

RESEARCH ARTICLE

Open Access

The immune gene repertoire of an important viral reservoir, the Australian black flying fox

Anthony T Papenfuss^{1,2†}, Michelle L Baker^{3,4*†}, Zhi-Ping Feng^{1,5†}, Mary Tachedjian^{3†}, Gary Cramer³, Chris Cowled³, Justin Ng³, Vijaya Janardhana³, Hume E Field^{6,7} and Lin-Fa Wang^{3,6*}

Abstract

Background: Bats are the natural reservoir host for a range of emerging and re-emerging viruses, including SARS-like coronaviruses, Ebola viruses, henipaviruses and Rabies viruses. However, the mechanisms responsible for the control of viral replication in bats are not understood and there is little information available on any aspect of antiviral immunity in bats. Massively parallel sequencing of the bat transcriptome provides the opportunity for rapid gene discovery. Although the genomes of one megabat and one microbat have now been sequenced to low coverage, no transcriptomic datasets have been reported from any bat species. In this study, we describe the immune transcriptome of the Australian flying fox, *Pteropus alecto*, providing an important resource for identification of genes involved in a range of activities including antiviral immunity.

Results: Towards understanding the adaptations that have allowed bats to coexist with viruses, we have *de novo* assembled transcriptome sequence from immune tissues and stimulated cells from *P. alecto*. We identified about 18,600 genes involved in a broad range of activities with the most highly expressed genes involved in cell growth and maintenance, enzyme activity, cellular components and metabolism and energy pathways. 3.5% of the bat transcribed genes corresponded to immune genes and a total of about 500 immune genes were identified, providing an overview of both innate and adaptive immunity. A small proportion of transcripts found no match with annotated sequences in any of the public databases and may represent bat-specific transcripts.

Conclusions: This study represents the first reported bat transcriptome dataset and provides a survey of expressed bat genes that complement existing bat genomic data. In addition, these data provide insight into genes relevant to the antiviral responses of bats, and form a basis for examining the roles of these molecules in immune response to viral infection.

Background

Bats make up approximately 20% of the extant mammalian diversity and are the second most species rich mammalian lineage after rodents [1]. The order Chiroptera is divided into two suborders: the Megachiroptera and Microchiroptera. These two lineages are estimated to have diverged approximately 58 million years ago [2]. Megachiroptera consists of a single family, the old world fruit bats, while Microchiroptera includes 17 families of echolocating bats. Bats have a wide geographic distribution and exploit a

variety of environmental niches, being absent only from the polar regions. Bats are also hosts to numerous viruses, many of which are highly pathogenic to humans and other mammals yet appear to cause no clinical consequences in bats [3-8]. This group of mammals also shares a variety of unique characteristics that likely facilitate the persistence and spread of the viruses they carry. Highly social species, bats live at much higher densities than other mammals. They are the only mammals capable of powered flight and have long lifespans relative to their body size [9]. Despite their diversity, unique characteristics and role as natural reservoirs for viruses, bats are also the least studied of all mammalian taxa and there is little information available on antiviral immunity in any bat species.

Bats are the natural reservoir hosts of more than 80 viruses, with new viruses or viral sequences of bat origin

* Correspondence: Michelle.Baker@csiro.au; Linfa.Wang@csiro.au

†Equal contributors

³CSIRO Livestock Industries, Australian Animal Health Laboratory, PO Bag 24, Geelong, VIC 3220, Australia

⁴Center for Evolutionary and Theoretical Immunology, Department of Biology, The University of New Mexico, Albuquerque, NM 87131, USA
Full list of author information is available at the end of the article

being discovered each year [9,10]. RNA viruses account for the overwhelming majority of known bat viruses, many of which are among the most deadly known to man, including Ebola, Hendra, Nipah and SARS-like coronaviruses [9]. Many of these viruses, which cause severe morbidity and mortality in humans and other mammals, appear to cause no clinical diseases in bats under natural or experimental infection. The most studied example is the henipaviruses (Hendra and Nipah viruses) which are members of the family Paramyxoviridae. Nipah virus has a mortality rate of 40-90% in humans and close to 100% in experimental animal models (cats and hamsters). Yet, infection of *Pteropus vampyrus* (the natural reservoir host of Nipah virus in Malaysia) and *P. poliocephalus* (a related bat species native in Australia) by a high dose of Nipah virus, failed to result in clinical signs of disease [7,8,11]. Other examples of experimental infections of bats including Ebola Zaire, Japanese encephalitis and St. Louis encephalitis viruses have not resulted in any symptoms of disease despite the presence of viral RNA in tissues [3-6]. Experimental infections of *P. poliocephalus* with Nipah virus have demonstrated the presence of serum antibody and viral shedding in the absence of clinical symptoms of disease [11]. The only viruses that have been demonstrated to cause clinical symptoms of disease in bats are rabies virus and the closely related Australian bat lyssavirus [12,13]. However, results of experimental infections are inconsistent with only a small proportion of bats succumbing to infection, and rates of sero-conversion and virus recovery from tissues were reported to be very low [13]. The long co-evolutionary history of bats with viruses has likely resulted in the adaptation of the bats immune system to cope with viral infection. One hypothesis is that the innate immune system rapidly controls viral replication to very low levels that cause no clinical consequences to bats, but still results in viral shedding and subsequent spillover to other species. However, as little information currently exists on any aspect of bat immunology and few bat-specific reagents exist, this hypothesis remains untested.

Recent years have seen a surge in the availability of whole genome sequence data. Bats were among the organisms sequenced as part of the US National Institutes of Health (NIH)-funded Mammalian Genome Project. These genomic resources are an important step forward in identifying the genes that are involved in antiviral immunity in bats and in providing insights into other unique life history characteristics. There are currently two publicly available bat genome sequences: one from the megabat *P. vampyrus* and a second from the microbat *Myotis lucifugus*. Both bat genomes were initially sequenced to low coverage (2.6X for *P. vampyrus* and 1.7X for *M. lucifugus*, though a draft quality assembly of the *M. lucifugus* genome based on 7X coverage

sequencing is now available). Additionally, the annotations were predominantly based upon comparative data. Despite these shortcomings, these projects have an important role to play in revealing the mechanisms that have evolved to allow bats to remain asymptomatic to so many viral diseases.

In order to understand bat-virus interactions, we are developing the Australian black flying fox, *P. alecto*, as a model bat species. *P. alecto* belongs to the family Pteropodidae and is closely related to *P. vampyrus* [14]. These two species are reservoirs for a variety of closely related viruses, the most important of which include the henipaviruses, Hendra virus in *P. alecto* and Nipah virus in *P. vampyrus* [10]. A number of important resources have now been developed for *P. alecto*, including cell lines from a variety of tissues [15]. We have also begun to identify some of the genes involved in immune responses in this species and carry out functional studies in bat cells [16-21]. To begin to characterise the immune gene repertoire of *P. alecto*, we sequenced the transcriptome of bat immune tissues and mitogen-stimulated cells using the Illumina platform. To our knowledge, this study represents the first analysis of the transcriptome of any species of bat. Our analysis of the *P. alecto* transcriptome provides information on a variety of immune genes not previously identified in any bat species and represents an important starting point for examining the antiviral activity of these molecules.

Results and discussion

Overview of the bat transcriptome

Two separate transcriptomic datasets were generated and raw sequences from each database were submitted to the Sequence Read Archive [SRA: SRR350710.3 and SRR351237.2]. The first was obtained using total RNA extracted from a juvenile male flying fox thymus. Due to its role in central tolerance, the thymus expresses a large proportion of the proteome and therefore allows for the identification of a broad range of genes, including those involved in the immune response. To enrich for sequences corresponding to cytokines and innate immune genes, the second dataset was derived from pooled total RNA obtained from mitogen-stimulated spleen, white blood cells and lymph node and unstimulated thymus and bone marrow obtained from one pregnant female and one adult male flying fox. Cells were stimulated with lipopolysaccharide (LPS) and Ionomycin, which stimulate the production of pro-inflammatory cytokines; PolyIC, a TLR3 ligand; PHA, which triggers T cell activation and PMA, which activates T and B cells.

About 12.5 million 65 bp long reads were obtained from the thymus dataset, while 23.9 million 76 bp long reads were generated from the stimulated pooled sample. Prior to assembly, the raw reads were trimmed of low

quality sequence and polyA/T tails, uninformative strings of 'N' and primer/adaptor contaminants were cleaned. The filtered dataset consisted of 12,399,095 reads from the thymus (between 20-63 bp) and 22,577,294 reads from the stimulated pooled dataset (between 20-73 bp). The filtered reads were *de novo* assembled using the software packages velvet and oases. The resulting oases assemblies consisted of 247,909 contigs (N50 1244 bp) from the thymus and 313,641 contigs (N50 733 bp) from the pooled samples. The largest contigs in the thymus and pooled samples were 11.8 kb and 8.9 kb respectively, both of which correspond to the DNA-dependent protein kinase catalytic subunit (DNA-PKcs) which is represented by a 12.4 kb transcript in other species, including horse. For comparative purposes, an assembly using MIRA was also generated. Summary statistics from the velvet, oases and MIRA assemblies are listed in Additional file 1: Table S1. All subsequent analyses were performed using the oases assemblies.

To identify orthologues of known mammalian protein coding genes, the bat contigs were used to search the KEGG and NCBI non-redundant (NR) protein databases with BLASTX (E-value < 0.001). Of the 247,801 contigs longer than 50 bp in the thymus sequence assembly, about 46% matched annotated proteins in the NR database. For the pooled dataset, about 51% of the 313,528 loci matched proteins in NR. Similar results were obtained for both assembled libraries against the KEGG database.

Of the assembled thymus transcripts annotated using KEGG, 36% of all transcripts were more similar to horse sequences than to any other species, followed by dog (16%) and cow (12%) (Figure 1). Similar results were obtained for the pooled tissue dataset (not shown). This result is consistent with the now generally accepted view that bats belong within Laurasiatheria, which includes Carnivora, Cetartiodactyla (whales and even toed ungulates), Eulipotyphla (moles and shrews), Pholidota (scaly anteater) and Perissodactyla (odd toed ungulates) [22-27]. However, until recently, the phylogenetic relationships within Laurasiatheria have been controversial. Conflicting results have been reported using complete mitochondrial genome sequences to infer phylogenetic relationships with support for a sister relationship between Chiroptera and Fereungulata (Carnivora, Pholidota, Perissodactyla and Cetartiodactyla) or a relationship between Chiroptera and Eulipotyphla [28-30]. Analysis of the nuclear gene, protamine P1, as well as large genomic datasets, has provided evidence that bats are sister to a clade containing Perissodactyla, Carnivora, and Cetartiodactyla [31,32]. The volume of sequence data generated by transcriptome sequencing provides the opportunity for larger scale sequence comparisons than previously possible using the few full length bat genes available or by comparison with the limited whole genome sequence data. Our results

support the comparative analysis of retroposon loci which has also demonstrated that bats share a sister relationship with horses, forming a clade with Carnivora [27].

