doi:10.1038/nature12511

Table S1: Species for which genome data were included in this study.

Common name	Superorder	Order	Family	Genus	Species
Elephant	Afrotheria	Proboscidea	Elephantidae	*Loxodonta*	*africana*
Rabbit	Euarchontoglires	Lagomorpha	Leporidae	*Oryctolagus*	*cuniculus*
American pika	Euarchontoglires	Lagomorpha	Ochotonidae	*Ochotona*	*princeps*
Human	Euarchontoglires	Primates	Hominidae	*Homo*	*sapiens*
Common chimpanzee	Euarchontoglires	Primates	Hominidae	*Pan*	*troglodytes*
Mouse	Euarchontoglires	Rodentia	Muridae	*Mus*	*musculus*
Cow	Laurasiatheria	Cetartiodactyla	Bovidae	*Bos*	*taurus*
Common bottlenose dolphin	Laurasiatheria	Cetartiodactyla	Delphinidae	*Tursiops*	*truncatus*
Alpaca	Laurasiatheria	Cetartiodactyla	Camelidae	*Vicugna*	*pacos*
Dog	Laurasiatheria	Carnivora	Canidae	*Canis*	*familiaris*
Cat	Laurasiatheria	Carnivora	Felidae	*Felis*	*catus*
Parnell's moustached bat	Laurasiatheria	Chiroptera (Yangochiroptera)	Mormoopidae	*Pteronotus*	*parnellii**
Little brown bat	Laurasiatheria	Chiroptera (Yangochiroptera)	Vespertilionidae	*Myotis*	*lucifugus*
Greater false vampire bat	Laurasiatheria	Chiroptera (Yinpterochiroptera)	Megadermatidae	*Megaderma*	*lyra**
Greater horseshoe bat	Laurasiatheria	Chiroptera: (Yinpterochiroptera)	Rhinolophidae	*Rhinolophus*	*ferrumequinum**
Straw-coloured fruit bat	Laurasiatheria	Chiroptera: (Yinpterochiroptera)	Pteropodidae	*Eidolon*	*helvum**
Large flying fox	Laurasiatheria	Chiroptera: (Yinpterochiroptera)	Pteropodidae	*Pteropus*	*vampyrus*
European hedgehog	Laurasiatheria	Erinaceomorpha	Erinaceidae	*Erinaceus*	*europaeus*
Horse	Laurasiatheria	Perissodactyla	Equidae	*Equus*	*caballus*
Common shrew	Laurasiatheria	Soricomorpha	Soricidae	*Sorex*	*araneus*
Short-tailed opossum	Marsupalia	Didelphimorphia	Didelphidae	*Monodelphis*	*domestica*
Nine-banded armadillo	Xenarthra	Cingulata	Dasypodinae	*Dasypus*	*novemcinctus*

Species sequenced *de novo* in this study are marked *. Echolocating taxa are coloured red.

Table S2: Top 5% of loci by ΔSSLS for H_1

Locus	Shortcode	# taxa	# amino acids	Mean $\Delta SSLS_{H0-H1}$ [a]	# strict sites [b]	$U_{corrected}$ [c]
ENSG00000136877	*FPGS*	22	133	-0.04973	17	0.402
ENSG00000136156	*ITM2B*	22	128	-0.04900	5	0.187
ENSG00000105771	*SMG9*	19	85	-0.04275	24	0.219
ENSG00000178922	*HYI*	21	108	-0.04190	30	0.210
ENSG00000100281	*HMGXB4*	22	107	-0.03762	0	0.329
ENSG00000162878	*PKDCC*	19	99	-0.03522	6	0.133
ENSG00000115365	*LANCL1*	19	111	-0.03355	6	0.413
ENSG00000150275	*PCDH15*[d]	20	784	-0.03162	13	0.118
ENSG00000175175	*PPM1E*	22	107	-0.03119	4	0.236
ENSG00000172613	*RAD9A*	18	133	-0.03100	0	0.100
ENSG00000167107	*ACSF2*	19	174	-0.02886	10	0.169
ENSG00000178385	*PLEKHM3*	21	163	-0.02758	2	0.277
ENSG00000100147	*CCDC134*	21	120	-0.02702	17	0.316
ENSG00000014919	*COX15*	22	166	-0.02655	7	0.323
ENSG00000127564	*PKMYT1*	20	118	-0.02492	1	0.171
ENSG00000213398	*LCAT*	20	191	-0.02365	10	0.421
ENSG00000132911	*NMUR2*	21	230	-0.02333	4	0.284
ENSG00000117477	*C1orf114*	21	236	-0.02323	2	0.145
ENSG00000106692	*FKTN*	22	138	-0.02241	11	0.197
ENSG00000165887	*ANKRD2*	21	101	-0.02200	5	0.225
ENSG00000196419	*XRCC6*	17	99	-0.02195	3	0.149
ENSG00000006453	*BAIAP2L1*	21	203	-0.02149	1	0.373
ENSG00000182010	*RTKN2*	21	297	-0.02071	9	0.172
ENSG00000180269	*GPR139*	19	154	-0.02038	11	0.050
ENSG00000071246	*VASH1*	20	73	-0.01960	10	0.209
ENSG00000054118	*THRAP3*	19	145	-0.01948	2	0.183
ENSG00000182134	*TDRKH*	21	215	-0.01946	0	0.242
ENSG00000164175	*SLC45A2*	21	245	-0.01890	2	0.098
ENSG00000146409	*SLC18B1*	19	138	-0.01843	2	0.203
ENSG00000205517	*RGL3*	19	194	-0.01814	9	0.253
ENSG00000154328	*NEIL2*	19	163	-0.01809	8	0.194
ENSG00000136271	*DDX56*	21	140	-0.01784	4	0.231
ENSG00000148444	*COMMD3*	21	152	-0.01745	0	0.314
ENSG00000204052	*LRRC73*	20	69	-0.01732	3	0.048
ENSG00000204311	*DFNB59*[d]	21	194	-0.01720	40	0.098
ENSG00000185133	*INPP5J*	19	120	-0.01717	0	0.009
ENSG00000138400	*MDH1B*	22	340	-0.01712	4	0.155
ENSG00000166881	*TMEM194A*	21	134	-0.01712	45	0.383
ENSG00000164089	*AGXT2L1*	21	134	-0.01670	9	0.468
ENSG00000040487	*PQLC2*	21	172	-0.01652	3	0.277
ENSG00000221838	*AP4M1*	21	219	-0.01586	7	0.122
ENSG00000138135	*CH25H*	16	115	-0.01533	6	0.279
ENSG00000124602	*UNC5CL*	20	127	-0.01531	12	0.219
ENSG00000170615	*SLC26A5*[d]	20	400	-0.01505	1	0.295
ENSG00000049883	*PTCD2*	22	241	-0.01491	12	0.150
ENSG00000143970	*ASXL2*	22	378	-0.01485	7	0.263
ENSG00000164749	*HNF4G*	18	188	-0.01480	5	0.130
ENSG00000128059	*PPAT*	21	183	-0.01456	11	0.238
ENSG00000137776	*SLTM*	18	152	-0.01456	7	0.441
ENSG00000005981	*ASB4*	20	198	-0.01447	68	0.316

Table S2 continued: Top 5% of loci by ΔSSLS for H_1

Locus	Shortcode	# taxa	# amino acids	Mean $\Delta SSLS_{H0-H1}$ [a]	# strict sites [b]	$U_{corrected}$ [c]
ENSG00000102575	*ACP5*	18	181	-0.01442	5	0.168
ENSG00000092295	*TGM1*	22	228	-0.01433	8	0.222
ENSG00000176170	*SPHK1*	20	205	-0.01416	0	0.272
ENSG00000156050	*FAM161B*	22	287	-0.01413	5	0.193
ENSG00000141434	*MEP1B*	22	300	-0.01393	16	0.267
ENSG00000152766	*ANKRD22*	20	125	-0.01380	0	0.211
ENSG00000163694	*RBM47*	19	105	-0.01380	10	0.161
ENSG00000174080	*CTSF*	22	122	-0.01374	4	0.217
ENSG00000129353	*SLC44A2*	21	130	-0.01355	4	0.192
ENSG00000137955	*RABGGTB*	21	175	-0.01349	1	0.300
ENSG00000158220	*ESYT3*	21	315	-0.01323	6	0.239
ENSG00000124155	*PIGT*	22	227	-0.01311	6	0.371
ENSG00000107581	*EIF3A*	22	177	-0.01280	3	0.152
ENSG00000165383	*LRRC18*	21	123	-0.01268	12	0.298
ENSG00000121316	*PLBD1*	22	185	-0.01210	3	0.331
ENSG00000089597	*GANAB*	19	418	-0.01198	3	0.140
ENSG00000144909	*OSBPL11*	22	193	-0.01189	7	0.256
ENSG00000118495	*PLAGL1*	21	214	-0.01188	4	0.453
ENSG00000006576	*PHTF2*	21	258	-0.01179	2	0.153
ENSG00000145029	*NICN1*	22	164	-0.01151	7	0.105
ENSG00000143369	*ECM1*	19	241	-0.01132	7	0.073
ENSG00000095383	*TBC1D2*	20	223	-0.01119	7	0.318
ENSG00000167720	*SRR*	19	218	-0.01115	19	0.255
ENSG00000101981	*F9*	21	130	-0.01094	4	0.164
ENSG00000108465	*CDK5RAP3*	21	217	-0.01084	6	0.175
ENSG00000107736	*CDH23*	19	1452	-0.01062	27	0.098
ENSG00000198001	*IRAK4*	21	143	-0.01061	21	0.295
ENSG00000149090	*PAMR1*	22	221	-0.01061	14	0.200
ENSG00000132259	*CNGA4*	19	176	-0.01045	35	0.103
ENSG00000112218	*GPR63*	22	142	-0.01025	2	0.188
ENSG00000175595	*ERCC4*	21	140	-0.01008	4	0.016
ENSG00000172269	*DPAGT1*	21	180	-0.00990	0	0.225
ENSG00000186889	*TMEM17*	20	120	-0.00972	3	0.352
ENSG00000197296	*FITM2*	21	164	-0.00951	24	0.299
ENSG00000131067	*GGT7*	22	213	-0.00948	6	0.413
ENSG00000133961	*NUMB*	22	95	-0.00947	4	0.015
ENSG00000118017	*A4GNT*	21	157	-0.00937	7	0.162
ENSG00000114200	*BCHE*	21	292	-0.00932	3	0.305
ENSG00000186009	*ATP4B*	19	145	-0.00924	6	0.301
ENSG00000213901	*SLC23A3*	19	236	-0.00919	0	0.253
ENSG00000064703	*DDX20*	22	135	-0.00912	0	0.188
ENSG00000111845	*PAK1IP1*	20	162	-0.00910	31	0.315
ENSG00000095739	*BAMBI*	21	103	-0.00876	3	0.252
ENSG00000005238	*FAM214B*	20	274	-0.00871	62	0.221
ENSG00000086189	*DIMT1*	22	197	-0.00871	12	0.256
ENSG00000116793	*PHTF1*	21	288	-0.00849	18	0.249
ENSG00000115239	*ASB3*	22	247	-0.00833	5	0.348
ENSG00000168389	*MFSD2A*	22	257	-0.00830	8	0.512
ENSG00000170385	*SLC30A1*	20	147	-0.00825	17	0.218
ENSG00000151092	*NGLY1*	21	231	-0.00822	1	0.342

