, Mossman JA, Sheldon BC, Slate J (2011) Characterisation of the transcriptome of a wild great tit *Parus major* population by next generation sequencing. *BMC Genomics*, **12**, 283.

Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**, 863–864.

Stapley J, Reger J, Feulner PGD *et al.* (2010) Adaptation genomics: the next generation. *Trends in Ecology and Evolution*, **25**, 705–712.

Strick PL, Dum RP, Fiez JA (2009) Cerebellum and nonmotor function. *The Annual Review of Neuroscience*, **32**, 413–434.

Suzuki R, Shimodaira H (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, **22**, 1540–1542.

Székely T, Cuthill IC (2000) Trade-off between mating opportunities and parental care: brood desertion by female Kentish plovers. *Proceedings of the Royal Society B*, **267**, 2087–2092.

Székely T, Lessells CM (1993) Mate change by Kentish Plovers *Charadrius alexandrinus*. *Ornis Scandinavica*, **24**, 317–322.

Székely T, Kosztolányi A, Kupper C, Thomas GH (2007) Sexual conflict over parental care: a case study of shorebirds. *Journal of Ornithology*, **148**, S211–S217.

Teranishi M, Shimada Y, Hori T *et al.* (2001) Transcripts of the MHM region on the chicken Z chromosome accumulate as non-coding RNA in the nucleus of female cells adjacent to the DMRT1 locus. *Chromosome Research*, **9**, 147–165.

Thiel T, Michalek W, Varshney RK, Graner A (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theoretical and Applied Genetics*, **106**, 411–422.

Tran AH, Uwano T, Kimura T *et al.* (2008) Dopamine D1 receptor modulates hippocampal representation plasticity to spatial novelty. *Journal of Neuroscience*, **28**, 13390–13400.

Vijay N, Poelstra JW, Kunstner A, Wolf JB (2012) 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.

Warren WC, Clayton DF, Ellegren H *et al.* (2010) The genome of a songbird. *Nature*, **464**, 757–762.

Weeber EJ, Levy M, Sampson MJ *et al.* (2002) The role of mitochondrial porins and the permeability transition pore in learning and synaptic plasticity. *Journal of Biological Chemistry*, **277**, 18891–18897.

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

Wong GK, Liu B, Wang J *et al.* (2004) A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. *Nature*, **432**, 717–722.

Wright AE, Moghadam HK, Mank JE (2012) Trade-off between selection for dosage compensation and masculinization on the avian Z chromosome. *Genetics*, **192**, 1434–1445.

Young LJ, Winslow JT, Wang Z *et al.* (1997) Gene targeting approaches to neuroendocrinology: oxytocin, maternal behavior, and affiliation. *Hormones and Behavior*, **31**, 221–231.

Zhao QY, Wang Y, Kong YM *et al.* (2011) Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics*, **12**, S2. doi: 10.1186/1471-2105-12-S14-S2.

T.S., J.E.M. contributed Experimental design. T.S., J.E.M., G.Z. contributed sample collection. J.E.M. supported laboratory work. H.K.M., P.W.H. contributed for the data analysis. H.K.M., P.W.H., T.S., J.E.M. wrote the manuscript.

Data Accessibility

Assembled contigs, read counts, SNP and SSR data reported in this study are accessible via DRYAD doi:10.5061/dryad.23vp5.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. A flowchart illustrating the main steps of our sequence processing, assembly, expression and variant call analysis.

Fig. S2. Length distribution of contigs with a significant pattern of gene expression.

Fig. S3. Hierarchical clustering of the gene expression data for the six males and six females investigated in this study for the A. autosomal and B. Z-linked genes. Clustering was carried out using Euclidean distance with complete linkage. The number at each node represents bootstrap support based on 1000 replicates. Values less than 50 are not shown.

Fig. S4. 2 Mb sliding average of the M/F ratios of expression in A. red jungle fowl and B. Kentish plover plotted against the median gene position along the chicken Z chromosome. The HMH locus is approximately at position 26 Mb.

Fig. S5. Count distribution of various types of nucleotide repeats throughout the autosomal contigs with putative avian origins.

Table S1. Examining the number of contigs that passed various filtering thresholds for different minimum number of reads per million mappable reads (PMMR) and minimum number of replicates (MNR) in either of sexes.

Table S2. Gene ontology categories enriched among female-biased transcripts.

Table S3. Gene ontology categories enriched among male-biased transcripts.