Alignment of contigs from the thymus and pooled datasets to the KEGG database identified 178,554 and 285,268 contigs respectively with homology to 16,863 and 16,927 unique human proteins. To explore gene function, Gene Ontology (GO) terms were used. Of contigs that matched proteins in the KEGG database, 86% were assigned GO terms and 78% could be mapped to GO slim terms using GO Term Mapper (Additional file 2: Figure S1). Genes with GO slim terms were further classified into twelve selected classes (Figure 2). The most abundant GO terms found in the thymus dataset were involved in cell growth and maintenance (16.8%), enzyme activity (14.8%), cellular components (14.3%) and metabolism and energy pathways (14.5%). Similar results were obtained for the pooled tissue dataset (data not shown). The GO classification demonstrates that a diverse range of genes were identified in each of our two datasets providing a broad survey of bat genes.

Immune-related transcripts

A goal of the present study was to identify immune transcripts, particularly those that may play a role in antiviral immunity. Only 3.5% of the bat transcribed genes from each of the datasets showed homology to genes associated with immune function. This represents about 500 different immune-related genes (Figure 2). The bat immune transcripts were further categorised using GO terms to annotate the transcripts into 40 immune categories. Represented in the datasets were genes involved in a broad range of immune activities with lymphocyte activation, cytokine production and T cell activation making up the largest proportions of immune transcripts (Figure 3). Using KEGG codes to identify immune genes, our data revealed 70 genes involved in Toll-like receptor (TLR) cascades, 50 genes involved in B cell activation, 79 involved in T cell activation, 72 involved in natural killer cell cytotoxicity and 41 involved in antigen presentation. Additional immune genes not identified in the KEGG database were obtained by searching sequences from the NR database. The sequences of all genes described in the text are provided in the Additional file 3.

Innate immunity

One hypothesis for the ability of bats to resist the pathological effects of viral infection is that they are able to rapidly control viral replication early in the immune response through innate antiviral mechanisms. The bat transcriptome contained representatives of a variety of immune genes including pattern recognition receptors, interferons, interferon stimulated genes and natural killer cell receptors.

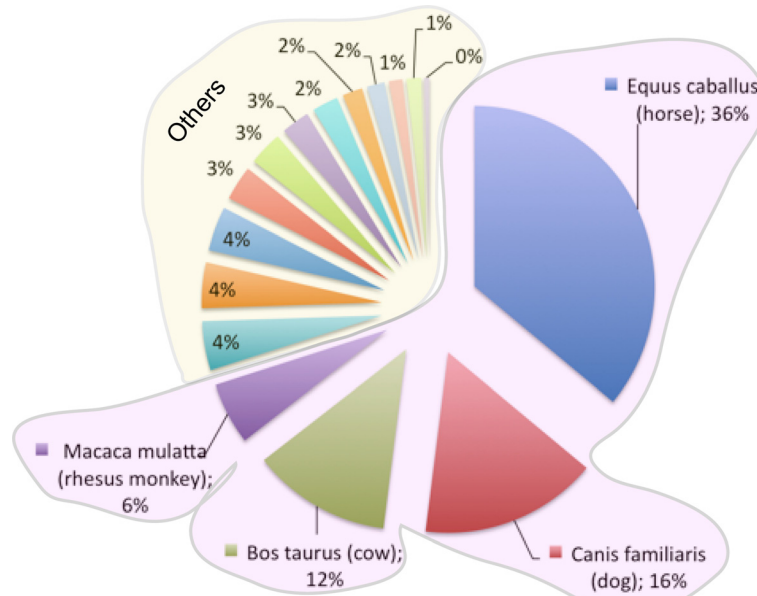


Figure 1 Bat transcripts are most closely related to horse genes. Summary of 183,788 hits from the thymus dataset to annotated genes from other species in the KEGG databases. Others from left to right in the pie chart include *Homo sapiens* (human), *Pan troglodytes* (chimpanzee), *Sus scrofa* (pig), *Monodelphis domestica* (opossum), *Rattus norvegicus* (rat), *Mus musculus* (mouse), *Taeniopygia guttata* (zebra finch), *Ornithorhynchus anatinus* (platypus), *Gallus gallus* (chicken), *Danio rerio* (zebrafish), *Xenopus tropicalis* (western clawed frog), *Xenopus laevis* (African clawed frog).

Pattern recognition receptors

Pattern recognition receptors (PRR) including TLRs, RIG-I like helicases (RLHs) and nucleotide oligomerisation domain (NOD) like receptors (NLRs) recognise conserved molecular patterns associated with a broad range of pathogens. Both TLRs and RLHs initiate signalling pathways that result in the induction of similar immune and inflammatory responses but are expressed

in different locations within the cell and differ in the pathogens they recognise. TLRs are transmembrane proteins expressed by the plasma membrane or endosome and recognise a broad range of pathogens including viruses, bacteria and fungi. Of eleven previously identified *P. alecto* TLR genes [18], only TLR5 was absent from the oases assemblies, however it was present in the MIRA assembly, which used a lower coverage cut-off and is useful

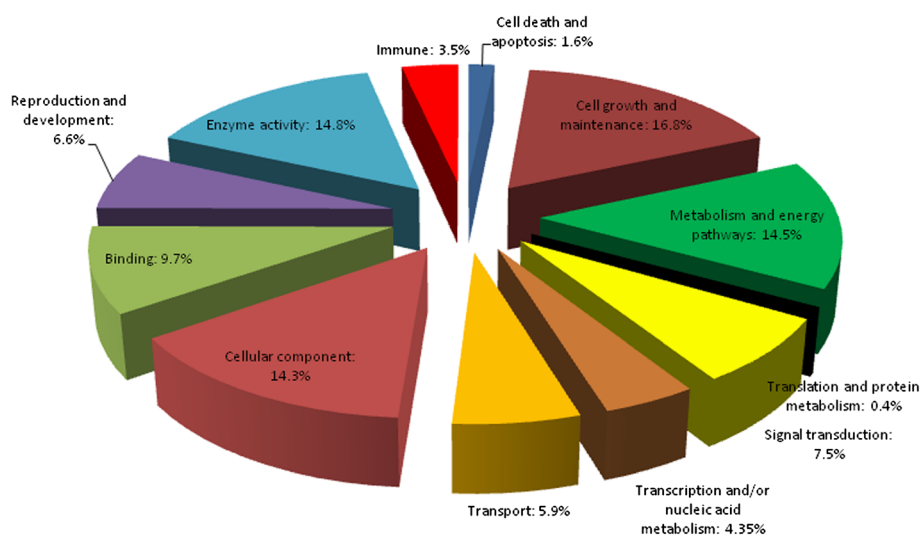
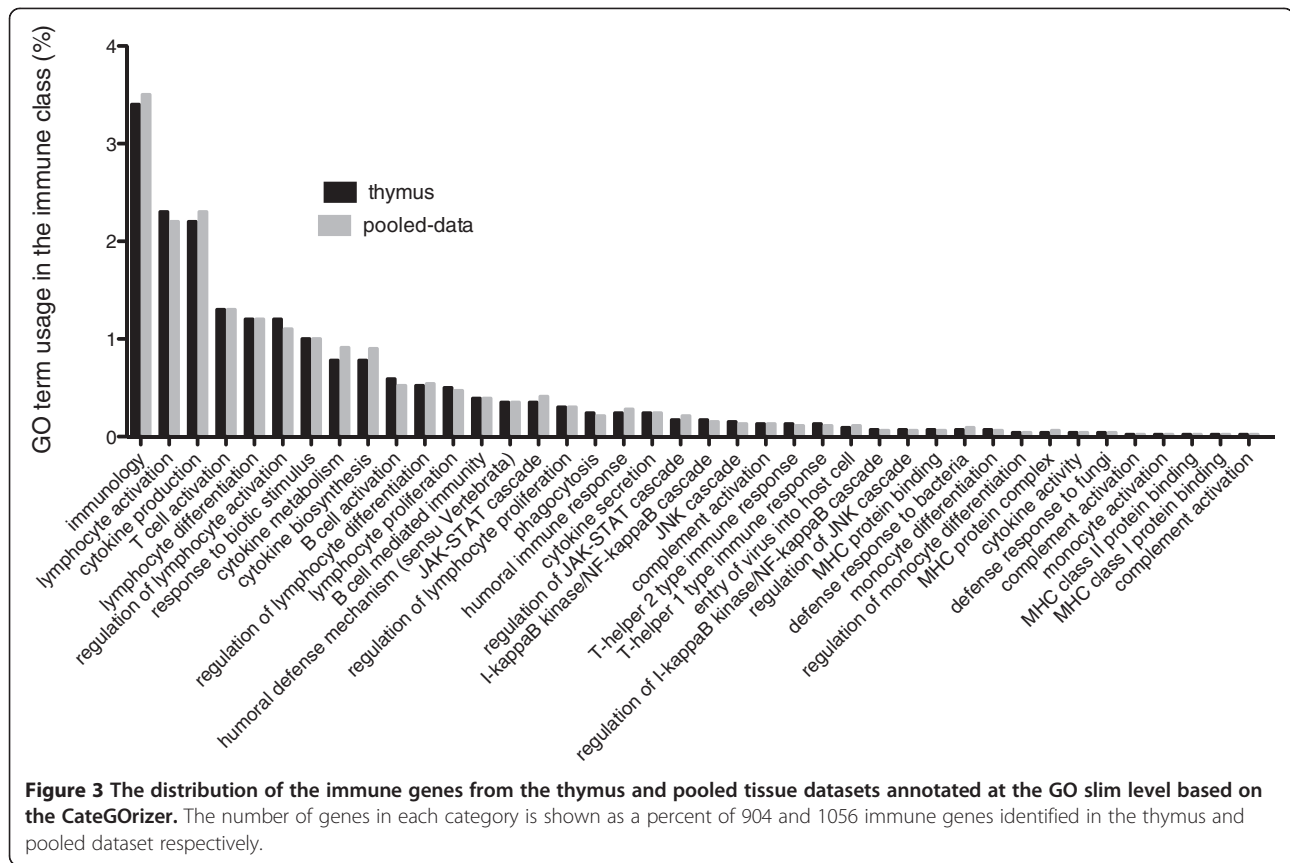


Figure 2 The distribution of 178,554 transcripts that have mapped to human genes from the bat thymus based on GO terms.



for identifying genes with low expression levels. RLHs are expressed in the cytoplasm where they recognise viral RNA and DNA [33,34]. Three bat RLH genes, retinoic-acid-inducible protein 1 (RIG-1), melanoma-differentiation-associated gene 5 (MDA5) and laboratory of genetics and physiology 2 (LGP2) were identified in our transcriptome datasets and have recently been described in *P. alecto* [17]. These results provide further evidence that bats are able to recognise a broad range of pathogens, similar to other species.