Table S2 continued: Top 5% of loci by ΔSSLS for H_1

Locus	Shortcode	# taxa	# amino acids	Mean $\Delta SSLS_{H0-H1}$ [a]	# strict sites [b]	$U_{corrected}$ [c]
ENSG00000162227	*TAF6L*	19	189	-0.00799	13	0.163
ENSG00000185019	*UBOX5*	18	336	-0.00786	16	0.421
ENSG00000123992	*DNPEP*	21	201	-0.00778	2	0.296
ENSG00000162702	*ZNF281*	18	191	-0.00749	4	0.183
ENSG00000129351	*ILF3*	21	219	-0.00746	11	0.116
ENSG00000122861	*PLAU*	21	276	-0.00745	11	0.160
ENSG00000106397	*PLOD3*	19	395	-0.00744	4	0.310
ENSG00000100099	*HPS4*	20	149	-0.00727	2	0.309
ENSG00000116406	*EDEM3*	21	243	-0.00709	8	0.097
ENSG00000130766	*SESN2*	22	227	-0.00701	5	0.093
ENSG00000143337	*TOR1AIP1*	22	164	-0.00687	1	0.361
ENSG00000104237	*RP1*	21	428	-0.00677	4	0.187
ENSG00000172594	*SMPDL3A*	21	201	-0.00674	12	0.215
ENSG00000134061	*CD180*	20	203	-0.00672	14	0.237
ENSG00000111879	*FAM184A*	22	274	-0.00668	13	0.181
ENSG00000108349	*CASC3*	22	227	-0.00667	67	0.307
ENSG00000112992	*NNT*	22	322	-0.00663	5	0.103

[a]Mean sitewise ΔSSLS; negative values imply support for the alternative hypothesis; see text for details.

[b]The count of sites with parallel amino acid substitutions present in echolocating taxa, away from the most recent common ancestor of all echolocators.

[c]The unexpectedness, *U*, of the genomic sequences data, compared to a null distribution drawn from simulated data. Corrected by a sample of 100 random phylogenies; positive values after correction indicate good signal for the alternative hypothesis; see text for details.

[d]These loci were manually curated (downloaded and aligned by MUSCLE then inspected) but analysed alongside genomic data.

Table S3: Top 5% of loci by ΔSSLS for H_2

Locus	Shortcode	# taxa	# amino acids	Mean $\Delta SSLS_{H0-H2}$[a]	# strict sites[b]	$U_{corrected}$[c]
ENSG00000120802	*TMPO*	18	444	-0.55960	17	0.069
ENSG00000054938	*CHRDL2*	21	273	-0.15506	7	0.224
ENSG00000137942	*FNBP1L*	22	237	-0.14072	2	0.145
ENSG00000176422	*SPRYD4*	21	105	-0.11159	7	0.232
ENSG00000129170	*CSRP3*	21	79	-0.10328	11	0.043
ENSG00000128595	*CALU*	21	94	-0.09968	6	0.446
ENSG00000232119	*MCTS1*	19	96	-0.09851	6	0.068
ENSG00000103429	*BFAR*	21	133	-0.09699	17	0.166
ENSG00000188175	*HEPACAM2*	20	166	-0.09296	11	0.210
ENSG00000163141	*BNIPL*	21	131	-0.08386	8	0.129
ENSG00000174080	*CTSF*	22	122	-0.07847	4	0.062
ENSG00000155367	*PPM1J*	NA	273	-0.07704	5	0.020
ENSG00000105401	*CDC37*	18	201	-0.07459	4	0.044
ENSG00000163959	*AC069257.9*	22	178	-0.07432	27	0.169
ENSG00000173653	*RCE1*	21	116	-0.06622	34	0.075
ENSG00000171135	*JAGN1*	20	89	-0.06430	9	0.123
ENSG00000106692	*FKTN*	22	138	-0.06077	11	0.133
ENSG00000079462	*PAFAH1B3*	19	86	-0.05752	2	0.090
ENSG00000115020	*PIKFYVE*	22	108	-0.05626	7	0.019
ENSG00000160396	*HIPK4*	22	155	-0.05488	3	0.038
ENSG00000121104	*FAM117A*	20	144	-0.05410	5	0.123
ENSG00000170615	*SLC26A5*[d]	20	400	-0.05322	1	0.261
ENSG00000136152	*COG3*	22	114	-0.05267	10	0.163
ENSG00000198794	*SCAMP5*	20	99	-0.05238	1	0.009
ENSG00000079785	*DDX1*	22	35	-0.05111	4	0.005
ENSG00000020922	*MRE11A*	22	200	-0.05004	13	0.071
ENSG00000108344	*PSMD3*	22	162	-0.04904	2	0.021
ENSG00000070495	*JMJD6*	20	80	-0.04769	2	0.081
ENSG00000100147	*CCDC134*	21	120	-0.04442	17	0.236
ENSG00000115459	*ELMOD3*	21	138	-0.04341	8	0.124
ENSG00000143756	*FBXO28*	19	81	-0.04319	9	0.075
ENSG00000105771	*SMG9*	19	85	-0.04275	24	0.216
ENSG00000221838	*AP4M1*	21	219	-0.04267	7	0.082
ENSG00000033170	*FUT8*	18	55	-0.04115	1	0.074
ENSG00000150275	*PCDH15*[d]	20	784	-0.04114	13	0.113
ENSG00000184302	*SIX6*	21	81	-0.04102	11	0.042
ENSG00000138795	*LEF1*	21	112	-0.04014	4	0.077
ENSG00000180269	*GPR139*	19	154	-0.04001	11	0.039
ENSG00000125991	*ERGIC3*	22	161	-0.03971	8	0.058
ENSG00000186767	*SPIN4*	18	41	-0.03966	11	0.128
ENSG00000084774	*CAD*	19	82	-0.03924	19	0.070
ENSG00000133961	*NUMB*	22	95	-0.03826	4	0.050
ENSG00000163161	*ERCC3*	22	72	-0.03635	4	0.163
ENSG00000025796	*SEC63*	21	153	-0.03619	27	0.050
ENSG00000179476	*C14orf28*	21	46	-0.03575	6	0.054
ENSG00000184845	*DRD1*	19	197	-0.03572	18	0.166
ENSG00000136156	*ITM2B*	22	128	-0.03492	5	0.181
ENSG00000119318	*RAD23B*	21	78	-0.03486	3	0.007
ENSG00000168060	*NAALADL1*	20	170	-0.03450	4	0.158

Table S3 continued: Top 5% of loci by ΔSSLS for H_2

Locus	Shortcode	# taxa	# amino acids	Mean $\Delta SSLS_{H0-H2}$[a]	# strict sites[b]	$U_{corrected}$[c]
ENSG00000172071	*EIF2AK3*	22	102	-0.03418	6	0.190
ENSG00000117751	*PPP1R8*	22	94	-0.03414	0	0.031
ENSG00000100281	*HMGXB4*	22	107	-0.03373	0	0.044
ENSG00000144468	*RHBDD1*	22	138	-0.03356	18	0.068
ENSG00000070214	*SLC44A1*	22	250	-0.03342	13	0.029
ENSG00000083782	*EPYC*	22	82	-0.03229	4	0.082
ENSG00000100220	*C22orf28*	22	145	-0.03226	1	0.054
ENSG00000163914	*RHO*	21	185	-0.03217	3	0.121
ENSG00000111845	*PAK1IP1*	20	162	-0.03137	31	0.179
ENSG00000115365	*LANCL1*	19	111	-0.03135	6	0.085
ENSG00000157500	*APPL1*	22	159	-0.03110	4	0.003
ENSG00000165091	*TMC1*[d]	19	336	-0.02990	6	0.279
ENSG00000159588	*CCDC17*	21	164	-0.02979	1	0.130
ENSG00000108018	*SORCS1*	22	173	-0.02965	9	0.186
ENSG00000129151	*BBOX1*	22	117	-0.02861	4	0.217
ENSG00000167098	*SUN5*	19	156	-0.02787	7	0.086
ENSG00000147526	*TACC1*	20	160	-0.02694	1	0.055
ENSG00000136875	*PRPF4*	19	104	-0.02680	6	0.060
ENSG00000244462	*RBM12*	17	79	-0.02666	4	0.153
ENSG00000073282	*TP63*	22	94	-0.02642	0	0.398
ENSG00000129159	*KCNC1*	18	55	-0.02632	6	0.435
ENSG00000198218	*QRICH1*	22	87	-0.02623	8	0.047
ENSG00000159086	*GCFC1*	22	74	-0.02608	14	0.084
ENSG00000159202	*UBE2Z*	22	124	-0.02604	0	0.059
ENSG00000125772	*GPCPD1*	22	152	-0.02575	3	0.103
ENSG00000143498	*TAF1A*	22	226	-0.02573	15	0.161
ENSG00000196584	*XRCC2*	18	108	-0.02540	8	0.124
ENSG00000157613	*CREB3L1*	22	133	-0.02517	8	0.119
ENSG00000116191	*RALGPS2*	22	112	-0.02506	6	0.090
ENSG00000108848	*LUC7L3*	22	88	-0.02468	4	0.138
ENSG00000171219	*CDC42BPG*	18	235	-0.02444	1	0.028
ENSG00000170312	*CDK1*	19	129	-0.02434	5	0.192
ENSG00000198836	*OPA1*	22	192	-0.02408	10	0.067
ENSG00000135472	*FAIM2*	20	95	-0.02339	6	0.087
ENSG00000111554	*MDM1*	21	149	-0.02320	3	0.129
ENSG00000198042	*MAK16*	22	135	-0.02210	9	0.108
ENSG00000099341	*PSMD8*	18	100	-0.02170	27	0.190
ENSG00000100473	*COCH*	22	228	-0.02165	1	0.046
ENSG00000113231	*PDE8B*	22	209	-0.02158	19	0.138
ENSG00000101361	*NOP56*	22	96	-0.02154	0	0.088
ENSG00000132612	*VPS4A*	20	133	-0.02082	11	0.077
ENSG00000149182	*ARFGAP2*	19	222	-0.02011	14	0.022
ENSG00000128285	*MCHR1*	21	96	-0.01992	1	0.137
ENSG00000130713	*EXOSC2*	17	109	-0.01989	2	0.105
ENSG00000168291	*PDHB*	19	59	-0.01984	14	0.071
ENSG00000188039	*NWD1*	20	377	-0.01960	3	0.161
ENSG00000089693	*MLF2*	19	106	-0.01957	5	0.151
ENSG00000112242	*E2F3*	22	105	-0.01888	2	0.054
ENSG00000116237	*ICMT*	21	104	-0.01878	5	0.075

