

Published in final edited form as:

*Nature*. 2013 October 10; 502(7470): . doi:10.1038/nature12511.

## Genome-wide signatures of convergent evolution in echolocating mammals

Joe Parker<sup>1,†,\*</sup>, Georgia Tsagkogeorga<sup>1,\*†</sup>, James A. Cotton<sup>1,2</sup>, Yuan Liu<sup>3</sup>, Paolo Provero<sup>4,5</sup>, Elia Stupka<sup>4</sup>, and Stephen J. Rossiter<sup>1,\*</sup>

<sup>1</sup>School of Biological and Chemical Sciences, Queen Mary, University of London, London E1 4NS, UK

<sup>2</sup>Current address: Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

<sup>3</sup>BGI-Europe, Ole Maaløes Vej 3, DK-2200 Copenhagen N, Denmark

<sup>4</sup>Center for Translational Genomics and Bioinformatics, San Raffaele Scientific Institute, Milano, Italy

<sup>5</sup>Department of Molecular Biotechnology and Health Sciences, University of Turin, Italy

### Abstract

Evolution is typically thought to proceed through divergence of genes, proteins, and ultimately phenotypes<sup>1-3</sup>. However, similar traits might also evolve convergently in unrelated taxa due to similar selection pressures<sup>4,5</sup>. Adaptive phenotypic convergence is widespread in nature, and recent results from a handful of genes have suggested that this phenomenon is powerful enough to also drive recurrent evolution at the sequence level<sup>6-9</sup>. Where homoplasious substitutions do occur these have long been considered the result of neutral processes. However, recent studies have demonstrated that adaptive convergent sequence evolution can be detected in vertebrates using statistical methods that model parallel evolution<sup>9,10</sup> although the extent to which sequence convergence between genera occurs across genomes is unknown. Here we analyse genomic sequence data in mammals that have independently evolved echolocation and show for the first time that convergence is not a rare process restricted to a handful of loci but is instead widespread, continuously distributed and commonly driven by natural selection acting on a small number of sites per locus. Systematic analyses of convergent sequence evolution in 805,053 amino acids within 2,326 orthologous coding gene sequences compared across 22 mammals (including four new bat genomes) revealed signatures consistent with convergence in nearly 200 loci. Strong and significant support for convergence among bats and the dolphin was seen in numerous genes linked to hearing or deafness, consistent with an involvement in echolocation. Surprisingly we also found convergence in many genes linked to vision: the convergent signal of many sensory genes was robustly correlated with the strength of natural selection. This first attempt to detect genome-wide convergent sequence evolution across divergent taxa reveals the phenomenon to be much more pervasive than previously recognised.

---

\*Corresponding author.

†Joint first authors

#### Authors' contributions

S.J.R. conceived the study and secured funding together with J.A.C. and E.S. J.P. conducted all phylogenetic, convergence and selection analyses with input from S.J.R., G.T. and J.A.C. Processing and analyses of sequence data was undertaken by G.T., with input from E.S., who also conducted gene ontology analyses with P.P. Raw sequence data was generated under direction of Y.L. and S.J.R. The paper was written and figures prepared by J.P. and S.J.R. with input from G.T., J.A.C. and E.S.

Echolocation is a complex phenotypic trait that has evolved independently in bats and whales, and which involves the production, reception and auditory processing of ultrasonic pulses for obstacle avoidance, orientation and hunting<sup>11,12</sup>. Recent phylogenetic studies have shown that echolocating bats are not a true group; one lineage also contains the non-echolocating Old World fruit bats (family Pteropodidae), implying echolocation has evolved at least twice in bats, or was lost early in the evolution of Old World fruit bats<sup>13-15</sup>. New evidence supports the former scenario: divergent clades of echolocating bats appear to have undergone convergent amino acid replacements in several genes implicated in hearing<sup>8,16-18</sup>. Furthermore, some candidate hearing genes also show parallel changes in echolocating bats and whales, again suggesting roles in high frequency hearing<sup>10,16,18,19</sup>. Other genes that might function in echolocation have been identified from screens for selection in bat<sup>20</sup> and cetacean sequence data<sup>21</sup>. Here, using the evolution of echolocation as a model of phenotypic convergence, we investigated the extent to which parallel changes have occurred across the genome during the independent evolution of echolocation in bats and cetaceans.

We undertook genome sequencing of four divergent bat species, including both echolocating and non-echolocating forms. From the suborder Yinpterochiroptera we sequenced the greater horseshoe bat *Rhinolophus ferrumequinum* and the greater false vampire bat *Megaderma lyra*, which exhibit ‘constant frequency’ (CF) and ‘frequency modulated’ (FM) echolocation, respectively (for details of calls, see<sup>15,22</sup>). From this suborder we also sequenced the non-echolocating straw-coloured fruit bat *Eidolon helvum*, to which we added published draft genome data from a second non-echolocating fruit bat, the large flying fox *Pteropus vampyrus*. From the second suborder, Yangochiroptera, we sequenced the CF echolocating Parnell’s moustached bat *Pteronotus parnellii*, and added published data from the FM echolocating little brown bat *Myotis lucifugus*.

For each of our four focal bat species, we generated paired-end short read sequence data on a Hi-Seq™ 2000 platform (Illumina Inc.), assembled the raw reads *de novo* into contigs using CLC bio (Cambridge, MA), and then built scaffolds in SOAPdenovo (see Supplementary Information for details). Short-read data have been deposited into the Short Read Archive under accession nos. SRR924356, SRR924359, SRR924361 and SRR924427. We conducted homology-based gene prediction and identified 20,424 genes for *R. ferrumequinum*, 20,043 for *M. lyra*, 20,455 for *E. helvum* and 20,357 for *P. parnellii*. Screening for single-copy (1-to-1) orthologous protein-coding nuclear genes conserved across eutherian mammals identified 7,612 genes present in each of our four draft genomes (see Supplementary Information for details).

To build a mammal-wide alignment of orthologous coding gene sequences (CDSs) we retrieved the CDS of each locus from 18 published mammal genomes from Ensembl (<http://www.ensembl.org/> Release 63), covering a broad taxonomic range and including *M. lucifugus* and *P. vampyrus* (see Supplementary Information) as well as the echolocating common bottlenose dolphin *Tursiops truncatus*. Individual gene datasets were built and aligned in frame as codons with all ambiguous sites and codons removed (see Supplementary Methods). To avoid potential errors that could arise either during sequencing or data assembly, which could adversely affect phylogenetic and molecular evolution analyses, we focused on all identified genes that, following clean up, contained no missing data or gaps in any of the newly sequenced bats. In total we generated alignments for 2,326 CDSs, each spanning at least 450 bp, and containing a minimum of six bat species (2% of alignments had missing data from *P. vampyrus* due to its lower coverage) as well as the bottlenose dolphin and the following five other mammals: dog *Canis familiaris*, horse *Equus caballus*, cow *Bos taurus*, mouse *Mus musculus* and human *Homo sapiens*.

To detect genome-wide sequence convergence between echolocating lineages, we built an analytical pipeline based on maximum likelihood (ML) phylogenetic reconstruction<sup>8,10,19</sup>. In this method, we examined each amino acid along the alignment of a given CDS, and measured its fit (sitewise log-likelihood support; SLS) to the commonly accepted species tree<sup>23-25</sup> (hereafter termed  $H_0$ ) and to two alternative topologies in which we forced echolocating taxa into erroneous monophyletic clades representing different convergence hypotheses (see Fig. 1a, for details see Fig. S1 and Supplementary Methods):  $H_1$  corresponds to all echolocating bats in a monophyletic group ('bat-bat convergence') and  $H_2$  to all echolocating mammals together in a monophyletic group ('bat-dolphin convergence'). Using this approach we obtained the SLS values of all amino acids under three different tree topologies. The difference in SLS for a single site under the species tree and a given convergent tree with an identical substitution model denotes the relative support for the convergence hypothesis; for example,  $\Delta SLS(H_1) = SLS(H_0) - SLS(H_1)$  (where negative  $\Delta SLS$  implies support for convergence; see Fig S2). We quantified the extent of sequence convergence at each locus by taking the mean of its  $\Delta SLS$  values, and found 824 loci with mean support for  $H_1$  and 392 for  $H_2$ . Using simulations we confirmed that these convergent signals were not due to neutral processes and were robust to the substitution model used (see Supplementary Methods).

We ranked the mean  $\Delta SLS$  for all 2,326 loci under both convergence hypotheses and, to assess the performance of our method, inspected the rank positions of seven hearing genes that have previously been shown to exhibit convergence and/or adaptation in echolocating mammals: *Prestin* (*SLC26A5*); *TMCI*; *KCNQ4* (*KQT-4*); *Pjvk* (*DFNB59*); *Otoferlin*; *Pcdh15*; *Cdh23* (see Supplementary Methods). *Prestin* was ranked #43 ( $H_1$ ) and #22 ( $H_2$ ), while several other loci were also ranked highly in the distribution of convergence support values (see Fig. 1b). In addition to these, we also found several other hearing genes in the top 5% supporting  $H_1$  (*ITM2B*, *SLC4A11*) and  $H_2$  (*COCH*, *ITM2B*, *ERCC3* and *OPA1*). Since bats and cetaceans are also known to have undergone shifts in spectral tuning and other adaptations in response to living in low light environments<sup>26-28</sup>, we also examined the position of genes implicated in vision and found four such loci in the top 5% of genes supporting  $H_1$  (*LCAT*, *SLC45A2*, *RABGGTB*, and *RPI*) and three supporting  $H_2$  (*JMJD6*, *SIX*, and *RHO*; see examples in Fig. 1b and Supplementary Tables S2 and S3).