NLRs are a diverse family of cytoplasmic PRRs involved in the activation of a variety of signalling pathways. NLRs are primarily involved in bacterial recognition, although more recently, evidence for recognition of viral RNA and DNA by some members of the NLR family has been reported [35-37]. The only NLRs identified in the bat transcriptome datasets were NOD-like receptor family CARD domain containing 5 (NLRC5) and NLR family, pyrin domain containing 3 (NLRP3). NLRC5 is a recently identified NLR proposed to function as a positive and negative regulator of antiviral immune responses [36]. NLRP3 (also known as NALP3) is activated by a variety of danger signals including viral and bacterial infections and environmental irritants. Activation of NLRP3 in turn activates caspase-1 in the inflammasome which proteolytically cleaves the cytokines IL-1 β and IL-18 into active mature

peptides [37]. The identification two NLRs with associations with antiviral immunity in the bat transcriptome is remarkable and provides a starting point for understanding the role of NLRs in antiviral immunity in bats.

Interferons and interferon stimulated transcripts

The interferon (IFN) response is a key component of the innate immune system and the first cytokines induced against viral infection. Since the IFN response is important in the control of viral replication in other mammals, we searched the bat transcriptome for IFNs and IFN stimulated genes (ISGs) that may be critical to the ability of bats to remain asymptomatic to viral infections. Type I (including IFN α and β) and III (IFN λ) IFNs are induced directly in response to viral infection and play a role in the earliest stages of the innate immune response. Type I (α) IFN and its receptor (IFNAR1 and IFNAR2) were identified in the bat transcriptome datasets (Additional file 3). Although type III IFNs, IFN λ 1 and IFN λ 2 are up-regulated in stimulated bat cells [21], neither of these genes were identified in our datasets, likely reflecting a low expression level in our samples. The IL-10R2 chain of the type III IFN receptor was present in the bat transcriptome, but its partner chain IFNLR1 was not found. Both IL-10R2 and IFNLR1 were recently

described in *P. alecto* and IFNLR1 was demonstrated to act as a functional receptor for IFN λ [20].

The induction of type I and type III IFNs results in the transcription of hundreds of ISGs including PRRs that detect viral RNA, transcription factors that result in the amplification of the IFN response and a small number of proteins that are directly responsible for inducing an antiviral state. The ISGs, myxovirus resistance (Mx) GTPases, Protein kinase R (PKR), 2'-5' oligoadenylate synthetases (OAS), ribonuclease L (RNaseL) and ISG15 are among the proteins with confirmed antiviral activity in other mammals [38]. The bat transcriptome datasets contained genes orthologous to mammalian Mx1, Mx2, OAS1, OAS2, OAS3, OAS-like (OASL), PKR, RNaseL and ISG15 consistent with the presence of an ISG repertoire in bats that is similar to that of other species. These results provide the first evidence that the pathways activated by the IFN response are likely similar in bats to those described in other mammals.

The Mx gene family is among the best characterised ISGs, first identified as antiviral proteins following the observation that the sensitivity of many inbred mouse strains to orthomyxovirus was solely due to mutations within the Mx locus [39]. The Mx family of GTPases trap essential viral components, and in so doing prevent viral replication at early time points. Although the full spectrum of Mx antiviral activity is unknown, representatives of both RNA and DNA viruses have been shown to be sensitive to the effects of Mx [40]. A full length transcript, encoding a 667 amino acid protein was identified in our bat transcriptome datasets and found to be orthologous to Mx1 based on comparison with known mammalian Mx1 and Mx2 family members (Figure 4a and data not shown). Bat Mx1 contained the highly conserved tripartite GTP-binding domain found in all mammalian Mx proteins. In addition, a dynamin family signature and putative leucine zipper motif were found near the C terminal end, represented by a stretch of evenly spaced leucine residues. The bat protein was also conserved in the region identified as the stalk of human MxA including Loop 2 which is associated with antiviral activity. Consistent with other species, Loop 4 of the MxA stalk is the least conserved region of the bat Mx protein [41]. Loop 4 has been reported to be proteinase K sensitive and may play a role in lipid binding [42,43] (Figure 4b). Bat Mx1 does not contain the stretch of basic amino acids (K/R) near the C terminal end associated with nuclear localisation of mouse Mx1, consistent with the bat protein remaining localised within the cytoplasm [44]. The conservation of key residues important in antiviral activity is consistent with the bat Mx1 playing a role in antiviral immunity similar to other species. The identification of the sequences of important ISGs will now allow us to determine whether functional differences in the initiation and

regulation of these proteins account for the differences in susceptibility of bats to viral infections compared to other mammals.

Natural killer cell receptors

Natural killer (NK) cells are an important component of the innate immune response, providing a first line of defence against viruses and tumours. To our knowledge, no investigations of NK cell receptors from any species of bat have been reported previously. NK cells express cell surface receptors that recognise major histocompatibility complex (MHC) class I or class I like molecules on the surface of cells and lyse infected or abnormal cells by cytotoxicity. Two families of NK receptors that bind classical MHC class I ligands have been identified: the killer immunoglobulin like receptors (KIRs), which are encoded by genes in the leukocyte receptor complex (LRC), and the killer cell lectin like receptors (KLRs), which are encoded by genes in the natural killer complex (NKC). Different lineages of mammals make use of genes from the two different superfamilies to carry out analogous functions. KIRs are used preferentially by primates, cattle, domestic cats, dogs and pigs [45,46]. Similarly, the KIR-like receptors, marsupial immunoglobulin-like receptors (MAIRs) and chicken immunoglobulin-like receptors (CHIRs), have expanded in marsupials and chickens respectively [47,48]. Although CHIR-AB binds IgY, the ligand for the majority of CHIRs is unknown and the presence of a charged transmembrane residue and a cytoplasmic immunoreceptor tyrosine-based inhibition motif (ITIM), are consistent with the possibility that they play a role in NK activity [49]. Rodents, horses and platypus are the only species so far described that have expanded the KLRs, represented by the Ly49 family [50-52]. In the bat transcriptome dataset, no transcripts with homology to KIRs or Ly49 receptors were identified. In bony fish, novel immune type receptors (NITRs) which contain an N terminal variable domain and a C terminal Ig domain have been identified as the primary activating and inhibitory receptors expressed by NK cells [53]. NITRs were also used to search the bat transcriptome but failed to identify any orthologous transcripts. The failure to find KIR or Ly49 like receptors in the bat transcriptome may reflect low expression levels of these genes resulting in their absence from our datasets. However, BLAST searches of the publicly available whole genome sequence of the closely related pteropid bat, *P. vampyrus* revealed no evidence of KIRs or Ly49 receptors. As this is a low coverage genome (2.63X), further work is required to determine whether pteropid bats have KIR and/or Ly49 receptors. Overall, the absence of these important NK receptors from our datasets warrants further investigation into the nature of NK cells in bats.

(See figure on previous page.)

Figure 4 a. Phylogenetic analysis based on amino acid alignments of bat Mx with representative vertebrate species. Branch support is indicated as the percentage of 1000 bootstrap replicates and is shown where support is >60%. Species names are: *B. taurus* (Bota), *E. caballus* (Eqca), *G. gallus* (Gaga), *H. sapiens* (Hosa), *M. musculus* (Mumu), *O. aries* (Ovar), *O. cuniculus* (Orcu), *R. norvegicus* (Rano), *S. scrofa* (Susc). Phylogenetic analysis was performed based on alignment of amino acid position 49 to the end of the *P. alecto* Mx sequence with the corresponding region of Mx from other species. **b.** Alignment of deduced amino acid sequences of bat Mx with human and mouse Mx sequences. The conserved tripartite GTP binding domain, dynamin family signature, leucine residues and Loops 2 and 4 are shown in bold. The nuclear localisation signal present in mouse Mx1 is shaded. Dashes indicated similarity and dots indicate gaps.

NK cells in a wide range of mammalian species additionally express CD94/NKG2 (also called KLRD1/KLRC) lectin-like receptor heterodimers. Unlike the KIR and Ly49 receptors, which bind (classical) MHC class Ia ligands, the CD94/NKG2 heterodimer binds the (non-classical) MHC class Ib ligands HLA-E and Qa-1 in humans and mice respectively [54]. The CD94/NKG2A heterodimer generates inhibitory signals whereas the CD94/NKG2C heterodimer generates activating signals within NK cells. Both CD94 and NKG2A were identified in the bat transcriptome, however NKG2C transcripts were not identified, possibly reflecting the low abundance of transcripts of this gene in our datasets. Two and 37 NKG2A transcripts were identified in the thymus and pooled datasets respectively and six transcripts corresponding to CD94 were identified in the pooled dataset. Two of the longest NKG2A sequences were aligned with NKG2A and NKG2C sequences from human and mouse. As shown in Figure 5a, the bat genes display highest conservation with other NKG2A genes including the presence of conserved ITIM motifs in their cytoplasmic domains, designated by I/V/L/SxYxxL/V indicating that they are likely functional inhibitory receptors [55]. The more divergent NKG2D, which binds MHC class I chain-related genes, MICA/B, and the UL16 binding proteins (ULBPs) in human [46], was also detected.

Two distinct bat CD94 contigs were identified, one of which is missing two conserved cysteines in the stalk region, the first of which forms an interchain disulfide bond with NKG2 and the second which forms an intrachain disulphide bond. The second bat CD94 sequence is missing a conserved cysteine in the extracellular domain that forms an intrachain disulphide bond (Figure 5b). The absence of key cysteines in both of the bat CD94 sequences may have implications for the formation of heterodimers with NKG2 and for the unique folding of the CD94 chain. Combined with our failure to detect KIRs or Ly49 receptors, our data may provide the first evidence for the presence of atypical NK cell responses in bats. However, confirmation of the nature of the NK response and the composition of receptors used by bat NK cells awaits further investigation.

Other NK receptors were also identified in our datasets including CD244 which acts as an activating or inhibitory receptor on human and mouse NK cells

respectively [56] and the natural cytotoxicity receptors expressed by NK cells. Co-receptors including CD16 and CD56 expressed by subsets of NK cells in other species were also identified in the bat transcriptome. Identification of NK cell receptors and co-receptors provides information for the development of reagents to identify bat NK cells and paves the way for further studies of NK cell function during viral infection in bats.

Adaptive immunity

Genes involved in the adaptive immune system, including MHC class I and II genes and T and B cell receptors and co-receptors were highly represented in both the thymus and pooled datasets providing evidence that bats have all of the components necessary to mount an adaptive immune response.

Major histocompatibility complex class I genes

MHC class I molecules play an important role in the initiation of the adaptive immune response through recognition of endogenously-derived peptides from viruses and other pathogens. In the thymus dataset, 46 contigs had homology to mammalian MHC class I proteins, while 24 were homologous in the pooled data. Other transcripts in the MHC class I antigen-loading and presentation pathway were also identified, including beta-2-microglobulin, transporter associated with antigen processing 1 (TAP1), calnexin and tapasin. Class I-related genes were also present in the bat transcriptome dataset including CD1a, CD1b, CD1d, MR1, HFE, FcRn and ULBPs, which have a variety of immune and non-immune functions. The presence of ULBPs is consistent with the expression of NKG2D, but orthologues of MICA/B or Mill were not observed. The presence of NKG2D suggests bats should have a MIC homologue, but these may not be detected possibly due to low or tissue-specific expression. To our knowledge, these sequences provide the first class I and class I-associated transcripts from any species of bat.