Table S3 continued: Top 5% of loci by ΔSSLS for H_2

Locus	Shortcode	# taxa	# amino acids	Mean $\Delta SSLS_{H0-H2}$[a]	# strict sites[b]	$U_{corrected}$[c]
ENSG00000163879	*DNALI1*	20	124	-0.01870	24	0.114
ENSG00000006704	*GTF2IRD1*	19	127	-0.01869	1	0.057
ENSG00000134222	*PSRC1*	22	165	-0.01859	6	0.206
ENSG00000152520	*PAN3*	22	156	-0.01807	10	0.079
ENSG00000074695	*LMAN1*	22	120	-0.01787	2	0.196
ENSG00000123610	*TNFAIP6*	19	114	-0.01780	2	0.095
ENSG00000107957	*SH3PXD2A*	20	126	-0.01779	7	0.120
ENSG00000188037	*CLCN1*	20	177	-0.01766	19	0.074
ENSG00000173039	*RELA*	19	186	-0.01739	7	0.101
ENSG00000198917	*C9orf114*	18	107	-0.01720	5	0.024
ENSG00000092931	*MFSD11*	22	153	-0.01713	3	0.166
ENSG00000169783	*LINGO1*	19	127	-0.01675	0	0.122
ENSG00000163322	*FAM175A*	19	141	-0.01657	4	0.250
ENSG00000091947	*TMEM101*	20	94	-0.01629	3	0.084
ENSG00000110066	*SUV420H1*	22	157	-0.01608	6	0.041
ENSG00000028528	*SNX1*	22	227	-0.01554	14	0.138
ENSG00000072415	*MPP5*	22	129	-0.01544	4	0.063
ENSG00000162510	*MATN1*	18	122	-0.01531	0	0.085
ENSG00000020129	*NCDN*	21	111	-0.01529	11	0.070

[a]Mean sitewise ΔSSLS; negative values imply support for the alternative hypothesis; see text for details.

[b]The count of sites with parallel amino acid substitutions present in echolocating taxa, away from the most recent common ancestor of all echolocators

[c]The unexpectedness, *U*, of the genomic sequences data, compared to a null distribution drawn from simulated data. Corrected by a sample of 100 random phylogenies; positive values after correction indicate good signal for the alternative hypothesis; see text for details.

[d]These loci were manually curated (downloaded and aligned by MUSCLE then inspected) but analysed alongside genomic data.

Table S4: Value of 0.5% ΔSSLS thresholds in empirical cumulative density functions.

Mean 0.5% cut-off by:	**H_1**	**H_2**
Empirical eCDF	-0.182	-0.126

Table S5: Relationship between convergence and selection. For each convergence hypothesis, the overall relationship between sitewise ω and sitewise convergence (ΔSSLS) is shown. In each hypothesis, sitewise ω values are estimated from Clade Model C (ref. 51), based on the subset of sites under diversifying selection in the foreground clade of echolocating taxa. Significance values are calculated using a general linear model (non-nested) and a general linear mixed model (nested) that takes account of locus identity.

	H_1		H_2	
	$\omega_{category=2}$ (not nested)	$\omega_{category=2}$ (nested by locus)	$\omega_{category=2}$ (not nested)	$\omega_{category=2}$ (nested by locus)
Significance	$P \leq 3.18 \times 10^{-16}$	$P \leq 0.0336$	$P \leq 10^{-16}$	$P \leq 10^{-4}$
95% CI (*lower, upper)*	n/a	0.023, 0.044	n/a	0.110, 0.176

Table S6: Loci predicted as convergent† under H_1

Locus	Shortcode	# taxa	# amino acids	Predicted $\Delta SSLS_{H0-H2}$ (upper 95% CI)
ENSG00000005844	*ITGAL*	22	199	-0.069895564
ENSG00000008086	*CDKL5*	22	351	-0.065927974
ENSG00000064225	*ST3GAL6*	21	217	-0.067540928
ENSG00000071553	*ATP6AP1*	17	326	-0.01678209
ENSG00000088836	*SLC4A11*	18	509	-0.412249256
ENSG00000100802	*C14orf93*	19	451	-0.089600522
ENSG00000121900	*TMEM54*	19	162	-0.88170854
ENSG00000129317	*PUS7L*	21	301	-0.092390644
ENSG00000129353	*SLC44A2*	21	238	-4.342244867
ENSG00000141642	*ELAC1*	22	350	-0.408438993
ENSG00000151445	*C14orf133*	22	320	-0.213546448
ENSG00000154856	*APCDD1*	20	358	-0.971202308
ENSG00000155846	*PPARGC1B*	20	227	-0.084438169
ENSG00000157856	*CCDC164*	21	162	-0.17684286
ENSG00000162066	*AMDHD2*	20	194	-0.049834835
ENSG00000163728	*TTC14*	22	522	-0.267828727
ENSG00000176170	*SPHK1*	20	281	-0.362903444
ENSG00000178971	*CTC1*	20	415	-0.07270797

†For definition of convergent in this table, see section 'Modelling convergence' in Supplementary Methods.

Table S7: Loci predicted as convergent† under H_2

Locus	Shortcode	# taxa	# amino acids	Predicted $\Delta SSLS_{H0-H2}$ (upper 95% CI)
ENSG00000168811	*IL12A*	20	178	-128.4342228
ENSG00000159579	*RSPRY1*	22	270	-53.27990152
ENSG00000133316	*WDR74*	20	192	-51.23810921
ENSG00000128342	*LIF*	14	178	-32.35141302
ENSG00000132386	*SERPINF1*	21	198	-27.83596811
ENSG00000064607	*SUGP2*	20	622	-24.32574224
ENSG00000111669	*TPI1*	19	211	-11.06309015
ENSG00000153292	*GPR110*	22	427	-10.17167636
ENSG00000149634	*SPATA25*	20	186	-6.876799493
ENSG00000011260	*UTP18*	18	236	-6.132208011
ENSG00000114204	*SERPINI2*	22	266	-6.005360888
ENSG00000244462	*RBM12*	17	356	-4.105273184
ENSG00000119321	*FKBP15*	22	458	-3.912682229
ENSG00000164089	*AGXT2L1*	21	229	-2.738789984
ENSG00000152223	*EPG5*	19	359	-2.459409088
ENSG00000108423	*TUBD1*	22	239	-1.610020216
ENSG00000116062	*MSH6*	21	821	-1.478653277
ENSG00000144460	*NYAP2*	21	259	-1.423851034
ENSG00000112365	*ZBTB24*	21	418	-1.169114405
ENSG00000184564	*SLITRK6*	19	813	-1.05091772
ENSG00000128203	*ASPHD2*	19	215	-1.003315718
ENSG00000114107	*CEP70*	19	217	-0.417963744
ENSG00000105088	*OLFM2*	19	428	-0.381619082
ENSG00000089693	*MLF2*	19	216	-0.233443653
ENSG00000112033	*PPARD*	21	246	-0.017125978
ENSG00000113812	*ACTR8*	22	567	-0.010582255
ENSG00000187003	*ACTL7A*	19	373	-0.01014298
ENSG00000181856	*SLC2A4*	21	407	-0.006228354
ENSG00000109689	*STIM2*	22	349	-0.002609717

†For definition of convergent in this table, see section 'Modelling convergence' in Supplementary Methods.