We tested statistically whether the strength of sequence convergence among echolocating bats, and between echolocating bats and the dolphin, is greater in hearing genes than in other genes (for locus selection see Supplementary Information). For each phylogenetic hypothesis, we averaged the mean  $\Delta SLS$  values of all 21 genes in our dataset that are listed as linked to either hearing and/or deafness in any taxon based on published functional annotations (see Supplementary Information). By comparing our observed values to null distributions of corresponding values obtained by randomisation, we found that hearing genes had significantly more negative average values than expected by chance for bat-dolphin convergence ( $H_2$ :  $z = -0.0194$ ,  $P < 0.05$ ). We repeated this method for 75 genes listed as involved in vision and/or blindness, and found support, although weaker, in both cases of phenotypic convergence ( $z = -0.0020$ ,  $P = 0.055$  and  $z = -0.0097$ ,  $P = 0.09$ ). Loci previously reported to have association with echolocation had strong support by randomisation for both hypotheses ( $P = 0.01$  in both cases).

We inspected direct protein-protein interactions among the 117 genes (top 5%) for  $H_2$  in networks of published interactions and found that 17 of these loci formed a single large network of direct interactions (for details see Supplementary Information and Fig. S9) centred around *TP63* (*p63*) and *CDK1* (*p34*); *p63* has been shown recently to be involved in cochlea development<sup>29</sup> and *CDK1* has been shown to be important for development and regeneration of hair cells in the inner ear<sup>30</sup>. We also scanned databases of tissue-specific

RNA expression in humans, finding some genes with elevated hypothalamus expression in human orthologues (Supplementary Information and Table S13).

Most of the loci supporting the monophyly of echolocating bats, or the clade of echolocating bats plus dolphin, have no known roles in the sensory perception of sound or light. Yet given that many of these loci encode proteins with poorly characterised functions, a role in hearing or vision cannot be ruled out and in this respect it is noteworthy that highly ranked genes for  $H_1$  (top 5%) included five solute carrier proteins related to Prestin, a motor protein that drives the cochlear amplifier<sup>8,19</sup> (see Supplementary Table S2). Other loci that appear to support convergence among echolocating taxa can be more confidently ruled out as having roles in sensory perception; however, some of these loci may instead be associated with phenotypic traits that are correlated with aspects of echolocation.

Sequence convergence in loci related to shared phenotypic traits provides compelling but indirect evidence that genome-wide examples of convergence reported here are likely to be due to adaptive selection rather than neutral evolution. Analysis of the distribution of support for convergence within each locus suggested mean support is driven by a few sites per locus (see Supplementary Information). For echolocating taxa under each convergent scenario, we directly estimated the strength of selection at each site as the ratio (omega,  $\omega$ ) of the rate of non-synonymous substitutions (dN) to the rate of synonymous substitutions (dS;  $\omega > 1$  being indicative of molecular adaptation; see Supplementary Information). To test whether sitewise support for a given hypothesis is driven by selection, we correlated absolute sitewise SLS and sitewise  $\omega$  and found a strong relationship after correcting for locus identity ( $H_1$ :  $P = 0.0336$ ;  $H_2$ :  $P < 0.001$ ).

Previous work on *Prestin* suggested that sequence convergence among echolocating taxa was a consequence of positive selection<sup>19</sup>. To determine the extent to which adaptive convergence occurs across the genome, for each locus we fitted the linear relationship between SLS and corresponding sitewise  $\omega$  in echolocating lineages for those sites that show different selection pressures between echolocators and other taxa. We hypothesized that loci with sites undergoing diversifying selection for convergence ('adaptive convergence') would exhibit a negative correlation, whereas those loci with sites undergoing divergence driven by diversifying selection, but under the accepted species topology ('adaptive divergence'), would exhibit a positive correlation. Based on whether the 95% confidence interval (c.i.) of the slope was below zero, for  $H_1$  we classified 92 loci as putatively adaptive convergent; conversely, in 111 loci, the 95% c.i. of the slope was greater than zero suggesting adaptive divergence. Using the same approach for  $H_2$ , we classified 59 loci as adaptive convergent and 212 loci as adaptive divergent (Fig. 2). The larger number of adaptive divergent loci seen under  $H_2$  reflects the stronger average relative support for  $H_0$  across the dataset in  $H_0H_2$  comparisons, as opposed to  $H_0H_1$  comparisons; to be expected since  $H_2$  is a more radical rearrangement of the species phylogeny.

To more fully explore the relationship between signatures of convergence and adaptive evolution, we predicted the level of convergent evolution expected (SLS) after prolonged diversifying selection ( $\omega = 2$ ) using the fitted regression models for each locus. Hearing and vision genes (including those encoding the solute carriers *SLC44A2* and *SLC4A11*, associated with deafness, and the integrin *ITGAL*, associated with blindness) showed significantly stronger convergence in these models than the background loci (see Supplementary Information). The genes identified in this analysis show a pattern of substitutions under differing selection regimes in echolocating and non-echolocating lineages that tend to have high  $\omega$  and be convergent in echolocating lineages; we therefore suggest that sustained adaptive selection in these loci is likely to reinforce the signal for

convergence as parallel changes accrue, in contrast to neutrally-driven or adaptive divergence, which would tend to reinforce support for the species phylogeny over time.

Our genome-wide analysis shows that natural selection has acted on three echolocating lineages (cetaceans and two separate bat lineages) to produce a complex pattern of changes in protein sequence, including both divergent and, more surprisingly, extensive convergent changes. Many of these changes are in genes that may be associated with the shift in primary sensory modality (between vision and echolocation), either directly or through the associated complex changes in ecology and natural history. Furthermore, this work identifies clear targets for future experimental work, for instance tissue-specific RNA expression analyses. This study represents the first systematic attempt to provide a framework for the genomic analysis of sequence convergence associated with independently shared phenotypes. Our findings strongly suggest that, despite the plethora of recent papers reporting sequence convergence in particular candidate genes, the importance of this mode of molecular evolutionary change is relatively underappreciated, and under-exploited in seeking to understand the genetic basis of complex traits such as echolocation.

## Methods Summary

### Taxonomic coverage, sequencing and assembly

Genome-wide short-read sequences (HiSeq, 500BP inserts) were generated and assembled for one non-echolocating (*Eidolon helvum*) and three echolocating bats (*Rhinolophus ferrumequinum*, *Megaderma lyra* and *Pteronotus parnelli*). The former three belong to the suborder Yinpterochiroptera and the latter to the Yangochiroptera. 7,612 coding sequences showed 1-to-1 orthology with *Homo* sequences and were present in all four taxa. Of these, 2,326 loci were aligned in MAFFT with published data from 17 additional taxa, including the non-echolocating bat *Pteropus vampyrus*, echolocating bat *Myotis lucifugus* and echolocating dolphin *Tursiops truncatus*.

### Convergence pipeline

We built a pipeline to estimate the strength of natural selection (dN/dS ratio,  $\omega$ ) and support for convergence (SSLS) for each amino-acid by maximum-likelihood.  $\omega$  values were estimated using published software; SSLS was calculated as the difference in fitted sitewise log-likelihoods under two competing hypotheses: a 'null' phylogenetic tree, reflecting the consensus species tree, and one of two 'alternative' phylogenies, in which either all echolocating bats, or all echolocating taxa (echolocating bats plus dolphin), were artificially forced into a monophyletic clade. Ancestral amino-acid states were inferred using parsimony.

### Null distributions

We compared observed SSLS values to SSLS estimated from (i) neutrally-evolving amino-acids simulated by Markov-chain Monte Carlo mixture models ( $n=1,000$ ), and (ii) random 'alternative' topologies ( $n=100$ ).

### Locus-wise adaptation / convergence regressions

For each locus the relationship between selection and convergence was modelled using least-squares regression. 95% confidence intervals were calculated for the line equation parameters. The fitted line equations were extrapolated to model the convergence signal under strong adaptive selection ( $\omega=2$ ).



## Hearing and vision genes

We used the gene ontology database DAVID (<http://david.abcc.ncifcrf.gov>) to identify which of our loci were putatively associated with hearing or vision. Simulation by randomisation ( $n=1,000$ ) was used to compare the observed mean convergence values among these sensory loci to null distributions.

## Methods

### Taxonomic coverage

We collected new genome-wide sequence data from four bat species, selected from the two suborders and encompassing the paraphyly of echolocating bat lineages (see ref. 13). From the suborder Yinpterochiroptera we studied the non-echolocating Old World fruit bat *Eidolon helvum* (family Pteropodidae) and two laryngeal echolocating species, *Megaderma lyra* (Megadermatidae) and *Rhinolophus ferrumequinum* (Rhinolophidae). From the suborder Yangochiroptera we studied the laryngeal echolocating species *Pteronotus parnellii* (Mormoopidae) that has independently evolved constant frequency echolocation.

From Ensembl (<http://www.ensembl.org/>), we also obtained sequence data from two additional bats - the laryngeal echolocating species *Myotis lucifugus* (Yangochiroptera; Broad Institute) and the non-echolocating Old World fruit bat *Pteropus vampyrus* (Yinpterochiroptera; Baylor College of Medicine Human Genome Sequencing Center) - as well as the echolocating bottlenose dolphin *Tursiops truncatus*. Genomic sequences from 15 additional mammal species were downloaded from Ensembl giving a total of 22 mammals (listed in Table S1).