Of the 46 contigs with homology to MHC class I genes in the thymus dataset, 29 contained in-frame stops. These may be expressed pseudogenes, represent assembly or sequencing errors or result from reading frame shifts due to the presence of unprocessed transcript. As the sequences were obtained from multiple individuals, it is not possible

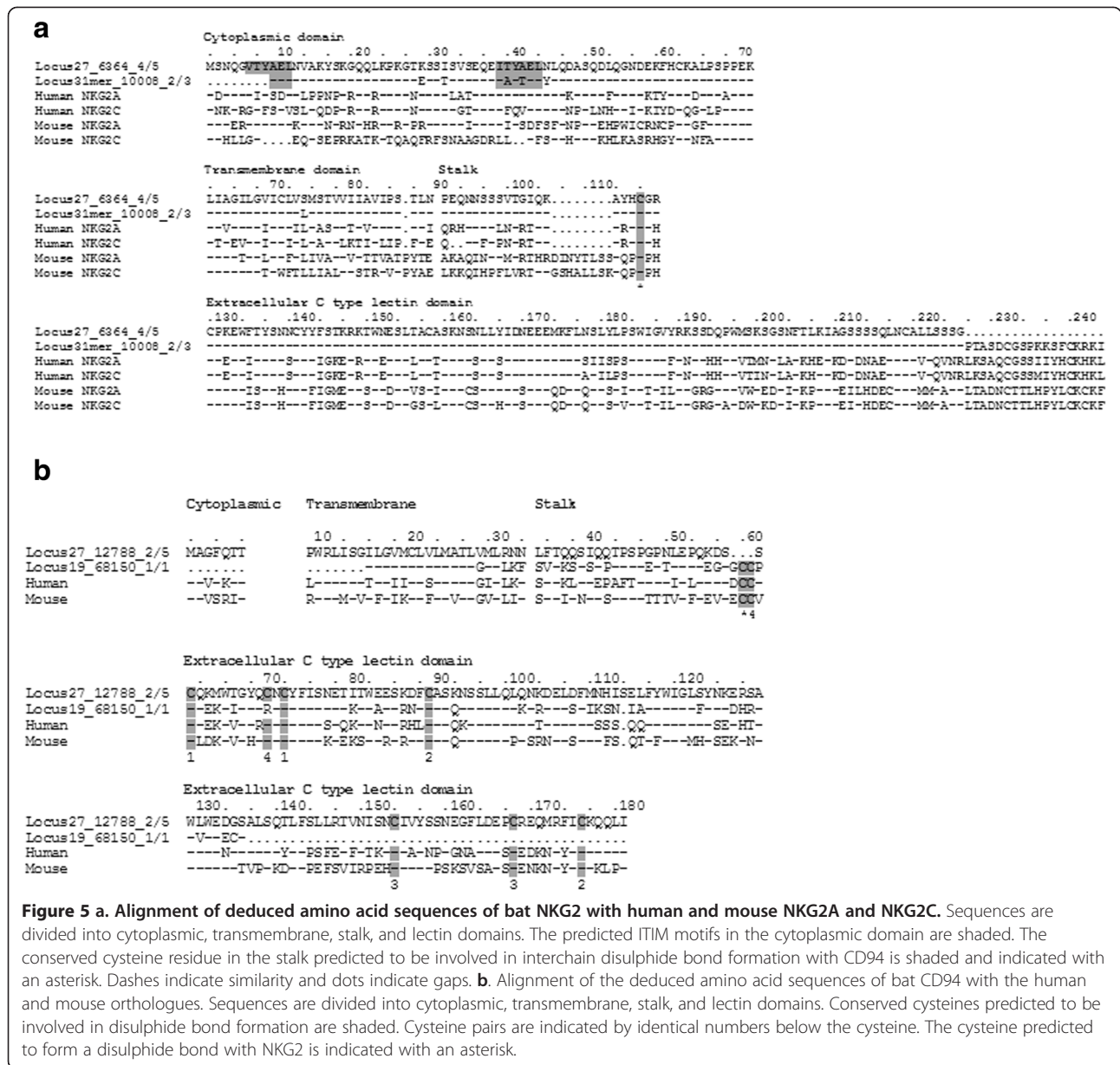


Figure 5 a. Alignment of deduced amino acid sequences of bat NKG2 with human and mouse NKG2A and NKG2C. Sequences are divided into cytoplasmic, transmembrane, stalk, and lectin domains. The predicted ITIM motifs in the cytoplasmic domain are shaded. The conserved cysteine residue in the stalk predicted to be involved in interchain disulphide bond formation with CD94 is shaded and indicated with an asterisk. Dashes indicate similarity and dots indicate gaps. **b.** Alignment of the deduced amino acid sequences of bat CD94 with the human and mouse orthologues. Sequences are divided into cytoplasmic, transmembrane, stalk, and lectin domains. Conserved cysteines predicted to be involved in disulphide bond formation are shaded. Cysteine pairs are indicated by identical numbers below the cysteine. The cysteine predicted to form a disulphide bond with NKG2 is indicated with an asterisk.

to confidently distinguish between alternative isoforms, alleles and in some cases, loci. However, clustering the remaining contigs with open reading frames (ORFs), there are clearly at least 9 distinct MHC class I genes expressed. The majority of class I contigs contained the $\alpha 1$ or $\alpha 2$ domain or partial sequence corresponding to both domains and were used for further sequence analysis. The deduced amino acid sequence of contigs with the most complete $\alpha 1$ or $\alpha 2$ domains were aligned with human HLA-A (Figure 6). All of the bat class I sequences contained a unique three amino acid insertion in the $\alpha 1$ domain that appears to be bat specific. As shown in Figure 6, the bat transcripts display amino acid variation in their $\alpha 1$ and $\alpha 2$ domains, corresponding

to the peptide binding region. However, they appear to be remarkably conserved from residues 131 to 175 of the $\alpha 2$ domain. These results may indicate that bats contain a very closely related class I gene repertoire that have coevolved with the specific viruses they carry.

Some of the class I transcripts represented in the thymus and pooled datasets contained an 84 bp insertion at the end of the $\alpha 1$ domain. The longest of these transcripts corresponded to the leader peptide through to 71 amino acids of the $\alpha 2$ domain and is shown in Figure 6. The insertion at the end of the $\alpha 1$ domain is not present in class I sequences from other mammals and includes two in frame stop codons that would prevent translation beyond the $\alpha 1$ domain (Figure 6). This sequence was

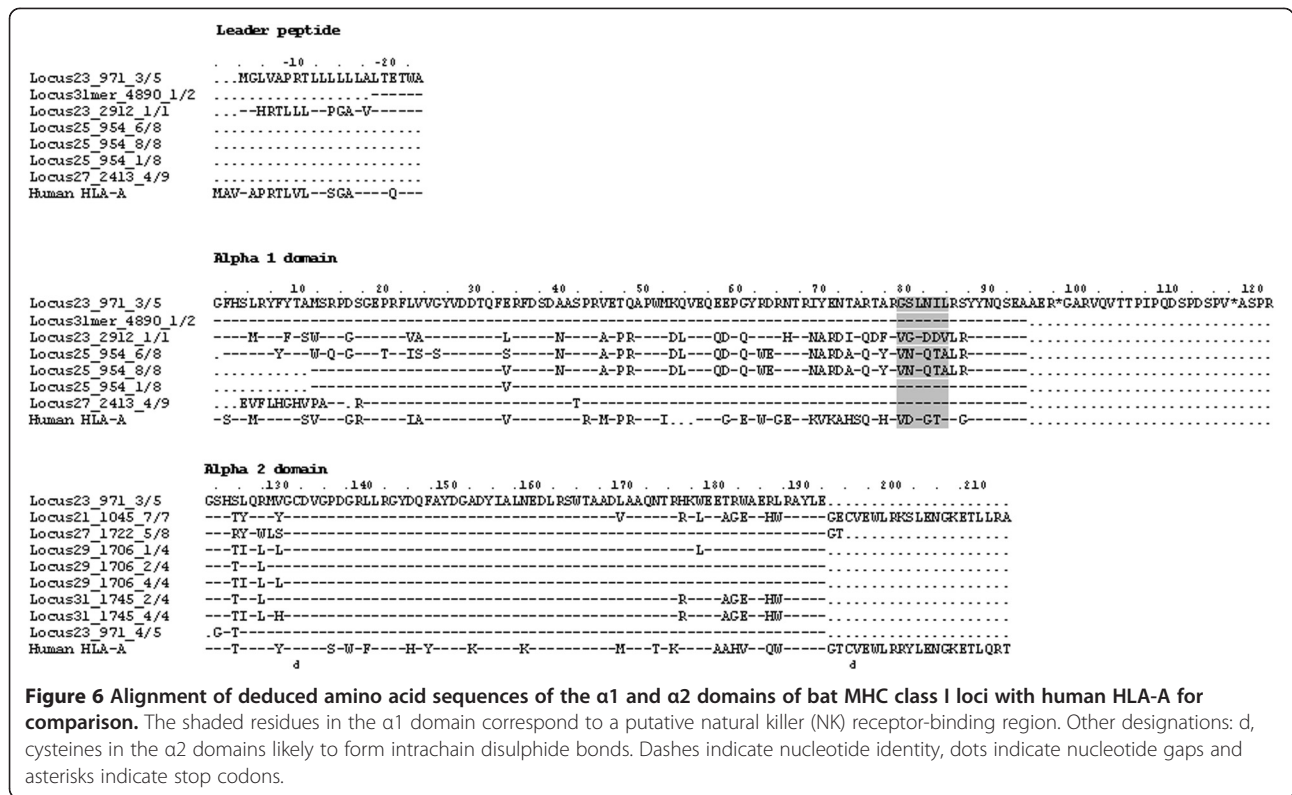


Figure 6 Alignment of deduced amino acid sequences of the $\alpha 1$ and $\alpha 2$ domains of bat MHC class I loci with human HLA-A for comparison. The shaded residues in the $\alpha 1$ domain correspond to a putative natural killer (NK) receptor-binding region. Other designations: d, cysteines in the $\alpha 2$ domains likely to form intrachain disulphide bonds. Dashes indicate nucleotide identity, dots indicate nucleotide gaps and asterisks indicate stop codons.

confirmed by RACE PCR and transcripts were detected in a variety of tissues including lymph node, spleen, liver, lung, heart, kidney, small intestine, brain and salivary glands, thus providing evidence that they are not an artefact of the transcriptome assembly (data not shown). Comparison with the closely related *P. vampyrus* whole genome sequence available in Ensembl revealed that the 84 bp insertion is identical to the beginning of intron 2 of a *P. vampyrus* class I gene. MHC class I splice variants that retain intron sequence and result in the translation of a truncated protein have been identified in other mammals, including soluble splice variants of human HLA-G that plays a role in immunoregulation at the fetal-maternal interface [57]. Further investigation will be required to determine whether the bat gene encodes a soluble protein corresponding only to the $\alpha 1$ domain or whether it represents a transcribed pseudo-gene. However, given the abundance of this transcript in our datasets it is possible that it plays a role in immune regulation in *P. alecto*.