Table S8: Top 50 loci predicted as 'possibly convergent'† under H_1

Locus	Shortcode	# taxa	# amino acids	Predicted $\Delta SSLS_{H0-H1}$ (upper 95% CI)
ENSG00000071246	*VASH1*	20	163	-229.561042
ENSG00000139567	*ACVRL1*	22	243	-15.7774201
ENSG00000105700	*KXD1*	18	154	-12.25772409
ENSG00000129353	*SLC44A2*	21	238	-11.92962605
ENSG00000131043	*C20orf4*	21	247	-11.80586063
ENSG00000214655	*KIAA0913*	20	658	-10.05081025
ENSG00000112118	*MCM3*	20	241	-9.799580222
ENSG00000001084	*GCLC*	21	387	-8.686815553
ENSG00000187726	*DNAJB13*	20	232	-7.980367887
ENSG00000130119	*GNL3L*	22	158	-7.314986577
ENSG00000116574	*RHOU*	18	149	-6.349564962
ENSG00000168538	*TRAPPC11*	19	381	-5.910413772
ENSG00000163661	*PTX3*	20	201	-5.316684388
ENSG00000154856	*APCDD1*	20	358	-5.168069912
ENSG00000109771	*LRP2BP*	22	250	-5.076442554
ENSG00000168876	*ANKRD49*	18	149	-5.041708728
ENSG00000071462	*WBSCR22*	18	199	-5.029199729
ENSG00000181830	*SLC35C1*	19	151	-4.916720929
ENSG00000118017	*A4GNT*	21	191	-4.903220192
ENSG00000121075	*TBX4*	21	263	-4.592456015
ENSG00000185278	*ZBTB37*	22	338	-4.23977918
ENSG00000176170	*SPHK1*	20	281	-4.079174033
ENSG00000149474	*CSRP2BP*	22	362	-3.736369935
ENSG00000173809	*TDRD12*	22	185	-3.707831464
ENSG00000162878	*PKDCC*	19	197	-3.594987378
ENSG00000100281	*HMGXB4*	22	161	-3.582726539
ENSG00000178764	*ZHX2*	19	641	-3.534863631
ENSG00000125772	*GPCPD1*	22	253	-3.317279982
ENSG00000121900	*TMEM54*	19	162	-3.251111689
ENSG00000108753	*HNF1B*	22	229	-3.239539591
ENSG00000188037	*CLCN1*	20	330	-3.117673238
ENSG00000005981	*ASB4*	20	258	-2.784343572
ENSG00000102575	*ACP5*	18	272	-2.742222021
ENSG00000117477	*C1orf114*	21	302	-2.700117017
ENSG00000011021	*CLCN6*	22	418	-2.627052039
ENSG00000112992	*NNT*	22	578	-2.580135746
ENSG00000101294	*HM13*	22	177	-2.530454534
ENSG00000110060	*PUS3*	18	303	-2.510405751
ENSG00000163728	*TTC14*	22	522	-2.472004537
ENSG00000157103	*SLC6A1*	20	448	-2.464005107
ENSG00000132872	*SYT4*	22	304	-2.463848677
ENSG00000180096	*SEPT1*	21	318	-2.251348075
ENSG00000014919	*COX15*	22	224	-2.224139232
ENSG00000040487	*PQLC2*	21	198	-2.198852341
ENSG00000163468	*CCT3*	22	230	-2.195194786
ENSG00000132906	*CASP9*	21	246	-2.191194873
ENSG00000124207	*CSE1L*	22	778	-2.181719017
ENSG00000013293	*SLC7A14*	21	413	-2.080115041
ENSG00000138030	*KHK*	21	233	-2.04697345
ENSG00000125484	*GTF3C4*	20	644	-2.029727711

†For definition of possibly convergent in this table, see section 'Modelling convergence' in Supplementary Methods.

Table S9: Top 50 loci predicted as 'possibly convergent'† under H_2

Locus	Shortcode	# taxa	# amino acids	Predicted $\Delta SSLS_{H0-H2}$ (upper 95% CI)
ENSG00000168811	*IL12A*	20	178	-225.5643999
ENSG00000133316	*WDR74*	20	192	-70.02695809
ENSG00000159579	*RSPRY1*	22	270	-64.75639316
ENSG00000183763	*TRAIP*	21	369	-63.82411505
ENSG00000128203	*ASPHD2*	19	215	-54.26976226
ENSG00000115657	*ABCB6*	20	665	-51.11948399
ENSG00000064607	*SUGP2*	20	622	-49.33173464
ENSG00000167962	*ZNF598*	19	221	-38.97026897
ENSG00000175115	*PACS1*	22	177	-33.65424179
ENSG00000128342	*LIF*	14	178	-32.66808687
ENSG00000132386	*SERPINF1*	21	198	-31.15590264
ENSG00000165805	*C12orf50*	21	241	-30.41497535
ENSG00000213380	*COG8*	21	473	-27.49925394
ENSG00000178691	*SUZ12*	21	324	-26.44763892
ENSG00000164089	*AGXT2L1*	21	229	-25.14091224
ENSG00000137824	*FAM82A2*	22	201	-23.676655
ENSG00000157181	*C1orf27*	19	389	-23.58274587
ENSG00000154920	*EME1*	21	257	-23.05070688
ENSG00000106144	*CASP2*	21	371	-22.46509414
ENSG00000049883	*PTCD2*	22	292	-21.93373628
ENSG00000152223	*EPG5*	19	359	-21.44784839
ENSG00000114204	*SERPINI2*	22	266	-21.22696999
ENSG00000166997	*CNPY4*	22	154	-20.55822135
ENSG00000186660	*ZFP91*	22	263	-20.52301715
ENSG00000081177	*EXD2*	22	311	-18.22607165
ENSG00000116062	*MSH6*	21	821	-18.13291693
ENSG00000137996	*RTCA*	22	234	-17.71329175
ENSG00000101391	*CDK5RAP1*	22	362	-17.68293644
ENSG00000155890	*TRIM42*	22	397	-17.38890268
ENSG00000120802	*TMPO*	18	546	-17.23699001
ENSG00000149634	*SPATA25*	20	186	-17.03364134
ENSG00000010318	*PHF7*	22	162	-16.85869445
ENSG00000158552	*ZFAND2B*	20	170	-16.71651213
ENSG00000168758	*SEMA4C*	21	495	-16.17044116
ENSG00000101282	*RSPO4*	21	155	-15.51121189
ENSG00000169641	*LUZP1*	21	641	-15.31100441
ENSG00000100034	*PPM1F*	19	275	-15.29183602
ENSG00000139908	*TSSK4*	21	312	-15.15284893
ENSG00000076382	*SPAG5*	19	206	-14.73972184
ENSG00000035403	*VCL*	22	228	-14.43451589
ENSG00000011478	*QPCTL*	19	200	-14.4009076
ENSG00000111669	*TPI1*	19	211	-14.4002126
ENSG00000121454	*LHX4*	21	198	-13.59648152
ENSG00000101224	*CDC25B*	19	356	-13.57127406
ENSG00000139269	*INHBE*	18	284	-13.32628638
ENSG00000011260	*UTP18*	18	236	-12.69252283
ENSG00000130985	*UBA1*	22	314	-12.04750572
ENSG00000119321	*FKBP15*	22	458	-11.95304955
ENSG00000153292	*GPR110*	22	427	-11.66759464
ENSG00000151500	*THYN1*	20	156	-11.48987376

†For definition of possibly convergent in this table, see section 'Modelling convergence' in Supplementary Methods.

Table S10: Randomisation analysis of significance of subsets of loci of interest and signal for convergence under H_1. * P < 0.05, ** P < 0.01

Subset of loci:	Sound	Vision	Published
Mean ΔSSLS	0.0006	0.0018	-0.0077
Expected mean ΔSSLS	0.0037	0.0038	0.0040
Significance by randomisation, P	0.09	0.055	0.003**

Table S11: Randomisation analysis of significance of subsets of loci of interest and signal for convergence under H_2. * P < 0.05, ** P < 0.01

Subset of loci:	Sound	Vision	Published
Mean ΔSSLS	0.0284	0.0380	-0.0057
Expected mean ΔSSLS	0.0478	0.0477	0.0472
Significance by randomisation, P	0.041*	0.07	0.006**

Table S12: Functional enrichment analysis of 117 genes showing the strongest signal for convergence between echolocating bats and dolphins (H_2).

List	Category	ID	Name	Hit Count in Query List	Expected hit count	P-value	Q-value	Loci
H1	GO: Biological process	GO:0006066	alcohol metabolic process	5	1.14	0.0045	0.27	*BCHE, CH25H, DPAGT1, LCAT, SPHK1*
H2	GO: Biological process	GO:0016192	vesicle-mediated transport	14	5.51	0.00086	0.27	*AP4M1, ARFGAP2, CALU, COG3, ELMOD3, ERGIC3, FNBP1L, HMGXB4, JMJD6, LMAN1, PIKFYVE, SCAMP5, SNX1, VPS4A*
H2	GO: Biological process	GO:0051641	cellular localization	22	11.28	0.0013	0.27	*AP4M1, APPL1, CALU, CDC37, CDK1, COG3, DRD1, EIF2AK3, FNBP1L, HMGXB4, ICMT, LMAN1, MPP5, OPA1, PDE8B, PIKFYVE, PSRC1, SCAMP5, SEC63, SNX1, TACC1, VPS4A*
H2	GO: Biological process	GO:0000077	DNA damage checkpoint	5	0.94	0.0017	0.27	*CDK1, FAM175A, PSMD3, PSMD8, TP63*
H2	GO: Biological process	GO:0010001	glial cell differentiation	5	0.94	0.0017	0.27	*CDK1, DRD1, LINGO1, MPP5, RELA*
H2	GO: Biological process	GO:0016071	mRNA metabolic process	11	4.05	0.0019	0.27	*DDX1, ERCC3, EXOSC2, JMJD6, LUC7L3, PAN3, PPP1R8, PRPF4, PSMD3, PSMD8, SMG9*
H2	GO: Biological process	GO:0009314	response to radiation	8	2.49	0.0027	0.27	*DRD1, ERCC3, FAM175A, RELA, RHBDD1, RHO, TP63, XRCC2*
H2	GO: Biological process	GO:0031570	DNA integrity checkpoint	5	1.09	0.0036	0.27	*CDK1, FAM175A, PSMD3, PSMD8, TP63*
H2	GO: Biological process	GO:0042063	gliogenesis	5	1.09	0.0036	0.27	*CDK1, DRD1, LINGO1, MPP5, RELA*
H2	GO: Biological process	GO:0009416	response to light stimulus	6	1.56	0.0037	0.27	*DRD1, ERCC3, RELA, RHBDD1, RHO, TP63*
H2	GO: Biological process	GO:0016197	endosomal transport	5	1.2	0.0055	0.27	*HMGXB4, LMAN1, PIKFYVE, SNX1, VPS4A*
H2	GO: Cellular component	GO:0012505	endomembrane system	21	11.9	0.0057	0.27	*AP4M1, APPL1, ARFGAP2, BFAR, COG3, CREB3L1, DRD1, EIF2AK3, ERGIC3, FKTN, FUT8, ICMT, ITM2B, JAGN1, LMAN1, MPP5, PIKFYVE, RCE1, SCAMP5, SEC63, TMPO*