### Phylogenetic hypotheses tested

To investigate the prevalence of convergent evolution at a genome-wide level associated with the independent evolution of echolocation in bats and cetaceans, we used a method that builds on maximum-likelihood phylogenetic reconstruction. This method compares, for a given sequence alignment of orthologous coding sequences (CDS), the goodness-of-fit of the accepted phylogenetic tree with that of an alternative convergent hypothesis (in this case, in which echolocating taxa were forced into a spurious monophyletic clade). From our dataset, we identified and tested three hypotheses: (a)  $H_0$ , the commonly-accepted species phylogeny (e.g.<sup>13,23-25</sup>) in which cetaceans (represented in our data set by the common bottlenose dolphin *Tursiops truncatus*) are nested within the even-toed ungulates in the order Cetartiodactyla, and the order Chiroptera is split into the suborders Yangochiroptera and Yinpterochiroptera, with paraphyly of bat laryngeal echolocation<sup>13</sup>; (b)  $H_1$ , or ‘bat-bat echolocation convergence’ (monophyly of all echolocating bats in the dataset); and (c)  $H_2$ , or ‘bat-dolphin convergence’ (monophyly of all echolocating mammals in the dataset). All three phylogenetic hypotheses are shown in Figure 1. The scale bar (in amino acid substitutions) is provided for approximate reference only, since branch lengths were optimised at runtime.

Because the  $H_2$  (bat-dolphin) hypothesis is necessarily a radical rearrangement of the any commonly-accepted species topology, and the concept of an ‘exact branching order’ or the ‘true’ topology does not apply in this case, we proposed a number of separate but related versions of this hypothesis, all of which were evaluated equally in the analysis. In each case the rest of the mammalian species phylogeny was fixed, as in the  $H_1$  hypothesis. In the first case we constrained all five echolocating taxa to a single ancestral node (‘hard polytomy’); secondly we enumerated the seven bifurcating trees that are possible where the position of *T. truncatus* is free to vary, but the suborders of echolocating bats – Yangochiroptera (*P. parnellii* & *M. lucifugus*) and Yinpterochiroptera (*R. ferrumequinum* and *M. lyra*) – were

preserved. A final topology was specified as a soft polytomy, with the resolution of the clade of echolocators being resolved by RAxML at runtime, with the rest of the phylogeny remaining constrained. A majority clade-consensus (MCC) summary phylogeny was constructed from these 2,326 inferred soft-polytomy  $H_2$  trees using TreeAnnotator v1.7.4 (in the BEAST v1.7.4 distribution<sup>31</sup>). This phylogeny recovered the Yangochiroptera and Yinpterochiroptera clades of echolocating bats with good (>50%) node support. When we compared the goodness-of-fit of all phylogenies (as opposed to pairwise comparison relative to the species phylogeny  $H_0$ ) we found the species phylogeny was preferred at 1,170 loci (55%), with the bat-bat phylogeny  $H_1$  preferred next most often (548 loci; 26%). The soft-polytomy version of  $H_2$  (resolved by RAxML) was the preferred phylogeny among 50% of the remaining loci, with remaining support equally split between the other  $H_2$  versions. We therefore adopted the soft polytomy, RAxML resolved version of  $H_2$  as our main bat-dolphin hypothesis.

### Sequencing and dataset assembly

Novel sequence data from the four bat species listed above were generated by BGI on an Illumina Genome Analyzer platform (Illumina Inc.), based on genomic libraries of 500 bp insert sizes. Using this method we obtained approximately 33 to 41 Gb of short read sequence data per species.

The CLC *de novo* algorithm (CLC bio, Cambridge, MA) was used for assembling raw reads into contigs using different *k*-mer size values ranging from 32 to 50. The assembled contigs from the CLC output were then processed using the module *prepare* of the SOAP package in order to do scaffold assembly using the *scaff* command of SOAPdenovo. Finally, gaps were filled using the GapCloser<sup>32</sup> tool. The resulting assemblies consisted of between 210,080 and 315,526 genomic sequences (depending on species), with an average depth of coverage of 17× to 18×. Estimated genome size was approximately ~2Gb in all four bats, whereas contiguity (as assessed by the N50 statistic) ranged from 16,292 bp (*M. lyra* genomic sequences) to 27,140 bp (*E. helvum*). Homology-based gene prediction analyses using the genBlastG<sup>33</sup> tool recovered 20,424 gene models for *R. ferrumequinum*, 20,043 for *M. lyra*, 20,455 for *E. helvum* and 20,357 for *P. parnellii*, in line with published gene content values for other mammals<sup>34</sup>. The completeness/contiguity of the gene representation was evaluated using the CEGMA (Core Eukaryotic Genes Mapping Approach) pipeline<sup>35,36</sup> and found ranging across species between 61.29 to 77.02 % and 90.32 to 96.77% for complete and partial genes, respectively. These compared well to the published *M. lucifugus* genome; when we analysed that genome using CEGMA the comparable completeness/contiguity scores for complete and partial genes were in the middle of this range (62.9% and 91.5%, respectively).

To identify genes adequate for systematic phylogenetic-based analyses of convergent sequence evolution, we next filtered the above predictions for single copy orthologous protein-coding genes conserved across the Eutheria. This was achieved by performing reciprocal blast searches against a database consisting of the gene models for the four bats, and using as queries the human sequences of 11,185 genes reported as 1-to-1 or apparent 1-to-1 orthologues between the human and *Myotis* genomes in Ensembl databases (<http://www.ensembl.org/> Release 63). In total we determined 7,612 1-to-1 orthologous genes, from which the longest coding sequences (CDSs) were then retrieved from Ensembl for the 18 additional mammalian genomes (Table S1).

### CDS alignment

Coding gene sequences (CDS) of individual loci were built and aligned as codons using a modified version of transAlign<sup>37</sup> incorporating MAFFT<sup>38</sup>, such that all sequences remained

in the correct reading frame. Any ambiguously aligned sites, and codons with excessive numbers of gaps, were removed from each gene alignment using Gblocks<sup>39</sup> under the following options:  $-t=c$   $-b1="\$b1"$   $-b2="\$b1"$   $-b3=1$   $-b4=6$   $-b5=h$ , where  $b1=70\%$  of the sequences sampled in the data set.

In order to avoid potential biases due to either sequencing or assembly errors, for all phylogenetic and molecular evolution analyses, we chose to focus on only a subset of the identified genes. Specifically, we restricted our downstream analyses on datasets, which after filtering out of ambiguous sites, showed no missing data in any of the sampled bats. The exception to this rule was *P. vampyrus*, which, because of its comparatively lower genome coverage, was missing in around 2% of CDS alignments. All final CDS alignments used in our analyses were characterised by a minimum length of 450bp (or 150 codons/ amino-acids) and included a minimum of six bat species, the dolphin and the additional following mammals as outgroups: *Canis familiaris*, *Equus caballus*, *Bos taurus*, *Mus musculus* and *Homo sapiens*. Of the 2,326 loci examined, 642 were also included in the analysis of Zhang *et al* (2013)<sup>20</sup>.

### Sets of genes associated with hearing and vision

We interrogated our full CDS dataset for loci that have been implicated in aspects of sensory perception of sound, including those that have been linked to deafness and/or ear development. For this we searched annotation databases hosted on the gene ontology (GO) site DAVID (<http://david.abcc.ncifcrf.gov><sup>40</sup>) and cross-referenced all of our CDSs with the terms ‘hearing’ or ‘deafness’. Using this approach, 23 ‘hearing loci’ were identified: *TYR*, *SLC4A11*, *NECAP1*, *COCH*, *JAG1*, *HOXA1*, *GTF3C2*, *PROX1*, *GGA3*, *MKKS*, *SLC44A2*, *ITM2B*, *EDNRB*, *FBXO11*, *PI4KB*, *DISP1*, *ERCC3*, *HESX1*, *FZD6*, *BDNF*, *NF2*, *OPA1* and *DFNB59*. We refer to this set of genes as “hearing”.

We repeated annotation searches of our full set of CDSs for associations with vision, this time using the keywords ‘vision’ or ‘blindness’. In total, 75 loci were classified: *PAX6*, *MED24*, *JMJD6*, *ATP6AP1*, *MCM2*, *TRAF4*, *GPC4*, *CIC*, *RDH8*, *IMPG2*, *CAD*, *RPGRIP1*, *HPS4*, *RABGGTA*, *RP2*, *CLN5*, *RPGRIP1L*, *VPS18*, *RP1*, *FZR1*, *GLI3*, *UNC119*, *GLRB*, *MYF5*, *COPZ1*, *MCM3*, *TTK*, *HMGCR*, *NPHP3*, *PDC*, *RPE65*, *PRPF3*, *ELOVL4*, *TGIF2*, *TCTN3*, *TGFBI*, *OPN4*, *ECD*, *NEUROD4*, *BBS2*, *PAICS*, *APC*, *LAMC1*, *SKIL*, *PDCL*, *RABGGTB*, *IFT172*, *BBS4*, *INTS7*, *LGSN*, *ZEB1*, *PELO*, *SMARCA5*, *KIT*, *CNNM4*, *GJD2*, *CCT3*, *RHO*, *RFC4*, *SLC45A2*, *VPS39*, *CTNNB1*, *STAT3*, *ADRA1B*, *GPRC5C*, *SP3*, *PRPF8*, *PVRL3*, *RRH*, *BCOR*, *SIX6*, *DRD1*, *NHS*, *TOPORS* and *LCAT*. We refer to this set of genes as “vision”.