Major histocompatibility complex class II genes

Unlike class I molecules, which are ubiquitously expressed, class II molecules are expressed only by antigen presenting and B cells and present exogenously derived peptides to T cells. The MHC class II molecules are composed of an α and a β chain encoded by A and B genes respectively [58,59]. Eutherians have three main

classical class II gene clusters: DP, DR, and DQ, as well as the nonclassical DM and DN/DO gene clusters [60,61]. Sequences corresponding to exon 2 of MHC class II DRB genes have been described in four species of microbats [62-64]. However, prior to the present study no class II genes have been reported from any species of megabat. Sequences corresponding to genes involved in the class II antigen processing and presentation pathway were also identified in our datasets including the class II invariant (CD74) chain and cathepsin S (Additional file 3).

In the *P. alecto* thymus and pooled datasets we identified 78 and 238 contigs respectively that were homologous to class II sequences. Phylogenetic analysis revealed that the alpha chain sequences were homologous to DMA, DOA, DQA and DRA from other mammals (Figure 7a) and the beta chain sequences were homologous to DMB, DOB, DQB and DRB (Figure 7b). These results are consistent with orthologous relationships between the bat class II genes and those from other mammals.

T and B cell receptors and co-receptors

T cell receptor (TCR) genes corresponding to all four chains of the T cell receptor were present in our datasets, consistent with bats having both $\alpha\beta$ and $\gamma\delta$ T cells. Sequences corresponding to the constant and variable domains of the TCR were identified including many TCR α related contigs, TCR β related contigs, a few TCR γ

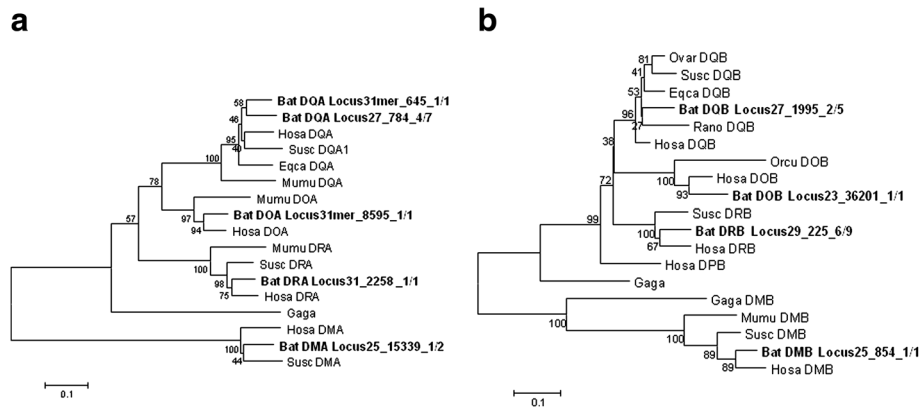


Figure 7 Phylogenetic Trees of the MHC Class II Genes. a. MHC class II A gene phylogeny based on alignment of 240 nucleotides corresponding to the end of the $\beta 2$ domain through to the cytoplasmic region of bat MHC class II sequences with class II sequences from other mammals. Branch support is indicated as the percentage out of 1000 bootstrap replicates and is shown where support is greater than 50%. **b.** MHC class II B gene phylogeny based on alignment of 567–594 nucleotides corresponding to the signal peptide through to the $\alpha 2$ domain of bat MHC class II sequences with class II sequences from other mammals. Branch support is indicated as the percentage out of 1000 bootstrap replicates and is shown where support is greater than 50%. Species names are: *E. caballus* (Eqca), *G. gallus* (Gaga), *H. sapiens* (Hosa), *M. musculus* (Mumu), *O. aries* (Ovar), *O. cuniculus* (Orcu), *R. norvegicus* (Rano), *S. scrofa* (Susc).

and TCR δ chain related contigs. In humans and mice approximately 95% of circulating T cells express the $\alpha\beta$ T cell receptor. In contrast, $\gamma\delta$ T cells account for up to 70% of circulating T cells in young ruminants, rabbits and chickens [65,66]. The low abundance of TCR γ and TCR δ related transcripts in our datasets is consistent with the possibility that $\alpha\beta$ T cells may be the predominant TCR present in bats. In addition, a variety of T cell co-receptors, including the accessory TCR ζ chain, CD3, CD4, CD8 and CD28 were identified in our datasets.

We previously described the immunoglobulin heavy chain diversity of *P. alecto*, revealing the presence of a highly diverse variable region gene repertoire [16]. Sequences encoding the variable and constant domains of immunoglobulin heavy and light chains were represented in our datasets. These included heavy chain genes encoding IgA, IgG, IgM and IgE, which have previously been described in the megabat, *Cynopterus sphinx*. No evidence for the transcription of IgD was observed in the *P. alecto* transcriptome, a result which is consistent with *C. sphinx* [67]. The two light chain subtypes, kappa and lambda and a variety of B cell co-receptors including CD19, CD22, CD72, CD79a and CD79b were also identified in our datasets (Additional file 3).

Conservation of bat immune genes

Many of the bat immune transcripts showed high levels of sequence similarity compared to homologues from other mammals. Among the most conserved bat innate immune genes were the PRRs; the TLRs, RIG-I helicases and NLRs, which displayed >80% amino acid identity with homologues. This likely reflects their roles in the recognition of conserved pathogen motifs. Members of

the OAS family were also highly conserved, in particular OAS1 which shared 87% amino acid identity with the dog OAS1 sequence. In addition, the NK co-receptor, CD56 shared 93% amino acid identity with mouse, hamster, guinea pig and human sequences. Among the adaptive immune genes, MHC associated proteins, calnexin, TAP1 and cathepsin S shared 89-95% identity with corresponding sequences from other mammals reflecting their conserved roles in the antigen processing and presentation pathway. Several members of the MHC class I and II families were also highly conserved, including CD1b and CD1d which shared 88 and 89% amino acid identity with horse and chimp sequences respectively. The bat MHC class II DOA and DRA shared 91 and 89% amino acid identity with orthologous sequences in other mammals. The T cell co-receptor, CD28 shared 90% identity with the rhinoceros CD28 sequence and the constant domain of IgM shared 92% identity with camel IgM.

Unannotated transcripts

There were >77,000 unannotated contigs in the thymus and pooled datasets. Only about 3% of these contigs matched predicted cDNAs from the *P. vampyrus* genome sequence, which are annotated using orthologous sequences from other species [68]. The unannotated contigs contained a total of 3266 open reading frames (ORFs) longer than 300 bp. Of these, 92.6% (E-value < 10^{-3}) aligned to the closely related *P. vampyrus* whole genome sequence and represent highly divergent homologues or bat specific genes. The remaining loci represent either misassembled contigs or bat-specific transcripts that are located in sequencing gaps in the low coverage *P. vampyrus* genome

sequence. The 3266 long (>300nt) ORFs were searched for conserved domains using profile hidden Markov models with hmmscan (HMMer v3; <http://hmmer.org/>) obtained from the Pfam database [69]. This identified 345 ORFs containing 214 unique domains, including several defensins, antimicrobial peptides and DNA-binding domains. Searches using domain models from the Pfam-B database, identified a further 437 unique, predicted-conserved domains in 733 ORFs. A further 2188 ORFs remained unannotated. A high proportion of these were rich in cysteine, tryptophan and proline, and prolines frequently appeared in low complexity regions (additional file 4: Figure S2a and b). Further characterisation of these unannotated transcripts will provide insight into whether they are functionally significant and in particular whether any unique bat specific transcripts are involved in the antiviral immune response.

Conclusion

Bats are a highly diverse, species rich group of mammals that have evolved a variety of distinctive characteristics since their divergence from other mammals [70]. Despite the central importance of bats in harbouring a variety of viruses with the potential to spillover to other species, very little is known about antiviral immunity in bats. Next generation sequencing provides the opportunity to survey genes that are conserved between distantly related species as well as to provide insights into novel adaptations through the identification of previously unidentified transcripts. To identify genes involved in the immune response, we carried out a transcriptome analysis of thymus and immune cells and tissues of the Australian black flying fox, *P. alecto*. This study represents the first survey of expressed bat immune genes and complements existing low coverage bat genome sequences. Our analysis provides a broad overview of the bat transcriptome and contains representatives of all of the major classes of immune genes. The results are consistent with bats having all of the components of the immune system present in other mammals. The majority of these correspond to genes that have not previously been described in any species of bat and thus represent an important resource for future investigations into antiviral immunity in bats.

Methods

Animals

P. alecto bats used in this study were wild caught from East Brisbane, Queensland, Australia. Bats were handled and euthanised as previously described [15]. All experiments were approved by the Australian Animal Health Laboratories animal ethics committee (protocol AEC1281).

Preparation of tissues and cells

The thymus was removed from a juvenile male bat and immediately stored in *RNAlater* (Ambion) for RNA

extraction. The spleen, lymph nodes (LN), thymus, bone marrow and peripheral blood were collected from one adult male and one pregnant female bat. Single cells were extracted from the spleen, thymus and LN by tissue extrusion through a 70 μ M sterile sieve (BD Biosciences) in the presence of DMEM supplemented with 15 mM L-glutamine, 100 units/ml penicillin and 100 units/ml streptomycin (Invitrogen). Splenocytes and peripheral blood lymphocytes (PBMLs) were isolated by density centrifugation over lymphoprep (Nycomed, Oslo) as described previously [21]. Cells were resuspended in DMEM with 10% FCS, 15 mM L-glutamine, 100 units/ml penicillin and 100 units/ml streptomycin and cell numbers were determined using a haemocytometer with trypan blue exclusion. The thymus and bone marrow cells were stored in *RNAlater* (Ambion) for RNA extraction and the spleen, thymus and LN were cultured with a variety of stimulants.

Culture of cells

The isolated splenocytes, LN and PBMLs from each bat were pooled and were then seeded at 1×10^7 cells per well in 24 well tissue culture plates (Nunc) with PHA (10ug/ml; Sigma) and LPS (10 μ g/ml; Sigma); PMA (50 μ g/ml; Sigma) and Ionomycin (2nM/ml; Sigma); or polyIC (30 μ g/ml; InvivoGen) and incubated in a humidified atmosphere of 5% CO₂ in air at 37°C. Cells were harvested in RLT buffer (Qiagen) at 0, 1, 4 and 18 hours and homogenised using a QIAshredder (Qiagen) following the manufacturer's instructions. The lysate was then stored at -80°C and total RNA extracted the next day (0, 1, and 4 hours) or processed immediately (18 hours).

Preparation of total RNA

RNA extraction was carried out as previously described using the RNeasy mini kit (Qiagen) with removal of genomic DNA with DNase I digestion [16]. Total RNA from the thymus of a juvenile male bat was used for Illumina sequencing separately from all other samples. Total RNA obtained from the stimulated and unstimulated cells from the two adult bats was pooled as follows: 22% thymus total RNA (11% from each bat) and 78% pooled total RNA from the rest of the mitogen stimulated and unstimulated cells/tissues (~3.45% for each sample; total of 22 samples).