List	Category	ID	Name	Hit Count in Query List	Expected hit count	P-value	Q-value	Loci
H2	GO: Biological process	GO:0006915	apoptotic process	19	10.39	0.0059	0.27	*APPL1, BFAR, BNIPL, CDK1, EIF2AK3, ERCC3, FAIM2, ITM2B, JMJD6, LEF1, LUC7L3, OPA1, PSMD3, PSMD8, RELA, RHBDD1, TP63, UBE2Z, XRCC2*
H2	GO: Biological process	GO:0008150	biological_process	103	93.01	0.0063	0.27	*AP4M1, APPL1, ARFGAP2, BBOX1, BFAR, BNIPL, C22orf28, C9orf114, CAD, CALU, CDC37, CDC42BPG, CDK1, CHRDL2, CLCN1, COCH, COG3, CREB3L1, CSRP3, CTSF, DDX1, DNALI1, DRD1, E2F3, EIF2AK3, ELMOD3, EPYC, ERCC3, ERGIC3, EXOSC2, FAIM2, FAM175A, FKTN, FNBP1L, FUT8, GCFC1, GPCPD1, GPR139, GTF2IRD1, HMGXB4, ICMT, ITM2B, JMJD6, KCNC1, LANCL1, LEF1, LINGO1, LMAN1, LUC7L3, MATN1, MCHR1, MCTS1, MDM1, MFSD11, MLF2, MPP5, MRE11A, NAALADL1, NCDN, NOP56, NUMB, OPA1, OSTalpha, PAFAH1B3, PAK1IP1, PAN3, PDE8B, PDHB, PIKFYVE, PPM1J, PPP1R8, PRPF4, PSMD3, PSMD8, PSRC1, QRICH1, RAD23B, RALGPS2, RCE1, RELA, RHBDD1, RHO, SCAMP5, SEC63, SH3PXD2A, SIX6, SLC44A1, SMG9, SNX1, SORCS1, SPIN4, SPRYD4, SUN5, SUV420H1, TACC1, TAF1A, TMEM101, TMPO, TNFAIP6, TP63, UBE2Z, VPS4A, XRCC2*
H2	GO: Biological process	GO:0051234	establishment of localization	30	19.33	0.0064	0.27	*AP4M1, APPL1, ARFGAP2, CALU, CDC37, CDK1, CLCN1, COG3, DRD1, EIF2AK3, ELMOD3, ERGIC3, FNBP1L, HMGXB4, ICMT, JMJD6, KCNC1, LMAN1, MCHR1, OPA1, OSTalpha, PDE8B, PIKFYVE, PSRC1, RHBDD1, SCAMP5, SEC63, SLC44A1, SNX1, VPS4A*

List	Category	ID	Name	Hit Count in Query List	Expected hit count	P-value	Q-value	Loci
H2	GO: Biological process	GO:0012501	programmed cell death	19	10.5	0.0066	0.27	*APPL1, BFAR, BNIPL, CDK1, EIF2AK3, ERCC3, FAIM2, ITM2B, JMJD6, LEF1, LUC7L3, OPA1, PSMD3, PSMD8, RELA, RHBDD1, TP63, UBE2Z, XRCC2*
H2	GO: Biological process	GO:0006916	anti-apoptosis	6	1.77	0.007	0.27	*BFAR, CDK1, FAIM2, OPA1, RELA, TP63*
H2	GO: Biological process	GO:0000075	cell cycle checkpoint	7	2.34	0.0075	0.27	*CDK1, ERCC3, FAM175A, MRE11A, PSMD3, PSMD8, TP63*
H2	GO: Biological process	GO:0051179	localization	35	23.9	0.0078	0.27	*AP4M1, APPL1, ARFGAP2, CALU, CDC37, CDK1, CLCN1, COG3, DRD1, EIF2AK3, ELMOD3, ERCC3, ERGIC3, FNBP1L, FUT8, HMGXB4, ICMT, JMJD6, KCNC1, LEF1, LMAN1, MCHR1, MPP5, OPA1, OSTalpha, PDE8B, PIKFYVE, PSRC1, RHBDD1, SCAMP5, SEC63, SLC44A1, SNX1, TACC1, VPS4A*
H2	GO: Biological process	GO:0010564	regulation of cell cycle process	9	3.59	0.0082	0.27	*CDK1, ERCC3, FAM175A, LEF1, MRE11A, PSMD3, PSMD8, PSRC1, TP63*
H2	GO: Biological process	GO:0051649	establishment of localization in cell	18	9.98	0.0085	0.27	*AP4M1, CALU, CDC37, CDK1, COG3, DRD1, EIF2AK3, HMGXB4, ICMT, LMAN1, OPA1, PDE8B, PIKFYVE, PSRC1, SCAMP5, SEC63, SNX1, VPS4A*
H2	GO: Cellular component	GO:0044422	organelle part	61	48.43	0.0097	0.27	*AP4M1, APPL1, ARFGAP2, BFAR, CAD, CALU, CDC42BPG, CDK1, COG3, CREB3L1, CSRP3, DDX1, DNALI1, DRD1, E2F3, EIF2AK3, ERCC3, ERGIC3, EXOSC2, FAM175A, FKTN, FUT8, GTF2IRD1, HEPACAM2, HMGXB4, ICMT, ITM2B, JAGN1, JMJD6, LEF1, LMAN1, LUC7L3, MAK16, MRE11A, NOP56, OPA1, PAK1IP1, PAN3, PDHB, PIKFYVE, PPP1R8, PRPF4, PSMD3, PSMD8, PSRC1, RAD23B, RCE1, RELA, RHO, SCAMP5, SEC63, SH3PXD2A, SLC44A1, SNX1, SUV420H1, TACC1, TAF1A, TMPO, TP63, UBE2Z, VPS4A*

List	Category	ID	Name	Hit Count in Query List	Expected hit count	P-value	Q-value	Loci
H2	GO: Biological process	GO:0006810	transport	29	19.07	0.0099	0.27	*AP4M1, APPL1, ARFGAP2, CALU, CDC37, CDK1, CLCN1, COG3, DRD1, EIF2AK3, ELMOD3, ERGIC3, FNBP1L, HMGXB4, ICMT, JMJD6, KCNC1, LMAN1, MCHR1, OPA1, OSTalpha, PDE8B, PIKFYVE, RHBDD1, SCAMP5, SEC63, SLC44A1, SNX1, VPS4A*
H2	GO: Biological process	GO:0008219	cell death	20	11.74	0.01	0.27	*APPL1, BFAR, BNIPL, CDK1, EIF2AK3, ERCC3, FAIM2, ITM2B, JMJD6, LEF1, LUC7L3, NOP56, OPA1, PSMD3, PSMD8, RELA, RHBDD1, TP63, UBE2Z, XRCC2*
H2	GO: Biological process	GO:0016265	death	20	11.74	0.01	0.27	*APPL1, BFAR, BNIPL, CDK1, EIF2AK3, ERCC3, FAIM2, ITM2B, JMJD6, LEF1, LUC7L3, NOP56, OPA1, PSMD3, PSMD8, RELA, RHBDD1, TP63, UBE2Z, XRCC2*
H2	GO: Biological process	GO:0071156	regulation of cell cycle arrest	7	2.49	0.011	0.27	*APPL1, ITM2B, PIKFYVE, SCAMP5, SNX1, VPS4A*
H2	GO: Cellular component	GO:0010008	endosome membrane	6	1.92	0.011	0.27	*AP4M1, APPL1, CDC37, CDK1, COG3, DRD1, ICMT, LMAN1, MPP5, PIKFYVE, SEC63, SNX1*
H2	GO: Biological process	GO:0034613	cellular protein localization	12	5.77	0.011	0.27	*BFAR, CDK1, FAIM2, LEF1, OPA1, RELA, RHBDD1, TP63, XRCC2*
H2	GO: Biological process	GO:0043066	negative regulation of apoptotic process	9	3.74	0.011	0.27	*CDK1, ERCC3, FAM175A, MRE11A, PSMD3, PSMD8, TP63*
H2	GO: Cellular component	GO:0005667	transcription factor complex	6	1.97	0.012	0.27	*E2F3, ERCC3, LEF1, RELA, TAF1A, TP63*
H2	GO: Biological process	GO:0042981	regulation of apoptotic process	15	8.05	0.012	0.27	*APPL1, BFAR, BNIPL, CDK1, ERCC3, FAIM2, ITM2B, LEF1, OPA1, PSMD3, PSMD8, RELA, RHBDD1, TP63, XRCC2*
H2	GO: Biological process	GO:0070727	cellular macromolecule localization	12	5.87	0.012	0.27	*AP4M1, APPL1, CDC37, CDK1, COG3, DRD1, ICMT, LMAN1, MPP5, PIKFYVE, SEC63, SNX1*
H2	GO: Cellular component	GO:0044446	intracellular organelle part	60	47.96	0.013	0.27	*BFAR, CDK1, FAIM2, LEF1, OPA1, RELA, RHBDD1, TP63, XRCC2*

List	Category	ID	Name	Hit Count in Query List	Expected hit count	P-value	Q-value	Loci
H2	GO: Biological process	GO:0043069	negative regulation of programmed cell death	9	3.85	0.013	0.27	*AP4M1, APPL1, ARFGAP2, BFAR, CAD, CALU, CDC42BPG, CDK1, COG3, CREB3L1, DDX1, DNALI1, DRD1, E2F3, EIF2AK3, ERCC3, ERGIC3, EXOSC2, FAM175A, FKTN, FUT8, GTF2IRD1, HEPACAM2, HMGXB4, ICMT, ITM2B, JAGN1, JMJD6, LEF1, LMAN1, LUC7L3, MAK16, MRE11A, NOP56, OPA1, PAK1IP1, PAN3, PDHB, PIKFYVE, PPP1R8, PRPF4, PSMD3, PSMD8, PSRC1, RAD23B, RCE1, RELA, RHO, SCAMP5, SEC63, SH3PXD2A, SLC44A1, SNX1, SUV420H1, TACC1, TAF1A, TMPO, TP63, UBE2Z, VPS4A*
H2	GO: Biological process	GO:0043067	regulation of programmed cell death	15	8.16	0.014	0.28	*APPL1, BFAR, BNIPL, CDK1, ERCC3, FAIM2, ITM2B, LEF1, OPA1, PSMD3, PSMD8, RELA, RHBDD1, TP63, XRCC2*
H2	GO: Cellular component	GO:0044440	endosomal part	6	2.03	0.014	0.28	*APPL1, ITM2B, PIKFYVE, SCAMP5, SNX1, VPS4A*
H2	GO: Biological process	GO:0006974	response to DNA damage stimulus	11	5.3	0.015	0.29	*CDK1, DDX1, ERCC3, FAM175A, MCTS1, MRE11A, PSMD3, PSMD8, RAD23B, TP63, XRCC2*

Table S13: Analysis of differential expression (data: RNA-Seq Atlas) in human tissues of 117 loci showing strongest signals of convergence between echolocating bats and dolphins (H_2).