In addition, several loci associated with the sensory perception of sound have previously been reported as convergent for bat-bat or bat-whale echolocation in the literature: *Prestin*<sup>19</sup>; *KCNQ4* (*KQT-4*)<sup>17</sup>; *Pcdh15*, *Cdh23* and *Otoferlin*<sup>18</sup>; and *Pjvk* and *Tmc1*<sup>16</sup>. We downloaded the sequences published in these studies (including focal echolocating taxa and background sequences from other mammalian species as in Table S1) from NCBI using the published accession numbers. We aligned each locus by MUSCLE<sup>41</sup>, and manually inspected them for quality. This included an initial phylogenetic step to check for long-branch effects where a *de novo* topology was estimated under ML (maximum likelihood) using RAxML (build 7.6.2; SSE3.MPI, compiled from source (<https://github.com/stamatak/standard-RAxML>)<sup>42,43</sup> and the PROTCATDAYHOFF model of amino acid substitution. All seven loci were retained following this step and analysed alongside the automatically assembled CDS datasets. We refer to these manually assembled alignments as “published”.



## Analysis pipeline

To detect signatures of molecular convergence in genomic data, we compiled an analysis pipeline consisting of previously released software for phylogenetic tree manipulation, phylogenetic reconstruction and codon model analyses in a Maximum Likelihood (ML) framework, as well as of a set of utility classes (available on request) for data handling, parsing and model/hypothesis testing. Our phylogenetic approach differs from genome-wide SNP comparisons for the detection of parallelism within intraspecific populations<sup>44</sup>, in that codons' phylogenetic histories are evaluated and compared separately, but aggregated over each locus; we also use simulation to establish a reference null distribution for each locus, and compare observed convergence values for a given test phylogenetic hypothesis to an expected distribution derived from 100 additional phylogenies. This framework incorporates RAxML (build 7.6.2; SSE.MPI, compiled from source (<https://github.com/stamatak/standard-RAxML><sup>42,43</sup>) and a modified build of version 4.4b of PAML<sup>45</sup>; available on request) where input/output (I/O) functions were adjusted to facilitate parallel cluster implementation. The main algorithms remained unchanged and in testing gave identical output to that produced using executables distributed by the authors; available on request. All analyses were conducted on a mixture of 32- and 64-bit processors at the 3000-node Queen Mary GridPP High Throughput Cluster hosted by the Physics Department at Queen Mary, University of London. Supplementary Figure S1 gives a schematic representation of the pipeline workflow:

**(i) Input**—For each locus, the multiple sequence alignment was first filtered to remove ambiguous or incomplete codons, as well as premature stop codons. Gaps were retained. The alignment was also checked to determine how many of the 22 possible species were present, and any absent taxa were pruned from the species tree  $H_0$  and the convergence trees  $H_1$  and  $H_2$  using NewickUtilities<sup>46</sup> (see Figure S1(i)).

**(ii) De novo phylogeny estimation**—We generated a separate *de novo* phylogeny for each locus using RAxML 7.6.2<sup>42,43</sup> under the model PROTCATDAYHOFF in rapid-search mode and 10 separate random start trees. The *de novo* tree was used as an independent estimate of the tree size. SSLS was subsequently calculated under this phylogeny as described below for  $H_1$  and  $H_2$ ; see Figure S1(ii). The soft polytomy present in the  $H_2$  hypothesis (the four echolocating bats plus *T. truncates*) was also resolved in a separate RAxML search using a constrained subtree, also under PROTCATDAYHOFF, in this step.

**(iii) Model fitting**—We fitted the checked alignment data to the  $H_0$ ,  $H_1$ ,  $H_2$  and *de novo* topologies using our modified build of the aaml program in PAML 4.4<sup>45</sup> under the WAG + model with estimated amino-acid frequencies. We also implemented the JONES and DAYHOFF models of amino-acid substitution. However, topology comparison requires marginalization of the site likelihoods with respect to substitution model, so that competing phylogenies' goodness-of-fit may be directly compared. Pilot studies of available alignments of orthologous coding sequences<sup>47</sup> showed congruence in relative SSLS estimates for each locus under different models of substitution (available on request). We also repeated our complete SSLS analyses separately under JONES and DAYHOFF; the Pearson's correlation coefficients between SSLS values for WAG-JONES, WAG-DAYHOFF, and JONES-DAYHOFF were 0.746, 0.979 & 0.742 respectively for  $H_0$ - $H_1$ ; and 0.917, 0.892 & 0.981 respectively for  $H_0$ - $H_2$ . We therefore determined to use the WAG model of substitution for all loci, optimising model parameters separately for each locus; see Figure S1(iii).

**(iv) Convergence hypotheses fitting**—The two hypothesised convergent phylogenies  $H_1$  and  $H_2$  were then fitted to the data as for the species tree  $H_0$  (Figure S1 (iv)). Since the

data and the substitution models were the same, the difference in likelihood between two phylogenies reflects the strength of support for each in the data (see below).

**(v) Comparison of sitewise log-likelihood support ( $\Delta$ SSLs)**—We used the mean SSLs of all sites in a locus as the primary measure of strength of support for convergence in this study. This statistic compares the goodness-of-fit of a pair of phylogenetic trees under a given model of evolution at every site in a DNA or amino acid alignment (see Supplementary Figure S2). Firstly the log-likelihood of the phylogeny (S2a) and substitution model, given the data (S2b), is calculated for every site in the alignment using ML (see above). Site-specific likelihood support, SSLs, was then calculated:

$$\Delta SSLS_i = \ln L_{i,H_0} - \ln L_{i,H_a} \quad (\text{Equation S1})$$

Where SSLs for the  $i^{\text{th}}$  site is given by the difference in log-likelihood units between the log-likelihood of the  $i^{\text{th}}$  site under  $H_0$  (the species tree) and  $H_a$  (the alternative tree; one of  $H_1$  or  $H_2$ ). By this definition, sites with a better model fit to  $H_0$  (the species tree) will have a positive SSLs, while sites with a better fit to the convergent topologies  $H_1$  or  $H_2$  will have a negative SSLs (Supplementary Figure S2c). The distribution of overall signal in the locus indicates the strength of support for convergence. In particular, loci with a negative mean SSLs show net sitewise signal for convergence (Supplementary Figure S2d). Boxplots showing the distribution of mean SSLs by hypothesis are shown in Figure S3; the top 5% of loci by mean SSLs for  $H_1$  and  $H_2$  are shown in Tables S2 and S3, respectively ( $n=805,053$ ). As expected, mean SSLs for the  $H_1$  and  $H_2$  hypotheses are positive, indicating that a majority of sites in the dataset are not convergent. Equally, mean SSLs for the hypotheses defined by the *de novo* tree comparison is negative, indicating that the topology that was directly fitted to the data has slightly better goodness-of-fit in many cases.

To investigate the nature of the convergence signal at all sites across this genomic dataset, we plotted the empirical cumulative distribution function of observed SSLs for  $H_1$  and  $H_2$  for the sitewise dataset (shown in Supplementary Figures S4 and S5, respectively). For both  $H_1$  and  $H_2$ , 95% of the sitewise SSLs observations were within  $\pm 0.5 \ln L$  units of the mean. However, some sites displayed large SSLs observations: absolute log-likelihood SSLs of  $>1 \ln L$  unit were observed for 1,828 and 26,342 amino acids in  $H_1$  and  $H_2$ , respectively. Furthermore, mean variance in sitewise convergence measured by locus was larger than the mean sitewise convergence measured across the whole dataset. Together, these indicate that the distribution of SSLs within a gene is aggregated; since the mean locus length in our dataset is relatively short but with a large variance (mean number of amino acids 346.8; s.d. 242.4) this suggests that small numbers of large sitewise SSLs may influence mean locus SSLs disproportionately in short loci.

**(vi) Simulation of expected sitewise  $\Delta$ SSLs distribution**—Homoplasious amino acid replacements may arise by neutral processes and, therefore, we used simulation to determine whether sitewise convergence was more significant than expected by chance. For each locus, we generated an expected distribution of likelihood differences as follows: using Phylobayes 3.3f<sup>48</sup>, we first used the ‘pb’ MCMC sampler to obtain the posterior distribution of substitution model parameters under the CAT GTR model, constraining the topology to the species ( $H_0$ ) tree. Each locus used one chain of 6000 steps, sampling every 10 and discarding the first 1000 as burn-in. We then used the ‘ppred’ function to simulate alignments using the model parameters from the samples in the stationary posterior distribution. Each simulated dataset therefore contained identical numbers of sites and taxa to the observed alignment; and since we fitted a mixture model, sites’ heterogeneity

parameters should reflect heterogeneity in the observed sequences. To generate an expected distribution, these replicates were analysed identically in the pipeline.