Sequencing

mRNA isolation from total RNA, library preparation and single-end read sequencing was performed by Geneworks Pty Ltd, Thebarton South Australia using the Illumina Genome Analyser Iix sequencing platform. Library preparation was performed as per Illumina's mRNA sequencing sample preparation guide (Part # 1004898 Rev. D) except 5 μ g of total RNA was used for mRNA selection

using poly-T oligo-attached magnetic beads. The thymus library was run on a single lane of a flow cell resulting in more than 12.5 million 65-base sequences for a total of about 0.82 Gigabases (Gb) of sequence. The pooled library consisted of 4 lanes resulting in 24 million 76 bp sequences for a total of about 1.8Gb of sequence.

Sequence pre-processing and de novo assembly

The quality of the sequences were evaluated using FastQC [71]. Sequences were pre-processed in two stages. First, all bases at the 3' end of the reads with quality scores of 3 or lower were removed. Second, poly A/T tails, uninformative sequences (Ns) and primer/adaptor contaminants were trimmed using SnoWhite (version 1.1.3) [72], a cleaning pipeline for next-generation cDNA sequences, which includes Seqclean [73] and TagDust [74]. We ran SnoWhite with two runs of Seqclean and one run of TagDust and a final minimal length cutoff of 20 bp was used.

The pre-processed sequences were *de novo* assembled using two different approaches. (1) The reads were assembled with velvet (version 1.0.12) [75] using individual kmers from 19 bp to 31 bp. Next, the contigs produced by velvet were processed using oases (version 0.1.15) [76]. Oases loci were then merged using cd-hit-est (version 4.0) [77] with a global sequence identity threshold of 1.0. Finally, a length cutoff was set to 50 bp and the default coverage cutoff of 3 was used. We term the final result of this process a contig (2). The reads were also assembled using MIRA 3 (V3.2.0rc3) [78] with default setting for EST and Illumina reads assembly, i.e. maximum front and end gap clip is 2 bp, maximum length of the possible vector leftover allowed is 18 bp, minimum quality score, window length and read length were all set to 20, allowed to clip poly A/T at ends, and minimum read coverage per contig was 2.

Annotation of the conserved protein coding transcripts and untranslated regions (UTRs)

The bat contigs were firstly annotated by using the best hits of BLASTX [79] search against NR protein database and KEGG pathway database with an E-value cutoff of 0.001 for annotating the protein coding contigs that were conserved with other species. Then the unannotated contigs were further annotated by using BLASTN search against Refseq_rna database with an E-value cutoff of 10^{-5} for the contigs containing conserved UTRs and without significant protein coding regions. The contigs not annotated by the above two steps were further analysed by using BLASTN against the cDNAs from megabat (*P. vampyrus*) and microbat (*M. lucifugus*).

Identification of open reading frames

We translated the un-annotated transcripts into protein sequences from 6 frames, extracted the ORFs longer

than 300 bp. This was performed separately for the 2 datasets. These ORFs were searched against Pfam-A and Pfam-B databases to identify conserved domains. The two sets of long ORFs were pooled and clustered based on cd-hit with sequence identity of 50% [77]. The amino acid compositions were further analysed for the non-redundant longer ORFs with Composition Profiler [80].

Annotation of the gene ontology (GO) terms

All the KEGG IDs of the human proteins identified by the BLASTX searches were extracted from the annotation process and were mapped to UniProt IDs. Then the GO analysis for the UniProt proteins (UniProtKB-GOA: gene_association.goa_human) was used to assign the GO terms for the transcripts. The number of genes in categories of the GO slim database was counted using the GO term classification counter, CateGORiser [81,82] and the immune category of the bat genes was annotated using the Generic Gene Ontology Term Mapper [83]. The GO classifications were further grouped into twelve broad categories as follows: Cell death and apoptosis (death, GO:0016265; cell death GO:0008219; apoptosis GO:0006915). Cell growth and maintenance (cell, GO:0005623; cell cycle, GO:0007049; cell differentiation, GO:0030154; cell proliferation, GO:0008283; growth, GO:0040007; cell homeostasis, GO:0019725; organelle organisation and biogenesis, GO:0006996; cytoskeleton organisation and biogenesis, GO:0030036; biosynthesis, GO:0009058; morphogenesis, GO:0009653; response to stress, GO:0006950; response to external stimulus, GO:0009605; protein modification, GO:0006464; response to abiotic stimulus, GO:0009628; response to biotic stimulus, GO:0009607; structural molecule activity, GO:0005198 and antioxidant activity, GO:0016209). Metabolism and energy pathways (metabolism, GO:0008152; generation of precursor metabolites and energy, GO:0006091; protein metabolism, GO:0019538; lipid metabolism, GO:0006629; oxygen binding, GO:0019825 and catabolism, GO:0009056). Translation and protein metabolism (protein biosynthesis, GO:0006412, translation regulator activity, GO:0045182 and translation factor activity, GO:0008135). Signal transduction (cell communication, GO:0007154; signal transduction, GO:0007165; signal transducer activity, GO:0004871 and cell to cell signaling, GO:0007267). Transcription and/or nucleic acid metabolism (nucleobase, nucleoside, nucleotide and nucleic acid metabolism, GO:0006139; transcription, GO:0006350; transcription regulator activity, GO:0030528; DNA metabolism, GO:0006259 and transcription factor activity, GO:0003700) Transport (transport, GO:0006810; ion transport, GO:0006811; protein transport, GO:0015031 and transporter activity, GO:0005215). Cellular component (intracellular, GO:0005622; cytoplasm, GO:0005737, plasma membrane, GO:0005886; cytoskeleton, GO:0005856; extracellular region, GO:0005576, cytoplasmic membrane bound vesicle, GO:0016023; endoplasmic

reticulum, GO:0005783; extracellular matrix, GO:0005578; ribosome, GO:0005840; nucleolus, GO:0005730; extracellular space, GO:0005615; chromosome, GO:0005694; endosome, GO:0005768; cytosol, GO:0005829; nucleus, GO:0005634; mitochondrion, GO:0005739; cell, GO:0005623 and nucleoplasm, GO:0005654). Binding (binding, GO:0005488; protein binding, GO:0005515; nucleic acid binding, GO:0003676; DNA binding, GO:0003677; nucleotide binding, GO:0000166; cytoskeletal protein binding, GO:0008092; receptor activity, GO:0004872; actin binding, GO:000377; calcium ion binding, GO:0005509; chromatin binding, GO:0003682; carbohydrate binding, GO:0030246 and RNA binding, GO:0003723). Reproduction and Development (development, GO:0007275; reproduction, GO:0000003 and embryonic development, GO:0009790). Enzyme activity (catalytic activity, GO:0003824; hydrolase activity, GO:0016787; transferase activity, GO:0016740; enzyme regulator activity, GO:0030234; kinase activity, GO:0016301; protein kinase activity, GO:0004672 and phosphoprotein phosphatase activity, GO:0004721). Immune Response (immunology, immune response, GO:0006955).

Sequence and phylogenetic analysis

Mx protein sequences and MHC and NK receptor nucleotide sequences were aligned using ClustalX [84]. Nucleotide sequences were aligned and gapped manually using Bioedit software version 7.0.9 (Tom Hall, Ibis Biosciences, Carlsbad, CA), based on the protein alignment to retain codon positions. Based on the nucleotide and protein alignments, phylogenetic trees were constructed by the neighbour joining method [85], maximum parsimony and minimum evolution using the MEGA4 program [86].

The GenBank accession numbers for sequences used in the sequence and phylogenetic analysis are as follows: MHC class I: (CAP58485) HLA-A; MHC class II: Human, *Homo sapiens* (Hosa) DMA (NM_006120), DOA (M26039), DQA (M26041), DRA (NM_019111); Mouse, *Mus musculus* (Mumu) DOA (M95514), DQA (M21931), DRA (U13648); Horse DQA (L33909); Pig, *Sus scrofa* (Susc) DQA (NM_001130224), DRA (NM_001113706), DMA (NM_001004039); Chicken, *Gallus gallus* (Gaga) B-LA (AY357253); MHC class IIB: Hosa DQB (M20432), DRB (NM_021983), DOB (L29472), DPB (M57466), DMB (U15085); Rat, *Rattus norvegicus* (Rano) DQB (X56596); Mumu DMB (U35332); Horse, *Equus caballus* (Eqca) DQB (L33910); Susc DQB (AY102478), DMB (DQ431246), DRB (AY191776); Sheep, *Ovis aries* (Ovar) DQB (L08792); Rabbit, *Oryctolagus cuniculus* (Orcu) DOB (XP_002714654); Gaga B-LBII (M29763), DMB (AB426148); NKG2A: Hosa (NM_002259), Mumu (NM_001136068); NKG2C: Hosa (NM_002260), Mumu (AF106010); CD94: Hosa (AJ000001), Mumu (AF030312). Mx1: Rano (NP_775119.2); Eqca (NP_001075961.1); Orca (XP_002723805.1); Susc (BAD11809.1);

Bota (NP_776365.1); Ovar (P33237.1); Mumu (NP_034976.1); Hosa (AF135187_1); Mx2: Rano (NP_599177); Eqca (XP_001491517); Susc (BAF76735); Bota (DAA33026); Ovar (AAW51454); Mumu (NP_038634); Hosa (NP_002454); Avian Mx: Gaga (NP_989940).

Analysis of repetitive or low complexity regions

Low complexity regions in protein sequences were detected with the seg program with default parameters [87].

Quantitative reverse transcription PCR

The transcription of a bat MHC class I gene was examined using quantitative PCR (qPCR) as described previously [18]. Briefly, total RNA was prepared from lymph node, spleen, liver, lung, heart, kidney, small intestine, brain and salivary glands using the RNeasy mini kit (Qiagen) as described above. cDNA was generated using a Quantitect reverse transcription kit for RT-PCR (Qiagen). qPCR primers were designed using Primer Express 3.0 (Applied Biosystems) with default parameter settings (5'-ACGACTCCTATTCCTCCAGGATAG-F and 5'-GAAAGC CACTGGTACCTGTGAGA-R). Reactions were carried out using EXPRESS SYBR[®] GreenER[™] qPCR Supermix Universal (Invitrogen) and an Applied Biosystems 7500 Fast Real-Time qPCR instrument.

Additional files

Additional file 1: Table S1. Summary of additive multiple-kmer velvet/oases/Mira3 assembly.

Additional file 2: Figure S1. Overview of the bat transcriptome. The distribution of 178,554 and 285,268 transcriptome sequences that have mapped to human orthologues from *P. alecto* thymus and pooled tissue datasets based on GO slim terms. Sequences within the three areas of Gene Ontology: molecular function, biological process and cellular component are further divided into subgroups at the GO Slim level.

Additional file 3: Sequences of all genes described in the manuscript.