MicroRNA family	Hit count - most convergent H_2 loci	Expected hit count	P-value	Q-value	Loci
miR-217	5	1.35	0.0095	0.1648	E2F3, LUC7L3, MRE11A, RAD23B, SH3PXD2A
miR-374ab	7	2.60	0.0132	0.1648	CALU, CHRDL2, COG3, PIKFYVE, RHBDD1, SEC63, TACC1
miR-34ac/34bc-5p/449abc/449c-5p	7	2.81	0.0198	0.1648	CREB3L1, E2F3, FAIM2, FUT8, LEF1, LMAN1, RALGPS2
miR-320abcd/4429	7	3.12	0.0334	0.1765	FBXO28, PAN3, RAD23B, SEC63, SORCS1, SUV420H1, UBE2Z
miR-96/507/1271	8	3.90	0.0379	0.1765	APPL1, CDC37, CDC42BPG, COG3, JMJD6, PDE8B, RALGPS2, TACC1
miR-200bc/429/548a	9	4.73	0.0434	0.1765	APPL1, CALU, E2F3, LMAN1, MPP5, NUMB, PAN3, SH3PXD2A, SPRYD4
miR-495/1192	8	4.10	0.0494	0.1765	C14orf28, FKTN, LUC7L3, RBM12, SH3PXD2A, SLC44A1, SORCS1, UBE2Z
let-7/98/4458/4500	8	4.42	0.0705	0.2203	C14orf28, GPCPD1, ICMT, KCNC1, LINGO1, LUC7L3, SPRYD4, TMPO
miR-15abc/16/16abc/195/322/424/497/1907	9	5.35	0.0829	0.2302	ARFGAP2, CALU, DRD1, E2F3, HMGXB4, RAD23B, SNX1, TACC1, VPS4A
miR-539/539-5p	5	2.65	0.1220	0.2831	MATN1, MRE11A, OPA1, RBM12, RCE1
miR-93/93a/105/106a/291a-3p/294/295/302abcde/372/373/428/519a/520be/520acd-3p/1378/1420ac	6	3.43	0.1246	0.2831	DRD1, E2F3, PAN3, RAD23B, RHO, TP63
miR-30abcdef/30abe-5p/384-5p	8	5.20	0.1442	0.2888	C14orf28, CALU, E2F3, GPCPD1, MCTS1, MFSD11, RALGPS2, SNX1
miR-590-3p	8	5.25	0.1502	0.2888	E2F3, GPCPD1, HMGXB4, KCNC1, LANCL1, MPP5, PAK1IP1, RBM12

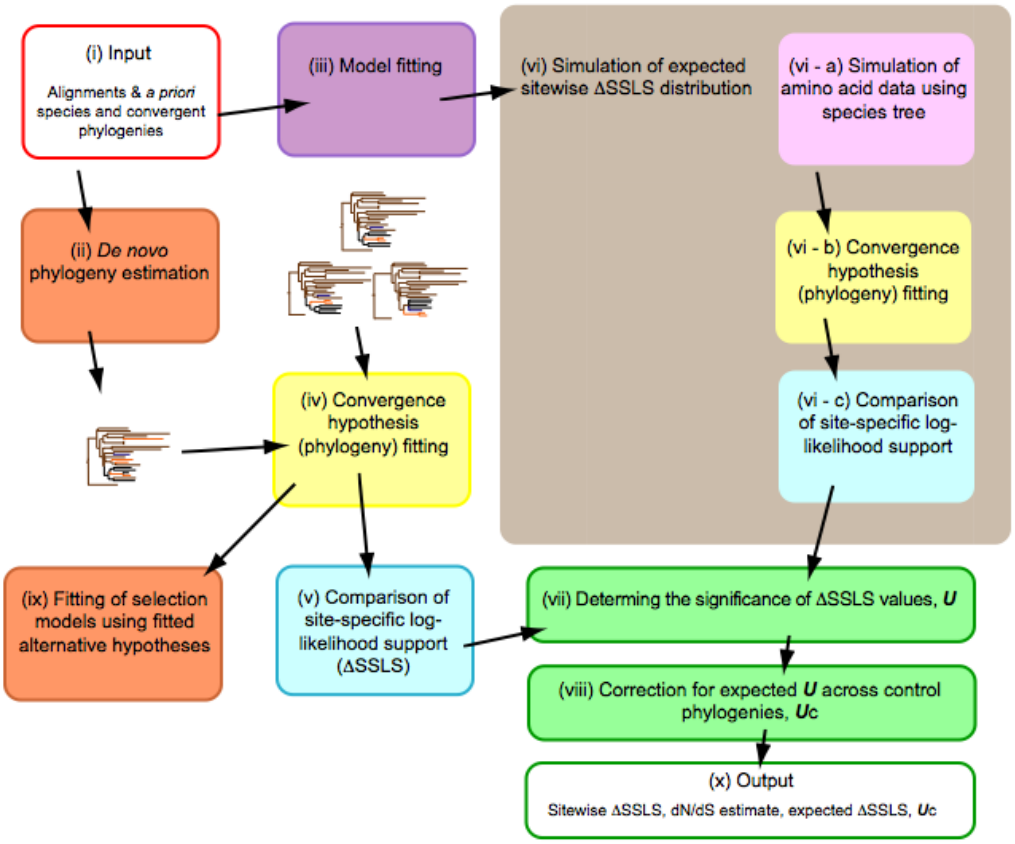

Figure S1: Schematic of high-throughput pipeline for the detection of molecular convergence *in silico*. Further details are given in Supplementary Methods, 'Analysis pipeline' (i) Job initialisation: CDS alignment parsed and checked for stop codons, ambiguity characters and missing taxa; pruning of branches from *a priori* phylogenies where necessary; (ii) Separate estimation of *de novo* phylogeny for *post-hoc* quality control, and separate resolution of soft polytomy in H_2 hypothesis; (iii) Substitution model fitting and branch length estimation under the species phylogeny (H_0); (iv) Fitting of proposed convergent phylogenies (H_1 and H_2); (v) Comparison of sitewise log-likelihood values to calculate ΔSSLS; (vi) The expected distribution of ΔSSLS values along the species phylogeny (H_0) is sampled by simulation: (a) Phylobayes used to sample posterior distribution of GTR CAT substitution model parameters under constrained species tree, H_0, by MCMC. Posterior samples used to simulate multiple replicate alignments under GTR CAT under H_0; (b) The hypotheses H_0, H_1 and H_2 are fitted to the simulated data, as in step (iv); (c) expected ΔSSLS calculated as in step (v); (vii) Calculation of the empirical cumulative density function (eCDF) of expected ΔSSLS values allows the cumulative probability of observed Δ SSLS values, *U*, to be determined; (viii) Cumulative probabilities of observed ΔSSLS values calculated on randomised control phylogenies are used to calculate the normalised CDF measure U_c. (ix) dN/dS estimated in the hypothesised convergent clades (bat-bat and bat-dolphin) using the optimised H_1 and H_2 phylogenies; (x) Output comprised: mean ΔSSLS value for the locus; the estimated mean dN/dS for the locus; sitewise estimates of dN/dS in the convergent clade, dN/dS across the phylogeny, and the observed ΔSSLS values (compared to the eCDF of the expected ΔSSLS values to assess significance via the normalised corrected measure U_c).