To determine how many replicates to use to form the expected distribution for each locus, we first simulated 50 alignments as described above from six loci that were representative of our dataset in terms of alignment length, heterogeneity, SSLS and tree length: ENSG00000008515, ENSG00000095906, ENSG00000121900, ENSG00000167671, ENSG00000170476 and ENSG00000173627. We calculated their sitewise SSLS values in the pipeline and then compared a single randomly-selected replicate's SSLS distribution to that of a variable number of replicates' pooled SSLS values in a Kolmogorov-Smirnov test. Replicates ( $1 < n < 50$ ) were sampled without replacement. Since the test's  $D$  statistic measures the largest difference between the two distributions, it will correlate with the smoothness of their step functions. We determined that smoothness monotonically increased as more replicates were used to construct the reference distribution, with 20-30 replicates providing a reference distribution that was comparably smooth to that derived from 50 replicates (available on request); we repeated these analyses to calculate means and variances. A purely numerical simulation in R using more replicates (10,000) but SSLS values simulated from a Normal distribution parametrised on the alignments' SSLS distribution's means and variances gave a similar result (not shown).

**(vii) Determining the significance of  $\Delta$ SSLS signals**—To ascribe confidence to our observed sitewise SSLS measurements for the trees of interest, we followed a two-stage process. First, we measured the sitewise SSLS for the tree comparison of interest  $H_0-H_a$ , described above. Next, we performed the same comparison on the simulated datasets, collated their SSLS values and calculated their stepwise empirical cumulative density function, with linear interpolation. This allowed us to calculate the cumulative probability of the observed SSLS under the null distribution. We define  $U$ , the unexpectedness of an observed site  $j$ 's SSLS comparison of the species topology,  $H_0$  and an alternative topology  $H_a$  as:

$$U = 1 - cdf(\Delta\text{SSLS}_{H_0-H_a | j}) \quad (\text{Equation S2})$$

**(viii) Correction for expected  $U$  across random (control) phylogenies**—The unexpectedness measure  $U$  quantifies the significance of a given site's convergence signal, or more explicitly, the cumulative probability of the observed log-likelihood difference between the two topologies, given the species topology is correct. However, in cases where the species topology itself is distant from the maximally-likely topology, or where the molecular signal is weak, SSLS scores  $< 0$ , indicating preference for the alternative topology, may arise spuriously. In this scenario, differential support for *any* alternative topology may be possible where the signal for the species topology is weak. To control for these cases, we generated 100 additional control phylogenies by resampling the  $H_1$  phylogeny taxon labels without replacement, obtained the mean sitewise  $U$  across this random set  $U_p$ , and calculated the controlled sitewise  $U$ ,  $U_c$  as:

$$U_c = U - \text{mean } U_c \quad (\text{Equation S3})$$

For each site  $j$ . These were summed across the locus and their arithmetic mean calculated. The random-tree controlled unexpectedness,  $U_c$ , therefore takes values on  $[-1, 1]$ , where values greater than zero indicate that the observed SSLS signal for the tree of interest,  $H_a$ , is both stronger than expected by chance assuming  $H_0$ , *and* that cumulative probability is itself greater than that seen in random SSLS comparisons.

We filtered loci with corrected unexpectedness  $U_c = 0$  from our principal gene lists (Figures 1 & 2; Supplementary Tables S2 & S3); from our randomisations of hearing, vision and curated genes; from Metacore analysis; and from functional enrichment analyses.

### (ix) Fitting of selection models

**Sitewise selection pressure:** The ratio  $dN/dS$  or  $\omega$  denotes the ratio of the rate of nucleotide substitutions that lead to a codon replacement (nonsynonymous substitutions, dN) to the rate of nucleotide substitutions that do not change the coding sequence (synonymous substitutions, dS). Under the neutral model, dN and dS are expected to occur at approximately the same rate ( $\omega \approx 1$ ), while sites constrained by purifying (negative) selection are expected to evolve with  $\omega < 1$ , and those under diversifying (positive) selection with  $\omega > 1$ . For each locus, we estimated  $\omega$  across the phylogeny, and also in the clade of taxa hypothesised to be convergent, for every site in the dataset using ML in our modified build of version 4.4b of the codeml program within PAML<sup>45</sup>.

The codon models M7 (the null model, with F3×4 frequencies, beta-distributed  $\omega$  and 10 site categories) and M8 (sitewise selection, also beta-distributed  $\omega$  with an additional category representing positive selection<sup>49</sup>) were fitted in our pipeline implementation of PAML 4.4b. The hypothesised phylogenies ( $H_0$ ,  $H_1$ ,  $H_2$ ) were fixed as the user tree and the models compared by likelihood ratio test (LRT). The individual LRTs' P-values were then corrected *post-hoc* in the complete dataset for multiple tests using the method of Benjamini and Hochberg<sup>50</sup>. We then considered the sitewise  $\omega$  estimate only from loci where the M8 model was favoured. In  $H_1$  and  $H_2$  comparisons, 2,235 and 2,234 loci (from the total set of 2,326) passed the LRT, respectively. The sitewise  $\omega$  estimates (and their mean for each locus) derived from this method were compared to clade-specific  $\omega$  estimates (see below, values available on request) and incorporated into the principal component analysis for site data (see below).

**Clade-specific  $\omega$  estimation and derived sitewise  $\omega$ :** Large  $\omega$  values might reflect true phylogenetic signal due to evolutionary divergence, or could alternatively arise from sequencing and/or alignment errors (such processing errors are addressed below).

Where  $\omega$  values represent phylogenetic signal this could be due to neutral drift or diversifying (adaptive) selection, in which case  $\omega$  should be proportional to sitewise  $\omega$ . To test this we fitted the clade-specific Clade Model C (and null model M1a)<sup>51,52</sup> and estimated  $\omega$  on the  $H_1$  and  $H_2$  topologies for the hypothesised clade of echolocating taxa in each case. In this model, three separate  $\omega$  ratios were estimated in the given convergent clade; 'category 0', denoted  $\omega_0$ , for sites under purifying selection where  $0 < \omega_0 < 1$ ; 'category 1', denoted  $\omega_1$ , for sites evolving neutrally, where  $\omega_1 = 1$ ; and 'category 2', denoted  $\omega_2$ , for sites under diversifying selection, in which  $\omega_2$  free to take values  $> 1$ . In this model, the three  $\omega$  ratios are estimated by ML and the Bayes empirical Bayes (BEB) posterior probabilities that each site falls into category 0, 1 or 2 calculated. For each site, the  $\omega$  estimated in each category is then weighted by the BEB posterior probabilities and summed to give an estimate of sitewise  $\omega$  (also see refs. 16 and 19). In subsequent analyses we treated sites with a BEB posterior for the divergent site category 2 ( $\omega_2$ ) of  $> 0.5$  as being under diversifying selection. As with the M8/M7 comparison above, we tested clade model C over the null (M1a) model by LRT, and the resulting complete set of P-values were treated with a *post-hoc* correction following Benjamini and Hochberg<sup>50</sup>. We considered the sitewise clade-specific  $\omega$  estimate only from loci where Model C was favoured.

We also examined the tree lengths for each CDS alignment under  $H_0$ ,  $H_1$  and  $H_2$ ; all were within the recommended limits for codon-based analyses, indicating that no large long branches had biased the ML fitting / optimisation.

**Comparison of ancestral sequences with convergent taxa:** Theoretically, signatures of sequence convergence might also arise under conditions of stronger purifying selection in echolocating taxa than in non-echolocating outgroups (e.g. Old World fruit bats). However,  $\omega$  in the fruit bats was not estimated directly, but instead was averaged over all non-echolocating taxa in the phylogeny. As a result, similar  $\omega$  estimates in both the background clade and the foreground (convergent) clade could arise from unrelated divergence elsewhere in the phylogeny, instead of in the non-echolocating taxa. In such cases, the likelihood support for the monophyly of lineages of echolocating taxa could arise due to the presence of derived amino acid substitutions in the non-convergent outgroup with retention of shared ancestral states in taxa comprising the putative clade of echolocators; misinterpreted as convergence.

Therefore we reconstructed the amino acid sequences at the internal nodes of the  $H_1$  and  $H_2$  phylogenies for each locus using unweighted parsimony<sup>53</sup>. We then examined every position in the alignment, comparing the sequence at the most recent common ancestor (MRCA) of echolocating taxa with the sequences of extant echolocating taxa sampled at the tips themselves. Where two or more echolocating taxa shared a divergent substitution from the MRCA at the same position, we termed this a ‘parallel’ substitution. Comparing the counts of parallel substitutions with the SSLS evidence for convergence at each locus, we found extremely good correspondence between the two measures; although 551 loci lacked any parallel substitutions for  $H_1$  (441 under  $H_2$ ) overall, only six of our 5% most convergent loci ( $n=118$ ) lacked parallel changes for  $H_1$  (seven in the most corresponding  $H_2$  comparison).

**Spatial distribution of selected sites:** Since it is known that estimated numbers of nonsynonymous substitutions, and thus dN/dS ratios ( $\omega$ ), can be inflated by potential sequencing, annotation, and alignment errors<sup>54</sup>, we performed analyses to assess the reliability of our selection results. Sequencing errors and misalignments (i.e. continuous blocks of poorly-aligned sites) will most commonly be characterised by spatially aggregated distributions of divergent amino acids (and hence correlations in the  $\omega$  estimates of adjacent sites). We therefore diagnosed the spatial distribution of sites in each locus estimated to have undergone diversifying selection to look for this signal. Within each locus that had passed the LRT M8/M7 (sitewise selection), we identified those sites with a Bayes empirical Bayes posterior  $> 0.5$  (likely to belong to the divergent site class). We then calculated the intragenic distance as the interval in alignment position between each codon under diversifying selection. For this calculation, sequences were treated as circular such that the distribution of intragenic distances would be identical regardless of the location on the gene of any putative group of misaligned codons.