Additional file 4: Figure 2. Amino acid composition of large unannotated ORFs. The horizontal axis shows amino acids sorted by flexibility index [88]. **a.** Amino acid composition of 1656 large unannotated non-redundant ORFs relative to proteins in the SwissProt database [89]. The amino acids Trp, Cys and Pro have twice the abundance in unannotated ORFs compared to SwissProt proteins. **b.** Amino acid composition of 1195 low complexity regions in unannotated ORFs relative to 1656 unannotated non-redundant ORFs. Prolines are abundant in low complexity regions, but Trp and Cys are not.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Conceived and designed the experiments: L-FW, HF, MT, MLB, CC. Performed the experiments: MT, VJ, JN, GC. De novo assembly and annotation: Z-PF. Analysed the data: Z-PF, ATP, MT, MLB. Contributed reagents/materials/analysis tools: HF. Wrote the paper: MLB, ATP, Z-PF, MT. All authors read and approved the final manuscript.

Acknowledgements

We thank Craig Smith, Carol de Jong, Deborah Middleton and Susanne Wilson for provision of bat tissues and peripheral blood used for this study and Andrew Bean for advice on mitogen stimulation of bat cells. The work was supported by funding from the Australian Biosecurity Cooperative Research Centre for Emerging Infectious Diseases to L-F.W and H.F. (Project 2.051R), a CSIRO CEO Science Leaders award to L-F.W and an NHMRC Career Development Fellowship to ATP.

Author details

¹Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, Melbourne, VIC 3052, Australia. ²Department of Mathematics and Statistics, University of Melbourne, Melbourne, VIC 3010, Australia. ³CSIRO Livestock Industries, Australian Animal Health Laboratory, PO Bag 24, Geelong, VIC 3220, Australia. ⁴Center for Evolutionary and Theoretical Immunology, Department of Biology, The University of New Mexico, Albuquerque, NM 87131, USA. ⁵Department of Medical Biology, University of Melbourne, Melbourne, VIC 3010, Australia. ⁶Australian Biosecurity Cooperative Research Centre for Emerging Infectious Diseases, Brisbane, Australia. ⁷Queensland Primary Industries and Fisheries, Biosecurity Queensland, Brisbane, Australia.

Received: 8 February 2012 Accepted: 16 May 2012

Published: 20 June 2012

References

1. Simmons NB: **Order Chiroptera**. In *Mammal species of the world: a taxonomic and geographic reference*. 3rd edition. Edited by Wilson DE, Reeder DM. Baltimore: Johns Hopkins University Press; 2005:312–529.
2. Teeling E: **Bats (Chiroptera)**. In *The timetree of life*. Edited by Hedges SB, Kumar S. New York: Oxford University Press; 2009.
3. Leroy EM, Epelboin A, Mondonge V, Pourrut X, Gonzalez JP, Muyembe-Tamfum JJ, Formenty P: **Human Ebola outbreak resulting from direct exposure to fruit bats in Luebo, Democratic Republic of Congo, 2007**. *Vector Borne Zoonotic Dis* 2009, **9**(6):723–728.
4. Leroy EM, Kumulungui B, Pourrut X, Rouquet P, Hassanin A, Yaba P, Delicat A, Paweska JT, Gonzalez JP, Swanepoel R: **Fruit bats as reservoirs of Ebola virus**. *Nature* 2005, **438**(7068):575–576.
5. Sulkin SE, Allen R, Sims R, Singh KV: **Studies of arthropod-borne virus infections in Chiroptera. IV. The immune response of the big brown bat (*Eptesicus f. fuscus*) maintained at various environmental temperatures to experimental Japanese B encephalitis virus infection**. *Am J Trop Med Hyg* 1966, **15**(3):418–427.
6. Swanepoel R, Leman PA, Burt FJ, Zachariades NA, Braack LE, Ksiazek TG, Rollin PE, Zaki SR, Peters CJ: **Experimental inoculation of plants and animals with Ebola virus**. *Emerg Infect Dis* 1996, **2**(4):321–325.
7. Williamson MM, Hooper PT, Selleck PW, Gleeson LJ, Daniels PW, Westbury HA, Murray PK: **Transmission studies of Hendra virus (equine morbillivirus) in fruit bats, horses and cats**. *Aust Vet J* 1998, **76**(12):813–818.
8. Williamson MM, Hooper PT, Selleck PW, Westbury HA, Slocombe RF: **Experimental Hendra Virus Infection in Pregnant Guinea-pigs and Fruit Bats (*Pteropus poliocephalus*)**. *J Comp Pathol* 2000, **122**(2–3):201–207.
9. Calisher CH, Childs JE, Field HE, Holmes KV, Schountz T: **Bats: important reservoir hosts of emerging viruses**. *Clin Microbiol Rev* 2006, **19**(3):531–545.
10. Wong S, Lau S, Woo P, Yuen K-Y: **Bats as a continuing source of emerging infections in humans**. *Rev Med Virol* 2007, **17**:67–91.
11. Middleton DJ, Morrissy CJ, van der Heide BM, Russell GM, Braun MA, Westbury HA, Halpin K, Daniels PW: **Experimental Nipah Virus Infection in Pteropid Bats (*Pteropus poliocephalus*)**. *J Comp Pathol* 2007, **136**(4):266–272.
12. Field H, McCall B, Barrett J: **Australian bat lyssavirus infection in a captive juvenile black flying fox**. *Emerg Infect Dis* 1999, **5**(3):438–440.
13. McColl KA, Chamberlain T, Lunt RA, Newberry KM, Middleton D, Westbury HA: **Pathogenesis studies with Australian bat lyssavirus in grey-headed flying foxes (*Pteropus poliocephalus*)**. *Aust Vet J* 2002, **80**(10):636–641.
14. Jones KE, Purvis A, MacLarnon A, Bininda-Emonds OR, Simmons NB: **A phylogenetic supertree of the bats (Mammalia: Chiroptera)**. *Biol Rev Camb Philos Soc* 2002, **77**(2):223–259.
15. Cramer G, Todd S, Grimley S, McEachern JA, Marsh GA, Smith C, Tachedjian M, De Jong C, Virtue ER, Yu M, *et al*: **Establishment, immortalisation and characterisation of pteropid bat cell lines**. *PLoS One* 2009, **4**(12):e8266.
16. Baker ML, Tachedjian M, Wang L-F: **Immunoglobulin heavy chain diversity in Pteropid bats: evidence for a diverse and highly specific antigen binding repertoire**. *Immunogenetics* 2010, **62**:173–184.
17. Cowled C, Baker ML, Zhou P, Tachedjian M, Wang L-F: **Molecular characterisation of RIGI-like helicases in the Black flying fox, *Pteropus alecto***. *Dev Comp Immunol* 2012, **36**(4):657–664.
18. Cowled C, Baker M, Tachedjian M, Zhou P, Bulach D, Wang LF: **Molecular characterisation of Toll-like receptors in the black flying fox *Pteropus alecto***. *Dev Comp Immunol* 2011, **35**(1):7–18.
19. Virtue ER, Marsh GA, Baker ML, Wang L-F: **Interferon Production and Signaling Pathways Are Antagonized during Henipavirus Infection of Fruit Bat Cell Lines**. *PLoS One* 2011, **6**(7):e22488.
20. Zhou P, Cowled C, Marsh GA, Shi Z, Wang L-F, Baker ML: **Type III IFN receptor expression and functional characterisation in the pteropid bat, *Pteropus alecto***. *PLoS One* 2011b, **6**(9):e25385.
21. Zhou P, Cowled C, Todd S, Cramer G, Virtue ER, Marsh GA, Klein R, Shi Z, Wang LF, Baker ML: **Type III IFNs in pteropid bats: differential expression patterns provide evidence for distinct roles in antiviral immunity**. *J Immunol* 2011, **186**(5):3138–3147.
22. Pumo DE, Finamore PS, Franek WR, Phillips CJ, Tarzami S, Balzarano D: **Complete Mitochondrial Genome of a Neotropical Fruit Bat, *Artibeus jamaicensis*; and a New Hypothesis of the Relationships of Bats to Other Eutherian Mammals**. *J Mol Evol*
23. Madsen O, Scally M, Douady CJ, Kao DJ, DeBry RW, Adkins R, Amrine HM, Stanhope MJ, de Jong WW, Springer MS: **Parallel adaptive radiations in two major clades of placental mammals**. *Nature* 2001, **409**(6820):610–614.
24. Murphy WJ, Eizirik E, Johnson WE, Zhang YP, Ryder OA, O'Brien SJ: **Molecular phylogenetics and the origins of placental mammals**. *Nature* 2001, **409**(6820):614–618.
25. Murphy WJ, Eizirik E, O'Brien SJ, Madsen O, Scally M, Douady CJ, Teeling E, Ryder OA, Stanhope MJ, de Jong WW, *et al*: **Resolution of the early placental mammal radiation using Bayesian phylogenetics**. *Science* 2001, **294**(5550):2348–2351.
26. Springer MS, Stanhope MJ, Madsen O, de Jong WW: **Molecules consolidate the placental mammal tree**. *Trends Ecol Evol* 2004, **19**(8):430–438.
27. Nishihara H, Hasegawa M, Okada N: **Pegasoferae, an unexpected mammalian clade revealed by tracking ancient retroposon insertions**. *Proc Natl Acad Sci U S A* 2006, **103**(26):9929–9934.
28. Mouchaty SK, Gullberg A, Janke A, Arnason U: **The phylogenetic position of the Talpidae within eutheria based on analysis of complete mitochondrial sequences**. *Mol Biol Evol* 2000, **17**(1):60–67.
29. Nikaido M, Harada M, Cao Y, Hasegawa M, Okada N: **Monophyletic origin of the order chiroptera and its phylogenetic position among mammalia, as inferred from the complete sequence of the mitochondrial DNA of a Japanese megabat, the Ryukyu flying fox (*Pteropus dasymallus*)**. *J Mol Evol* 2000, **51**(4):318–328.
30. Nikaido M, Kawai K, Cao Y, Harada M, Tomita S, Okada N, Hasegawa M: **Maximum likelihood analysis of the complete mitochondrial genomes of eutherians and a reevaluation of the phylogeny of bats and insectivores**. *J Mol Evol* 2001, **53**(4–5):508–516.
31. Prasad AB, Allard MW, Green ED: **Confirming the Phylogeny of Mammals by Use of Large Comparative Sequence Data Sets**. *Mol Biol Evol* 2008, **25**(9):1795–1808.
32. Van Den Bussche RA, Hooper SR, Hansen EW: **Characterization and phylogenetic utility of the mammalian protamine p1 gene**. *Mol Phylogenet Evol* 2002, **22**(3):333–341.
33. Kato H, Takeuchi O, Sato S, Yoneyama M, Yamamoto M, Matsui K, Uematsu S, Jung A, Kawai T, Ishii KJ, *et al*: **Differential roles of MDA5 and RIG-I helicases in the recognition of RNA viruses**. *Nature* 2006, **441**(7089):101–105.
34. Yoneyama M, Kikuchi M, Matsumoto K, Imaizumi T, Miyagishi M, Taira K, Foy E, Loo Y-M, Gale M, Akira S, *et al*: **Shared and Unique Functions of the DExD/H-Box Helicases RIG-I, MDA5, and LGP2 in Antiviral Innate Immunity**. *J Immunol* 2005, **175**(5):2851–2858.
35. Franchi L, Warner N, Viani K, Nuñez G: **Function of Nod-like receptors in microbial recognition and host defense**. *Immunol Rev* 2009, **227**(1):106–128.
36. Lamkanfi M, Kanneganti T-D: **Regulation of immune pathways by the NOD-like receptor NLRC5**. *Immunobiology* 2012, **217**(1):13–16.