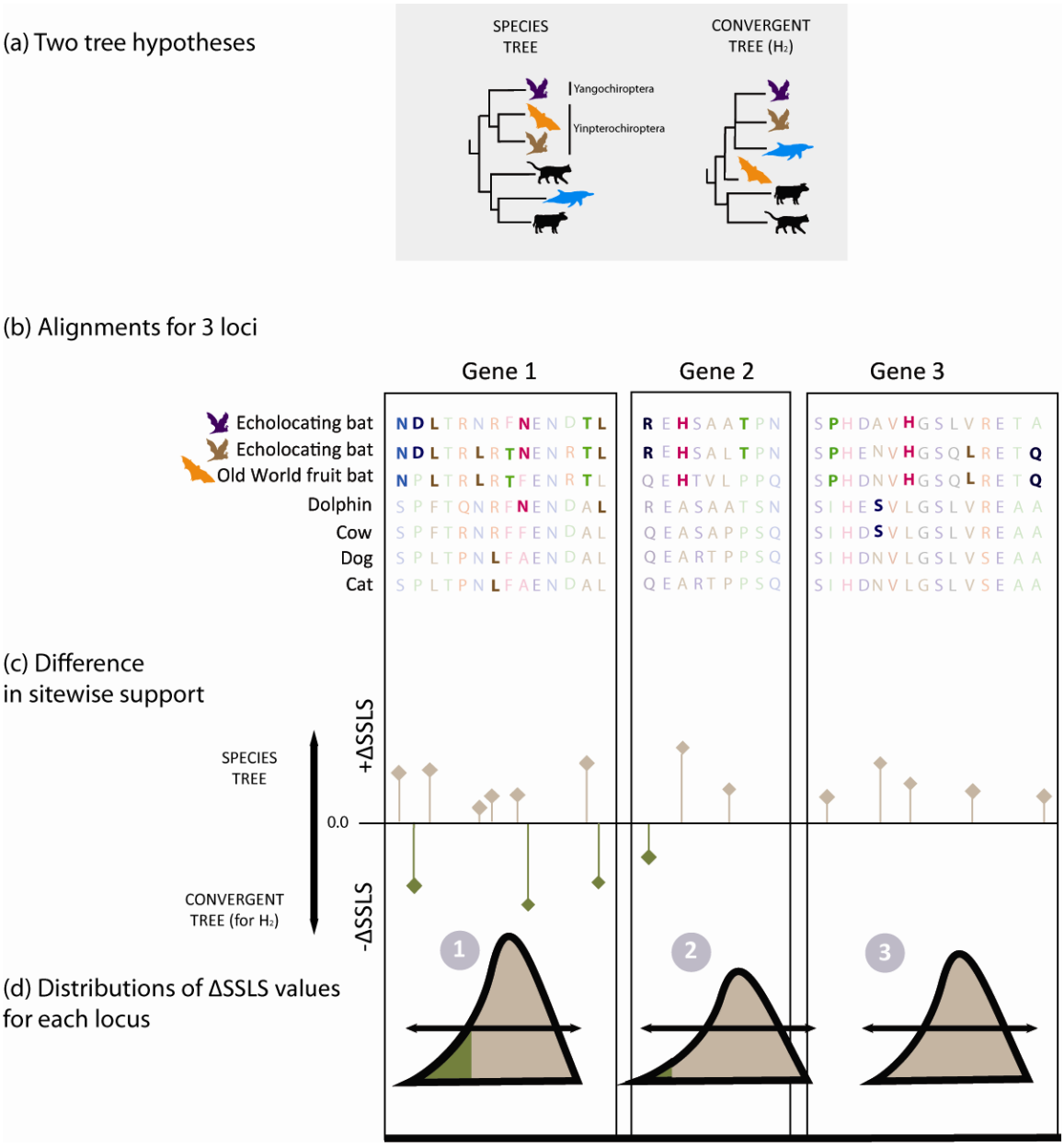

Figure S2: Example of ΔSSLS calculation based on (a) two candidate phylogenies (species and convergent trees H_0 and H_2, in this case) that contain the same taxa but differ in their topology. (b) For these taxa, multiple sequence alignments for several loci contain amino acid substitutions, some of which are consistent with H_0; and others with H_2. (c) For every amino acid the goodness-of-fit of each topology, given the data, is estimated under maximum likelihood: the difference in goodness-of-fit ('ΔSSLS'; see Supplementary Methods) gives the direction of support, where positive ΔSSLS values indicate support for the first topology (species tree H_0), while negative ΔSSLS values indicate support for the convergent phylogeny, H_2. (d) Finally, the distribution of ΔSSLS values for each locus indicates the overall support for convergence; loci with significant support for convergence will have a negative mean ΔSSLS (1); signal in others will be mixed and approximately zero (2); positive mean ΔSSLS indicates a locus supporting the species tree H_2 (3).

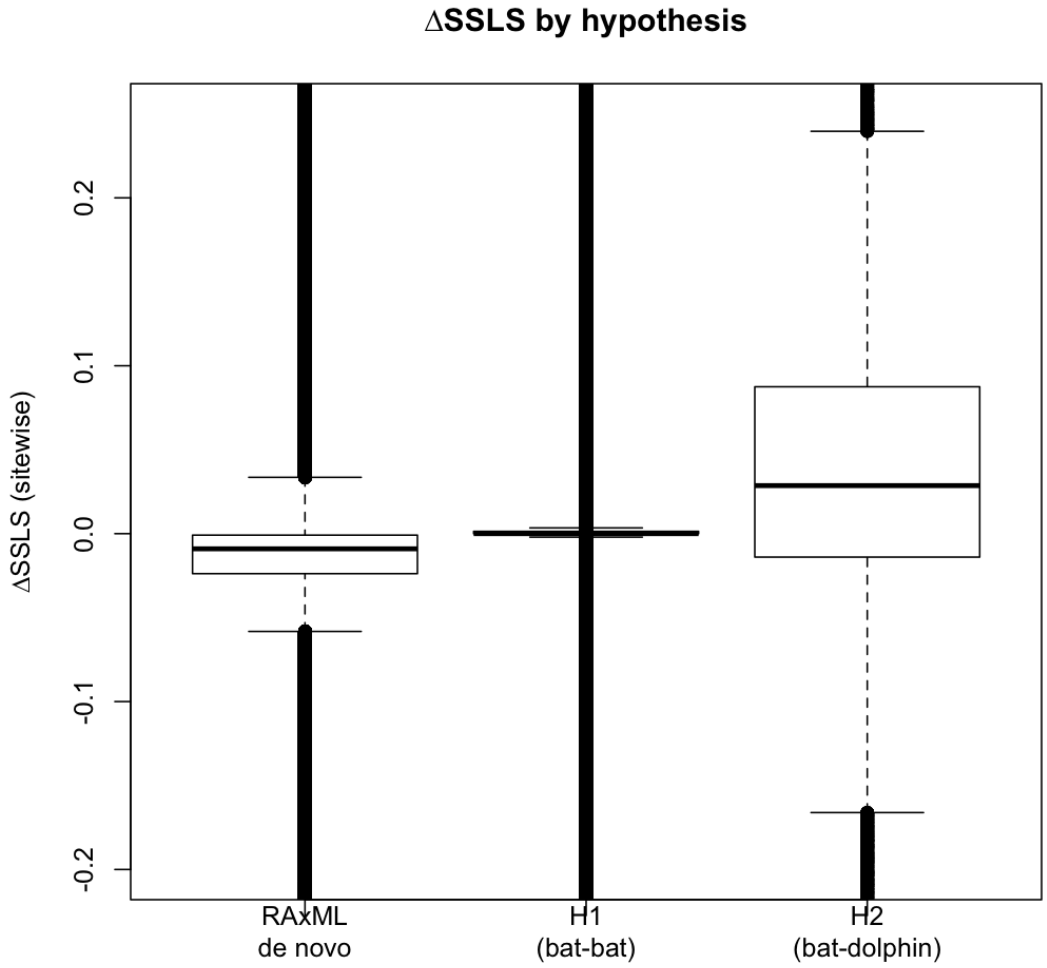

Figure S3: Distribution of ΔSSLS values for all amino acids (n=805,053) calculated under the accepted species phylogeny (see refs. 23-25) compared with a) a *de novo* phylogeny inferred using the software RAxML; b) the H_1 convergent ('bat-bat') phylogeny; and c) the H_2 convergent ('bat-dolphin') phylogeny.

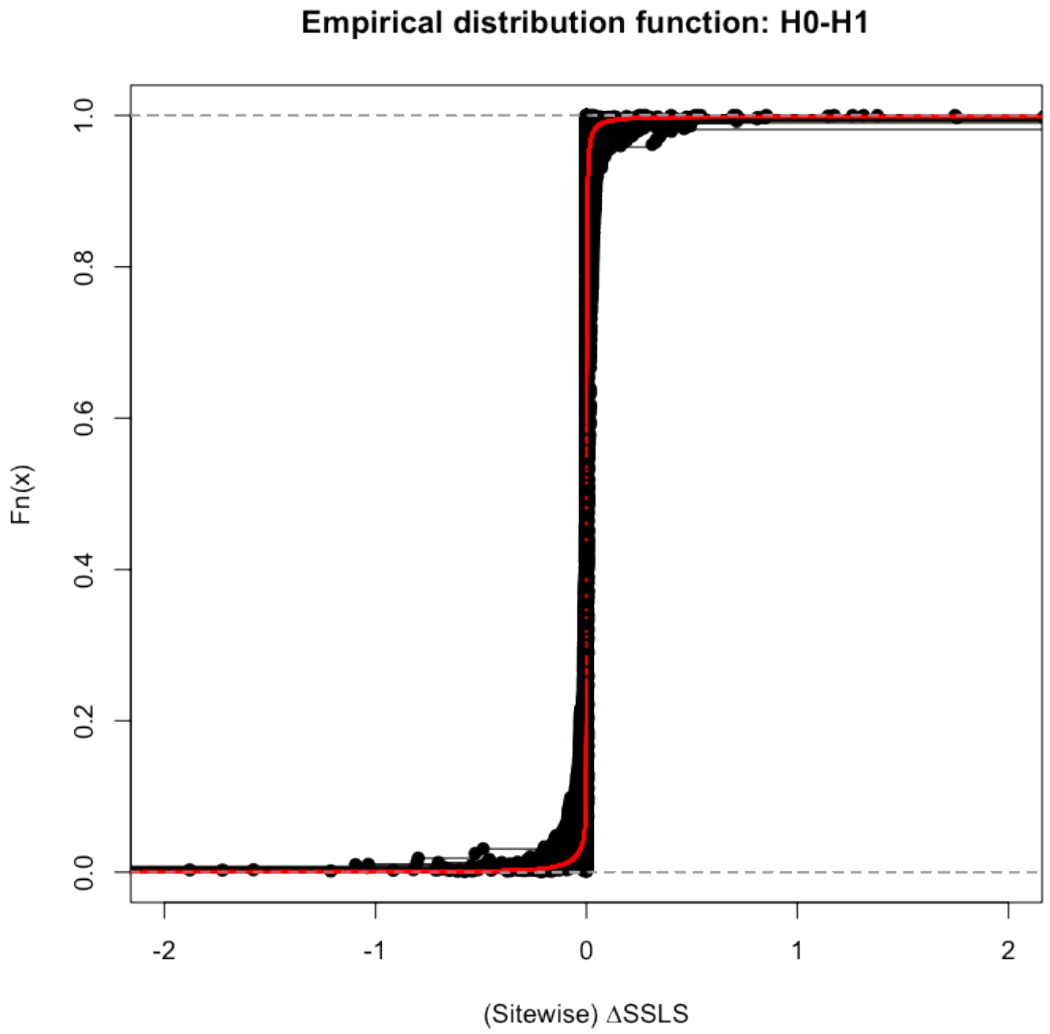

Figure S4: Empirical cumulative distribution function for ΔSSLS values under H_0-H_1. Black lines: individual loci. Red line: all loci.