Based on this definition of intragenic distances, we developed three measurements of the aggregation of codons of interest along a locus. They were: the minimum distance between two codons; the range (max-min distance); the  $k$  distance (defined as the largest integer  $k$  such that no more than  $k+1$  intragenic distances were of length  $k$ ); and the ‘exponent-coefficient.’ The exponent-coefficient assumes distances are exponentially distributed; the ranked distances are log-transformed and a linear regression fitted; this coefficient is the ‘exponent-coefficient’. We fitted linear, generalized linear and generalized linear mixed models to see if these measures of intragenic aggregation correlated (with or without number of taxa or total CDS length) with SSLS. Results of these analyses (available on request) showed that no CDSs were significant.

**Principal component analysis:** To explore broader patterns of association among signatures of convergence, selection and putative gene function, for both CDS (locus) and sitewise data, we undertook principal component analyses (PCA). Data were transformed to



approximate a Standard Normal Distribution. The PCA was repeated for  $H_1$  and for  $H_2$  with the mean and sitewise  $\omega$ SSLs (in CDS and sitewise analyses, respectively) measured with respect to  $H_1$  and  $H_2$ . In the CDS dataset we analysed the principal component weightings of mean  $\omega$ SSLs, number of taxa present, total amino acid count, count of amino acids with significant support (see ‘Simulation of expected sitewise  $\omega$ SSLs distribution’, above), mean  $\omega$ (dN/dS) for sites in the divergent site class in the convergent clade (see ‘Estimation of  $\omega$ (dN/dS ratio)’, above) and loglinear exponent coefficient (a measure of the spatial aggregation of positively-selected sites, see ‘Spatial distribution of selected sites’, above). In the dataset of all amino acids, we analysed the principal component weightings of site-specific likelihood support ( $\omega$ SSLs), number of taxa present, estimated  $\omega$ (dN/dS; see ‘Estimation of  $\omega$ (dN/dS ratio)’, above) and estimated  $\omega$  in the convergent clade (see ‘Estimation of  $\omega$ (dN/dS ratio)’, above). For the locus and sitewise analyses, variance was explained approximately equally across all components; example component loadings for the loci-based PCA of  $H_1$  and  $H_2$  are shown in Supplementary Figure S7(a) and (b) respectively. Example component loadings for the PCA of sitewise data in  $H_1$  and  $H_2$  are shown in Supplementary Figure S8(a) and (b), respectively.

**Locus-wise adaptation/convergence regressions:** High-magnitude  $\omega$ SSLs sites were positively correlated with estimated sitewise  $\omega$  in the clade of echolocating mammals in both  $H_1$  and  $H_2$ . A positive correlation was previously found between sitewise  $\omega$  and  $\omega$ SSLs for the *Prestin* gene<sup>19</sup>, i.e. sites under stronger positive selection also showed larger  $\omega$ SSLs for the alternative (convergent) topology. To test whether site-wise support was driven by selection under each of the given convergent hypotheses, we fitted the following generalized linear mixed model, treating locus as a random effect:

$$|\Delta\omega\text{SSLs}| = \alpha + (\beta \times \omega_2) \quad (\text{Equation S4})$$

Where  $|\Delta\omega\text{SSLs}|$  is the absolute value of the site-specific log-likelihood support;  $\alpha$  the line constant;  $\beta$  the regression coefficient;  $\omega_2$  is the clade-specific estimated sitewise  $\omega$ (dN/dS rate ratio) for the divergent site class (site category 2) in the hypothetical clade of echolocating taxa (i.e. the foreground clade of Clade Model C). For details, see ‘Clade-specific  $\omega$  estimation and derived sitewise  $\omega$ ’, above. Due to computational constraints arising from the size of the dataset, we fitted the model above to replicate datasets directly subsampled without replacement from the original sitewise dataset. For this, we jackknifed 1,000 replicates, each containing 4,000 sites. Mean jackknife estimates of the model parameters are given in Table S5.

**Modelling convergence:** Several loci displayed significant regressions with a negative correlation between  $\omega$  and  $\omega$ SSLs, and also showed support for convergence based on a negative mean  $\omega$ SSLs, but were not categorised as convergent in Figure 2 because their  $\omega$ SSLs did not exceed the threshold we established above. i.e., the value of the empirical cumulative density function (eCDF) of mean locus-wise  $\omega$ SSLs at the 5% significance level (thresholds were  $-0.01035$  and  $-0.205$  for  $H_1$  and  $H_2$ , respectively). However, given their regression suggested a convergent evolutionary trajectory, we predicted the impact of continued selection on these loci by modelling the convergent signal following continued diversifying selection by estimating the  $\omega$ SSLs value at  $\omega = 2$  using the fitted linear relationship in each locus. We predicted two trajectories for each locus: an upper estimate of  $\omega$ SSLs incorporating the uncertainty in the regression parameter estimates (95% confidence intervals of the slope and the intercept); and a central estimate predicted from the mean slope and intercept. Loci were modelled as ‘convergent’ if the upper  $\omega$ SSLs estimate passed the  $\omega$ SSLs threshold and ‘possibly convergent’ if the upper estimate failed the  $\omega$ SSLs threshold, but the central estimate passed it.

‘Convergent’ (Table S6 and S7 for H<sub>1</sub> and H<sub>2</sub>, respectively) or ‘possibly convergent’ predictions (top 50 loci are shown in Tables S8 and S9 for H<sub>1</sub> and H<sub>2</sub>, respectively) were more common in the datasets of CDSs with *a priori* indicators for convergence (hearing, vision and published loci) than in randomly-chosen loci. Predicted convergence signals were significantly stronger for vision loci compared to the background set in H<sub>1</sub> (one-tailed  $T = -2.11$ ;  $P = 0.019$  with ~159 d.f.) and significantly stronger for hearing loci compared to the background set in H<sub>2</sub> (one-tailed  $T = -2.23$ ;  $P = 0.016$  with ~41.7 d.f.).

**Randomisation analysis of convergence signal in hearing and vision genes:** We analysed the strength of the signal for convergence, measured by mean  $\chi^2$  and by the number of sites per locus with significant site-wise  $\chi^2$ , under the H<sub>1</sub> (bat-bat) and H<sub>2</sub> (bat-dolphin) hypotheses among our manually-curated subsets (see above) of loci that were associated with hearing ( $n = 23$ ) or vision ( $n = 75$ ), or previously reported (published) as convergent in the literature ( $n = 7$ ). We measured the mean  $\chi^2$  and number of significantly convergent sites in each of these three subsets. To assess the significance of each result, we performed a randomisation separately for each subset of loci as follows: 1,000 replicate datasets were simulated by reshuffling the observed values without replacement, and the expected mean  $\chi^2$  or number of significantly convergent sites for the subset of loci recomputed. The set of 10,000 expected mean values were taken together to form a null distribution; the position of the observed value in the empirical cumulative density function (eCDF) of the expected values is given as the significance,  $P$ . See Tables S10 (H<sub>1</sub>) and S11 (H<sub>2</sub>).

**Functional enrichment analysis:** Having screened these genomes for convergence, we compiled lists of the most convergent (strongest  $\chi^2$  plus credible  $U_c$ , and alignment heterogeneity / signal: noise scores) loci in our dataset. These genes represent prime candidates for molecular convergence, but direct experimental validation of their function has only been conducted in a handful of cases (*e.g. Prestin; Tmc-1; Pcdh15*, etc). Because generating tissue-specific expression data for the relevant organs is fraught with difficulties (tropical bat cochleae being tiny, highly mineralised, and their storage in the field presents significant logistical challenges) we elected to analyse these target lists for functional enrichment *in silico* as an initial step towards inferring their function.

Functional enrichment analysis for all gene lists under investigation (top 5% of loci by signals for convergence in H<sub>1</sub> and H<sub>2</sub>) was carried out using Fisher’s exact test and Benjamini-Hochberg<sup>50</sup> correction. All enrichments were computed with respect to the background defined by the 2,204 loci under study, so as to avoid biases due to possible functional enrichments of the latter gene list. We thus investigated enrichment of our lists for Gene Ontology annotation terms<sup>55</sup>, KEGG pathways<sup>56</sup>, common interactors according to the HPRD database<sup>57</sup>, microRNA targets according to Targetscan<sup>58</sup> and genes associated to human diseases according to the Genetic Association Database<sup>59</sup>.

At a Benjamini-Hochberg<sup>50</sup> FDR of 30%, and retaining only functional categories represented by at least 5 genes in our H<sub>1</sub> and H<sub>2</sub> lists, only Gene Ontology annotations and microRNA targets showed significant enrichment. However when a control gene list generated ranked convergence signals from random phylogenies was used, approximately equivalent levels of enrichment were seen. The functional enrichment we found and present here should therefore be considered putative, and further work carried out to validate the functional roles of the loci with strongest evidence for convergence.

Overall, one significant enrichment was found for the top 5% most convergent H<sub>1</sub> genes, pertaining to alcohol metabolic processes. For H<sub>2</sub> lists, on the other hand, significant enrichments were found for several Gene Ontology categories and microRNA target sets. The comprehensive results obtained for both the functional enrichment analyses are collated

in Supplementary Table S12 (117 loci representing the top 5% of loci by signal for H<sub>1</sub> and H<sub>2</sub>).