37. Tschopp J, Schroder K: **NLRP3 inflammasome activation: the convergence of multiple signalling pathways on ROS production?** *Nat Rev Immunol* 2010, **10**(3):210–215.
38. Sadler AJ, Williams BR: **Interferon-inducible antiviral effectors.** *Nat Rev Immunol* 2008, **8**(7):559–568.
39. Arnheiter H, Skuntz S, Noteborn M, Chang S, Meier E: **Transgenic mice with intracellular immunity to influenza virus.** *Cell* 1990, **62**(1):51–61.
40. Haller O, Stertz S, Kochs G: **The Mx GTPase family of interferon-induced antiviral proteins.** *Microbes and Infection* 2007, **9**(14–15):1636–1643.
41. Bernasconi D, Schultz U, Staeheli P: **The Interferon-Induced Mx Protein of Chickens Lacks Antiviral Activity.** *J Interferon Cytokine Res* 1995, **15**:47–53.
42. von der Malsburg A, Abutbul-Honita I, Haller O, Kochs G, Danino D: **Stalk Domain of the Dynamitin-like MxA GTPase Protein Mediates Membrane Binding and Liposome Tubulation via the Unstructured L4 Loop.** *J Biol Chem* 2011, **286**(43):37858–37865.
43. Gao S, von der Malsburg A, Paeschke S, Behlke J, Haller O, Kochs G, Daumke O: **Structural basis of oligomerization in the stalk region of dynamitin-like MxA.** *Nature* 2010, **465**(7297):502–506.
44. Noteborn M, Arnheiter H, Richter-Mann L, Browning H, Weissmann C: **Transport of the murine Mx protein into the nucleus is dependent on a basic carboxy-terminal sequence.** *J Interferon Res* 1987, **7**(5):657–669.
45. Storset AK, Slettedal IÖ, Williams JL, Law A, Dissen E: **Natural killer cell receptors in cattle: a bovine killer cell immunoglobulin-like receptor multigene family contains members with divergent signaling motifs.** *Eur J Immunol* 2003, **33**(4):980–990.
46. Kelley J, Walter L, Trowsdale J: **Comparative Genomics of Natural Killer Cell Receptor Gene Clusters.** *PLoS Genet* 2005, **1**(2):e27.
47. Belov K, Sanderson CE, Deakin JE, Wong ES, Assange D, McColl KA, Gout A, de Bono B, Barrow AD, Speed TP, Trowsdale J, Papenfuss AT: **Characterization of the opossum immune genome provides insights into the evolution of the mammalian immune system.** *Genome Res* 2007, **17**:982–991.
48. Laun K, Coggill P, Palmer S, Sims S, Ning Z, Ragoussis J, Volpi E, Wilson N, Beck S, Ziegler A, et al: **The Leukocyte Receptor Complex in Chicken Is Characterized by Massive Expansion and Diversification of Immunoglobulin-Like Loci.** *PLoS Genet* 2006, **2**(5):e73.
49. Viertboeck BC, Schweinsberg S, Schmitt R, Herberg FW, Göbel TW: **The Chicken Leukocyte Receptor Complex Encodes a Family of Different Affinity Fcγ Receptors.** *J Immunol* 2009, **182**(11):6985–6992.
50. Anderson SK, Ortaldo JR, McVicar DW: **The ever-expanding Ly49 gene family: repertoire and signaling.** *Immunol Rev* 2001, **181**(1):79–89.
51. Takahashi T, Yawata M, Raudsepp T, Lear TL, Chowdhary BP, Antczak DF, Kasahara M: **Natural killer cell receptors in the horse: evidence for the existence of multiple transcribed LY49 genes.** *Eur J Immunol* 2004, **34**(3):773–784.
52. Wong E, Sanderson C, Deakin J, Whittington C, Papenfuss A, Belov K: **Identification of natural killer cell receptor clusters in the platypus genome reveals an expansion of C-type lectin genes.** *Immunogenetics* 2009, **61**(8):565–579.
53. Yoder JA, Litman GW: **The phylogenetic origins of natural killer receptors and recognition: relationships, possibilities, and realities.** *Immunogenetics* 2011, **63**(3):123–141.
54. Brown MG, Scalzo AA: **NK gene complex dynamics and selection for NK cell receptors.** *Semin Immunol* 2008, **20**(6):361–368.
55. Tomasello E, Blery M, Vely E, Vivier E: **Signaling pathways engaged by NK cell receptors: double concerto for activating receptors, inhibitory receptors and NK cells.** *Semin Immunol* 2000, **12**(2):139–147.
56. Vaidya SV, Mathew PA: **Of mice and men: Different functions of the murine and human 2B4 (CD244) receptor on NK cells.** *Immunol Lett* 2006, **105**(2):180–184.
57. LeMaout J, Le Discorde M, Rouas-Freiss N, Moreau P, Menier C, McCluskey J, Carosella ED: **Biology and functions of human leukocyte antigen-G in health and sickness*.** *Tissue Antigens* 2003, **62**(4):273–284.
58. Bodmer JG, Marsh SGE, Parham P, Erlich HA, Albert E, Bodmer WF, Dupont B, Mach B, Mayr WR, Sasazuki T, et al: **Nomenclature for factors of the HLA system, 1989.** *Hum Immunol* 1990, **28**(3):326–342.
59. Brown JH, Jardetzky TS, Gorga JC, Stern LJ, Urban RG, Strominger JL, Wiley DC: **Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1.** *Nature* 1993, **364**(6432):33–39.
60. Beck S, Trowsdale J: **Sequence organisation of the class II region of the human MHC.** *Immunol Rev* 1999, **167**(1):201–210.
61. Hughes AL, Nei M: **Evolutionary relationships of class II major-histocompatibility-complex genes in mammals.** *Mol Biol Evol* 1990, **7**(6):491–514.
62. Richman AD, Herrera M LG, Ortega-García S, Flores-Martínez JJ, Arroyo-Cabrales J, Morales-Malacara JB: **Class II DRB polymorphism and sequence diversity in two vesper bats in the genus *Myotis*.** *Int J Immunogenet* 2010, **37**(5):401–405.
63. Mayer F, Brunner A: **Non-neutral evolution of the major histocompatibility complex class II gene DRB1 in the sac-winged bat *Saccopteryx bilineata*.** *Heredity* 2007, **99**(3):257–264.
64. Schad J, Dechmann DKN, Voigt CC, Sommer S: **MHC class II DRB diversity, selection pattern and population structure in a neotropical bat species, *Noctilio albigentris*.** *Heredity* 2011, **107**(2):115–126.
65. Hein WR, Mackay CR: **Prominence of gamma delta T cells in the ruminant immune system.** *Immunol Today* 1991, **12**(1):30–34.
66. Six A, Rast JP, McCormack WT, Dunon D, Courtois D, Li Y, Chen C-IH, Cooper MD: **Characterization of avian T-cell receptor γ genes.** *Proc Natl Acad Sci U S A* 1996, **93**(26):15329–15334.
67. Butler JE, Wertz N, Zhao Y, Zhang S, Bao Y, Bratsch S, Kunz TH, Whitaker JO Jr, Schountz T: **The two suborders of chiropterans have the canonical heavy-chain immunoglobulin (Ig) gene repertoire of eutherian mammals.** *Dev Comp Immunol* 2011, **35**(3):273–284.
68. Flicek P, Armode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, et al: **Ensembl 2011.** *Nucleic Acids Res* 2011, **39**(Database issue):D800–806.
69. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, et al: **The Pfam protein families database.** *Nucleic Acids Res* 2012, **40**(D1):D290–D301.
70. Teeling EC, Springer MS, Madsen O, Bates P, O'Brien SJ, Murphy WJ: **A Molecular Phylogeny for Bats Illuminates Biogeography and the Fossil Record.** *Science* 2005, **307**(5709):580–584.
71. FastQC: <http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/>.
72. SnoWhite: <http://kdlugosch.net/software/>.
73. SeqClean: <http://compbio.dfci.harvard.edu/tgi/software>.
74. Lassmann T, Hayashizaki Y, Daub CO: **TagDust—a program to eliminate artifacts from next generation sequencing data.** *Bioinformatics* 2009, **25**(11):2839–2840.
75. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res* 2008, **18**(5):821–829.
76. Oases: *De novo transcriptome assembler for very short reads:* <http://www.ebi.ac.uk/~zerbino/oases/>.
77. Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.** *Bioinformatics* 2006, **22**(13):1658–1659.
78. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WE, Wetter T, Suhai S: **Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs.** *Genome Res* 2004, **14**(6):1147–1159.
79. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**(17):3389–3402.
80. Vacic V, Uversky VN, Dunker AK, Lonardi S: **Composition Profiler: a tool for discovery and visualization of amino acid composition differences.** *BMC Bioinformatics* 2007, **8**:211.
81. CateGOrizer: <http://www.animalgenome.org/tools/catego/>.
82. Hu Z-L, Bao J, Reecy JM: **CateGOrizer: A Web-Based Program to Batch Analyze Gene Ontology Classification Categories.** *Online J Bioinformatics* 2008, **9**(2):108–112.
83. *Generic gene ontology (GO) term mapper.* <http://go.princeton.edu/cgi-bin/GOTermMapper>.
84. Thompson JD, Gibson TJ, Higgins DG: **Multiple Sequence Alignment Using ClustalW and ClustalX.** In *Current Protocols in Bioinformatics*. Edited by John Wiley & Sons, Inc; 2002.
85. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4**(4):406–425.
86. Tamura K, Dudley J, Nei M, Kumar S: **MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0.** *Mol Biol Evol* 2007, **24**(8):1596–1599.
87. Wootton JC, Federhen S: **Statistics of local complexity in amino acid sequences and sequence databases.** *Comput Chem* 1993, **17**:149–163.

88. Vihinen M, Torkkila E, Riihonen P: **Accuracy of protein flexibility predictions.** *Proteins* 1994, **19**(2):141–149.
89. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, *et al*: **The Universal Protein Resource (UniProt).** *Nucleic Acids Res* 2005, **33**(Database issue):D154–D159.

doi:10.1186/1471-2164-13-261

Cite this article as: Papenfuss *et al.*: The immune gene repertoire of an important viral reservoir, the Australian black flying fox. *BMC Genomics* 2012 **13**:261.

**Submit your next manuscript to BioMed Central
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