Several loci with the most strongest signals for H<sub>2</sub> convergence were found to be associated to the GO annotation “sensitivity to light stimuli” (GO:0009416). Other categories enriched in the most convergent genes are related to vesicle-mediated transport, regulation of cell cycle and cell death.

We also compared our lists to sets of genes highly expressed (defined as RPKM > 10) in human tissues according to the RNA-Seq Atlas, and found the genes from the most convergent H<sub>2</sub> list were enriched in genes highly expressed in the hypothalamus (P=0.029, Fisher’s exact test; Supplementary Table S13).

We further analysed the most convergent genes from the H<sub>1</sub> and H<sub>2</sub> hypotheses, as well as those from random trees with weakest species tree support, using the Thompson Reuters Metacore database of interactions documented in the literature. For each list we extracted all networks in which the query convergence genes were directly connected (were adjacent vertices.) A large network of 17 genes was discovered matching genes from the H<sub>2</sub> list (see Supplementary Figure S9), including *TP63* (*p63*), *CDK1* (*p34*) and several other genes. *CDK1* (*p34*) has been reported to be closely involved both regeneration and development of hair cells in the cochlea, a key cell cycle process required to replenish hair cell populations for efficient hearing<sup>30</sup>. Similarly, recent work by Terrinoni *et al.* (2013)<sup>29</sup> suggests that *p63* is involved in cochlear development. Networks extracted from H<sub>1</sub> and random tree query lists were much smaller, with a maximum of six loci present.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

We are grateful to K. Baker, L. Davalos, D. Hayman, E. Koilmani, Y. Liu, A. Peel, R. Ransome and A. Rodriguez for providing material for sequencing. We thank C. Walker (Queen Mary GridPP High Throughput Cluster) and A. Terry and C. Mein (Barts and the London Genome Centre) for providing access to computing facilities, and for assistance with running analyses. We are also grateful to S. Dodsworth, R. Buggs, K. Davies, J. Kirkpatrick, R. Nichols, Y. Wurm and S. Young, for comments on the manuscript. This work was funded by Biotechnology and Biological Sciences Research grant BB/H017178/1 awarded to S.J.R., E.S. and J.A.C. (UK).

## References

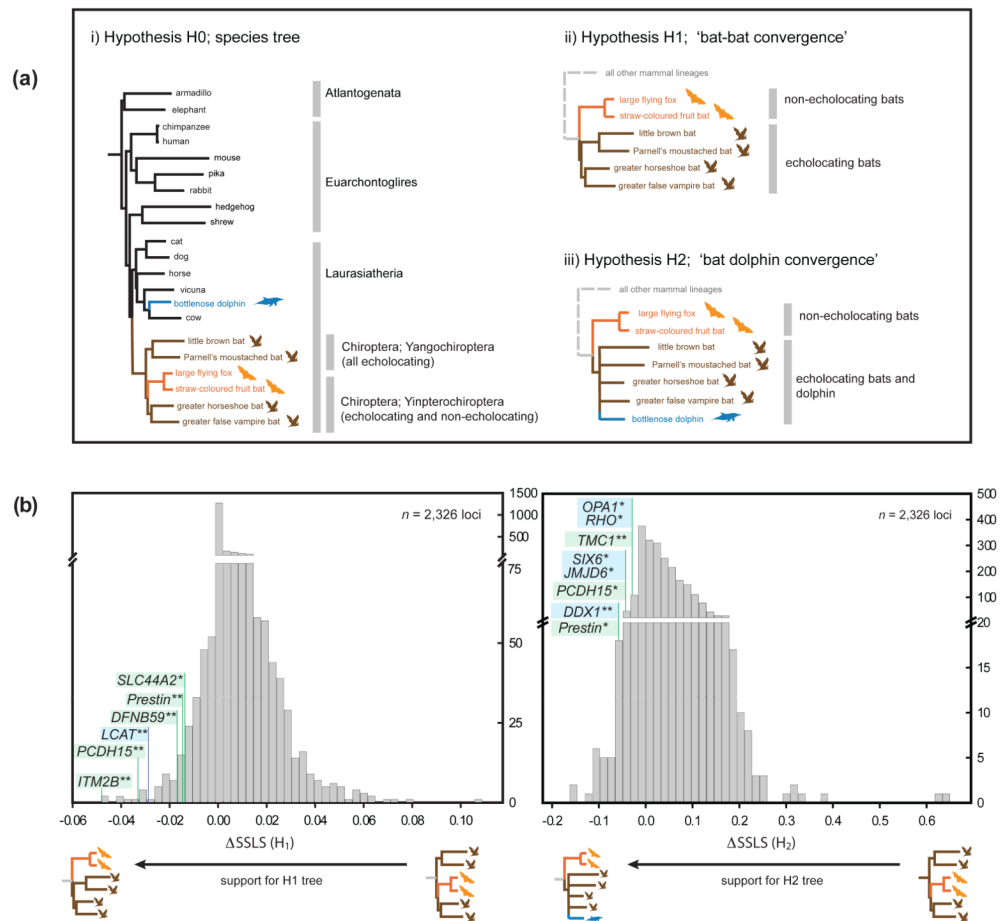
1. Soskine M, Tawfik DS. Mutational effects and the evolution of new protein functions. *Nat. Rev. Genet.* 2010; 11:572–582. [PubMed: 20634811]
2. Clark AG, et al. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature.* 2007; 450:203–218. [PubMed: 17994087]
3. Hughes JF, et al. Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature.* 2010; 463:536–539. [PubMed: 20072128]
4. Hoy RR. Evolution. Convergent evolution of hearing. *Science (New York, N.Y.).* 2012; 338:894–895.
5. Grant PR, Grant BR, Markert JA, Keller LF, Petren K. Convergent evolution of Darwin’s finches caused by introgressive hybridization and selection. *Evolution.* 2004; 58:1588–1599. [PubMed: 15341160]
6. Zhang JZ, Kumar S. Detection of convergent and parallel evolution at the amino acid sequence level. *Mol. Biol. Evol.* 1997; 14:527–536. [PubMed: 9159930]

7. Kriener K, O'Huigin C, Tichy H, Klein J. Convergent evolution of major histocompatibility complex molecules in humans and New World monkeys. *Immunogenetics*. 2000; 51:169–178. [PubMed: 10752625]
8. Li G, et al. The hearing gene *Prestin* reunites echolocating bats. *P. Natl. Acad. Sci. USA*. 2008; 105:13959–13964.
9. Castoe TA, et al. Evidence for an ancient adaptive episode of convergent molecular evolution. *P. Natl. Acad. Sci. USA*. 2009; 106:8986–8991.
10. Liu Y, Rossiter SJ, Han X, Cotton JA, Zhang S. Cetaceans on a molecular fast track to ultrasonic hearing. *Curr. Biol*. 2010; 20:1834–1839. [PubMed: 20933423]
11. Vater, M.; Kössl, M. Echolocation in bats and dolphins. Thomas, JT.; Moss, CF.; Vater, M., editors. University of Chicago Press; Chicago: 2004. p. 89-98.
12. Au WWL, Simmons JA. Echolocation in dolphins and bats. *Phys. Today*. 2007; 60:40–45.
13. Teeling EC, et al. Microbat paraphyly and the convergent evolution of a key innovation in Old World rhinolophoid microbats. *P. Natl. Acad. Sci. USA*. 2002; 99:1431–1436.
14. Teeling EC, et al. Molecular evidence regarding the origin of echolocation and flight in bats. *Nature*. 2000; 403:188–192. [PubMed: 10646602]
15. Jones G, Holderied MW. Bat echolocation calls: adaptation and convergent evolution. *P. Roy. Soc. Lond. B Bio*. 2007; 274:905–912.
16. Davies KTJ, Cotton JA, Kirwan JD, Teeling EC, Rossiter SJ. Parallel signatures of sequence evolution among hearing genes in echolocating mammals: an emerging model of genetic convergence. *Heredity*. 2012; 108:480–489. [PubMed: 22167055]
17. Liu Y, et al. The voltage-gated Potassium channel subfamily KQT Member 4 (*KCNQ4*) displays parallel evolution in echolocating bats. *Mol. Biol. Evol*. 2012; 29:1441–1450. [PubMed: 22319145]
18. Shen Y-Y, Liang L, Li G-S, Murphy RW, Zhang Y-P. Parallel evolution of auditory genes for echolocation in bats and toothed whales. *PLoS Genet*. 2012; 8
19. Liu Y, et al. Convergent sequence evolution between echolocating bats and dolphins. *Curr. Biol*. 2010; 20:R53–R54. [PubMed: 20129036]
20. Zhang G, et al. Comparative analysis of bat genomes provides insight into the evolution of flight and immunity. *Science*. 2013; 339:456–460. [PubMed: 23258410]
21. Sun Y-B, et al. Genome-wide scans for candidate genes involved in the aquatic adaptation of dolphins. *Genome Biology and Evolution*. 2013; 5:130–139. [PubMed: 23246795]
22. Jones G, Teeling EC. The evolution of echolocation in bats. *Trends Ecol. Evol*. 2006; 21:149–156. [PubMed: 16701491]
23. Murphy WJ, Pringle TH, Crider TA, Springer MS, Miller W. Using genomic data to unravel the root of the placental mammal phylogeny. *Genome Res*. 2007; 17:413–421. [PubMed: 17322288]
24. Lindblad-Toh K, et al. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*. 2011; 478:476–482. [PubMed: 21993624]
25. Zhou X, et al. Phylogenomic analysis resolves the interordinal relationships and rapid diversification of the laurasiatherian mammals. *Syst Biol*. 2012; 61:150–164. [PubMed: 21900649]
26. Zhao HB, et al. The evolution of color vision in nocturnal mammals. *P. Natl. Acad. Sci. USA*. 2009; 106:8980–8985.
27. Zhao HB, et al. Rhodopsin molecular evolution in mammals inhabiting low light environments. *PLoS One*. 2009; 4:e8326. [PubMed: 20016835]
28. Fasick JI, Robinson PR. Spectral-tuning mechanisms of marine mammal rhodopsins and correlations with foraging depth. *Visual Neurosci*. 2000; 17:781–788.
29. Terrinoni A, et al. Role of p63 and the Notch pathway in cochlea development and sensorineural deafness. *PNAS*. 2013 Epub ahead of print April 15, 2013; retrieved 22 June, 2013. doi:10.1073/pnas.1214498110.
30. Ryan AF. The cell cycle and the development and regeneration of hair cells. *Curr. Top. Dev. Biol*. 2003; 57:449–466. [PubMed: 14674489]

31. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* 2009; 29:1969–1973. [PubMed: 22367748]
32. Li R, et al. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics.* 2009; 25:1966–1967. [PubMed: 19497933]
33. She R, et al. genBlastG: using BLAST searches to build homologous gene models. *Bioinformatics.* 2011; 27:2141–2143. [PubMed: 21653517]
34. Kim EB, et al. Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nature.* 2011; 479:223–227. [PubMed: 21993625]
35. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics.* 2007; 23:1061–1067. [PubMed: 17332020]
36. Parra G, Bradnam K, Ning Z, Keane T, Korf I. Assessing the gene space in draft genomes. *Nucleic Acids Research.* 2009; 37:298–297. [PubMed: 19042972]
37. Bininda-Emonds OR. transAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. *BMC Bioinformatics.* 2005; 6:156. [PubMed: 15969769]
38. Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* 2005; 33:511–518. [PubMed: 15661851]
39. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 2000; 17:540–552. [PubMed: 10742046]
40. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 2009; 4:44–57. [PubMed: 19131956]
41. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics.* 2004; 5:113. [PubMed: 15318951]
42. Stamatakis A. RAXML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics.* 2006; 22:2688–2690. [PubMed: 16928733]
43. Stamatakis, A. Proceedings of 20th IEEE/ACM International Parallel and Distributed Processing Symposium (IPDPS2006); High Performance Computational Biology Workshop, Rhodos, Greece. 2006;
44. Hancock AM, et al. Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *PNAS.* 2010; 107:8924–8930. [PubMed: 20445095]
45. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 2007; 24:1586–1591. [PubMed: 17483113]
46. Junier T, Zdobnov EM. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics.* 2010; 26:1669–1670. [PubMed: 20472542]
47. Ranwez V, et al. OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. *BMC Evol. Biol.* 2007; 7:241. [PubMed: 18053139]
48. Lartillot N, Lepage T, Blanquart S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics.* 2009; 25(17):2286–8. [PubMed: 19535536]
49. Yang Z, Nielsen R, Goldman N, Pedersen AM. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics.* 2000; 155:431–449. [PubMed: 10790415]
50. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J. Roy. Stat. Soc. B Met.* 1995; 57:289–300.
51. Bielawski JP, Yang ZH. A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. *J. Mol. Evol.* 2004; 59:121–132. [PubMed: 15383915]
52. Wong WS, Yang Z, Goldman N, Nielsen R. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics.* 2004; 168:1041–1051. [PubMed: 15514074]
53. Fitch WM, Margoliash E. Construction of phylogenetic trees. *Science.* 1967; 155:279–284. [PubMed: 5334057]
54. Schneider A, et al. Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome Biol. Evol.* 2009; 1:114–118. [PubMed: 20333182]
55. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nat. Genet.* 2000; 25:25–9. [PubMed: 10802651]



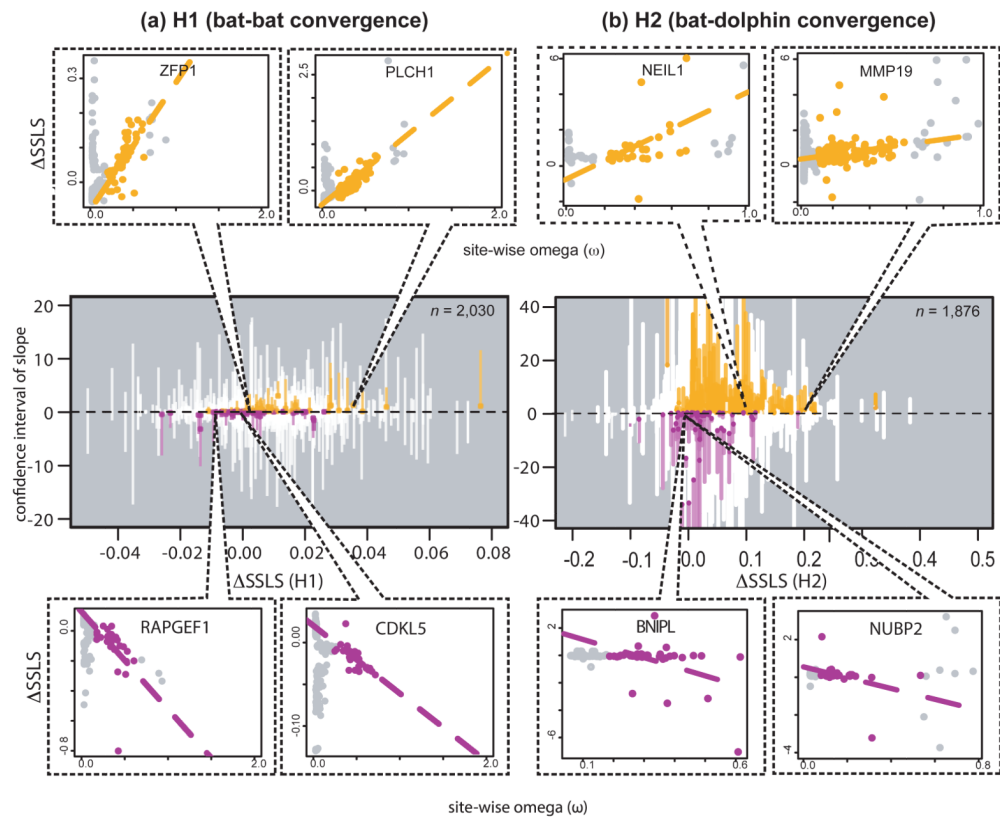
56. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular datasets. *Nucleic Acids Res.* 2012; 40:D109–D114. [PubMed: 22080510]
57. Prasad TSK, et al. Human Protein Reference Database - 2009 Update. *Nucleic Acids Research.* 2008; 37:D767–72. [PubMed: 18988627]
58. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell.* 2005; 120:15–20. [PubMed: 15652477]
59. Becker KG, Barnes KC, Bright TJ, Wang SA. The Genetic Association Database. *Nature Genetics.* 2004; 36:431–432. [PubMed: 15118671]



**Figure 1. Convergence hypotheses, and genomic distribution of support**

**a)** For each locus, the goodness-of-fit of three separate phylogenetic hypotheses was considered: (i)  $H_0$ , the accepted species phylogeny based on recent findings e.g.<sup>14,23-25</sup>; (ii)  $H_1$ , or 'bat-bat convergence', in which echolocating bat lineages (shown in brown) are forced to form a monophyletic group to the exclusion of non-echolocating Old World fruit bats (shown in orange); and (iii)  $H_2$ , or 'bat-dolphin convergence', in which the echolocating bat lineages and the dolphin (blue) form a monophyletic group to the exclusion of all non-echolocating mammals. See Supplementary Information for details of model fitting and topologies.

**b)** The distribution of convergence signal across 2,326 loci in 14–22 representative mammalian taxa, as measured by locus-wise mean site-specific likelihood support for the species topology ( $H_0$ ) over (i) the 'bat-bat' hypothesis uniting echolocating bats (i.e. SSLS ( $H_1$ )) and (ii) bat-dolphin hypothesis (i.e. SSLS ( $H_2$ )). Representative hearing and vision loci are shown in green and blue, respectively; for each locus significance levels based on simulation denote whether it had significant counts of convergent sites after correcting for expected counts in random (control) phylogenies (\*) and whether strength of positive selection (dN/dS) and convergence (SSLS) at sites under selection in echolocators were correlated (\*\*); see Table S4, Supplementary Information.



**Figure 2. Relationship between strength of convergence signal and adaptive selection**  
 For hypotheses (a)  $H_1$  and (b)  $H_2$  ( $n = 2,030$  and  $1,876$  loci, respectively), the 95% confidence intervals of the coefficient (slope) for locus-wise regressions between site-wise support for convergence and site-wise  $\omega$  for sites under diversifying selection are plotted. In each plot, loci showing a negative relationship, as characterised by a slope significantly below zero, are consistent with an evolutionary trajectory of adaptive convergence (purple line; filled circle indicating upper 95% limit) and loci showing a positive relationship, with a slope of greater than zero, are consistent with an evolutionary trajectory adaptive divergence (orange line with filled circle indicating lower 95% limit). Insets show two examples of adaptive convergence and divergence under each hypothesis. Full details of estimation and regression fitting are given in the Supplementary Information.