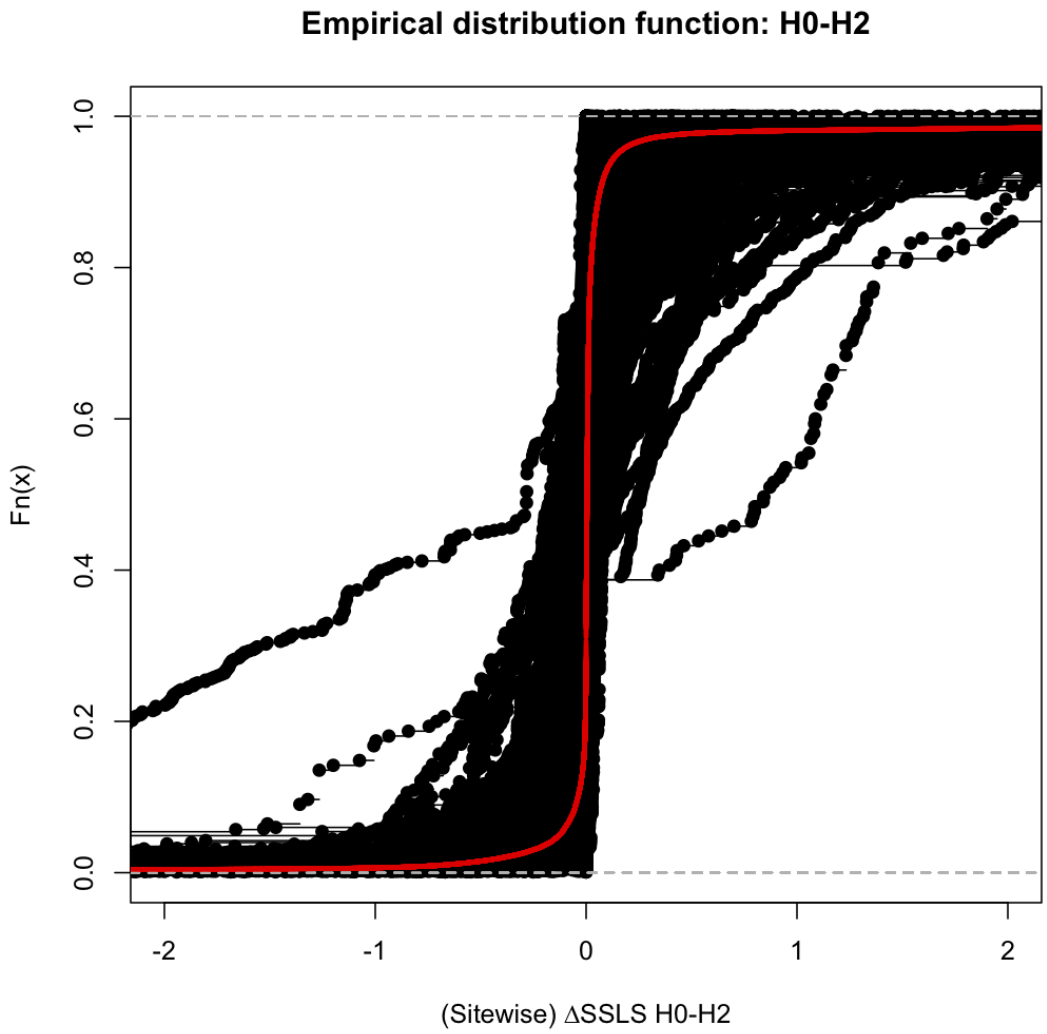

Figure S5: Empirical cumulative distribution function for ΔSSLS values under H_0-H_2. Black lines: individual loci. Red line: all loci.

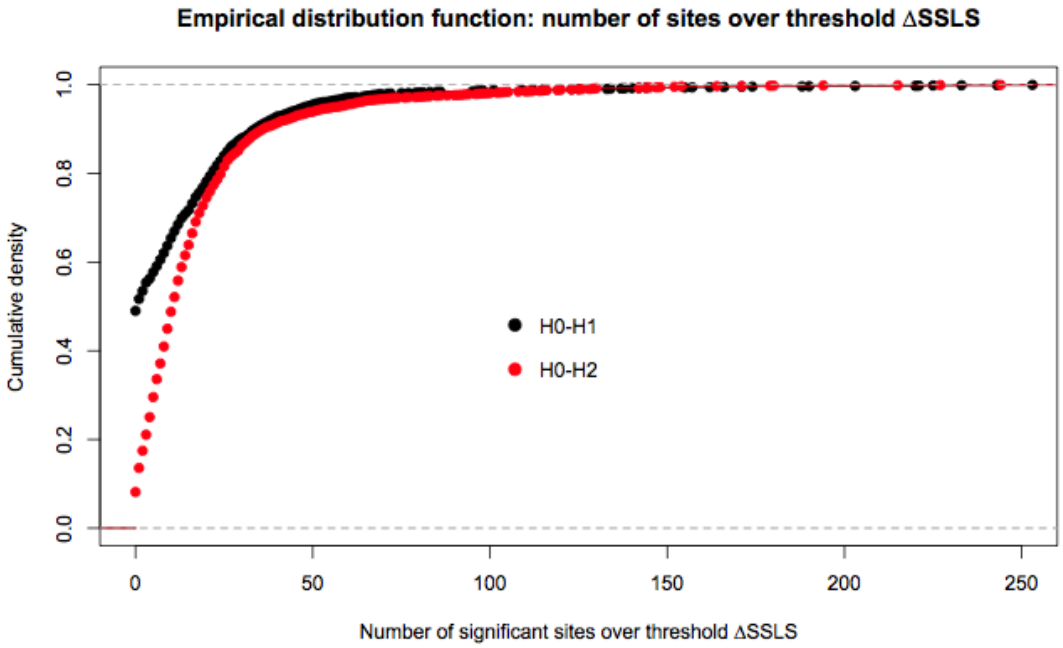

Figure S6: Empirical cumulative distribution function of loci by number of sites with ΔSSLS support greater than threshold for the convergence topology. Within each locus, the total number of amino acids showing significant convergence was counted by comparing observed sitewise ΔSSLS values, with a threshold for ΔSSLS determined by simulation and empirically with reference to the complete set of 805,053 amino acids. Black: H_1 (threshold = -0.182); Red: H_2 (threshold =-0.126). 95% percentiles for these distributions (including ties) occur at 7 and 6 significant site counts for H_1 and H_2, respectively.

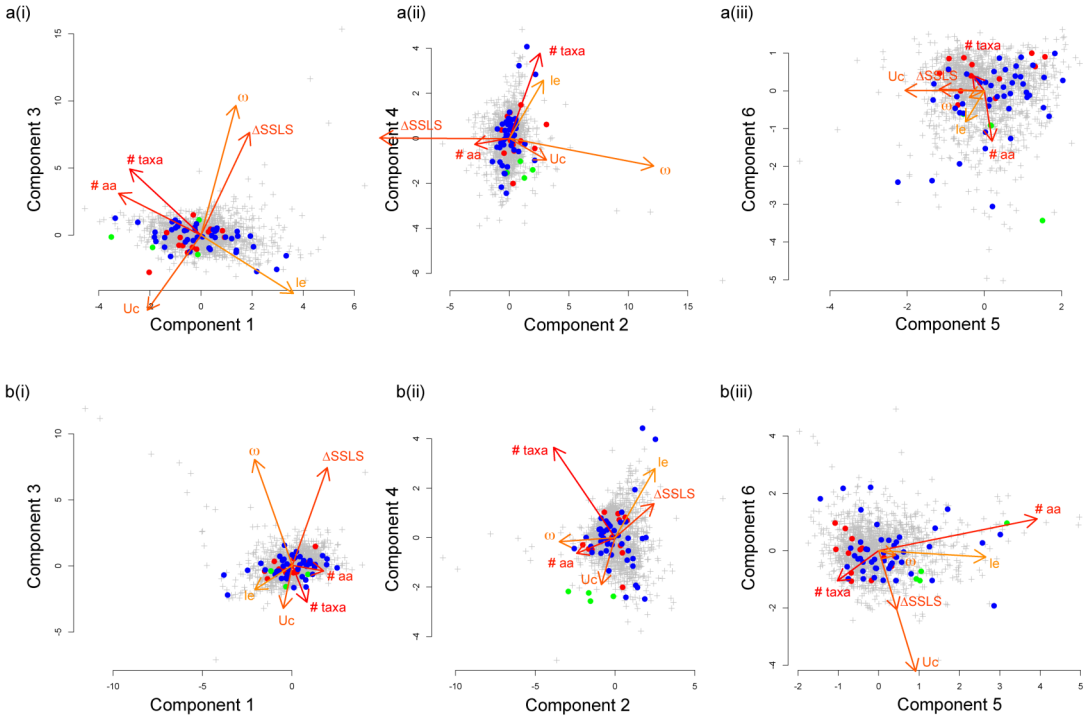

Figure S7: Principal component analysis of locus data. Component loadings for numbers of taxa and amino acids, mean ΔSSLS, corrected U ('Uc'), estimated dN/dS ratio in the convergent clade ('ω') and log-linear exponent ('le'; a measure of the spatial aggregation of selected sites) are shown. Highlighted red, blue and green points show loci we classified as associated with sound or vision, or published in the literature, respectively. a) For H_1, components 1 & 3 (i), 2 & 4 (ii) and 5 & 6 (iii) are shown. b) For H_2, components 1 & 4 (i), 2 & 4 (ii) and 5 & 6 (iii) are shown.

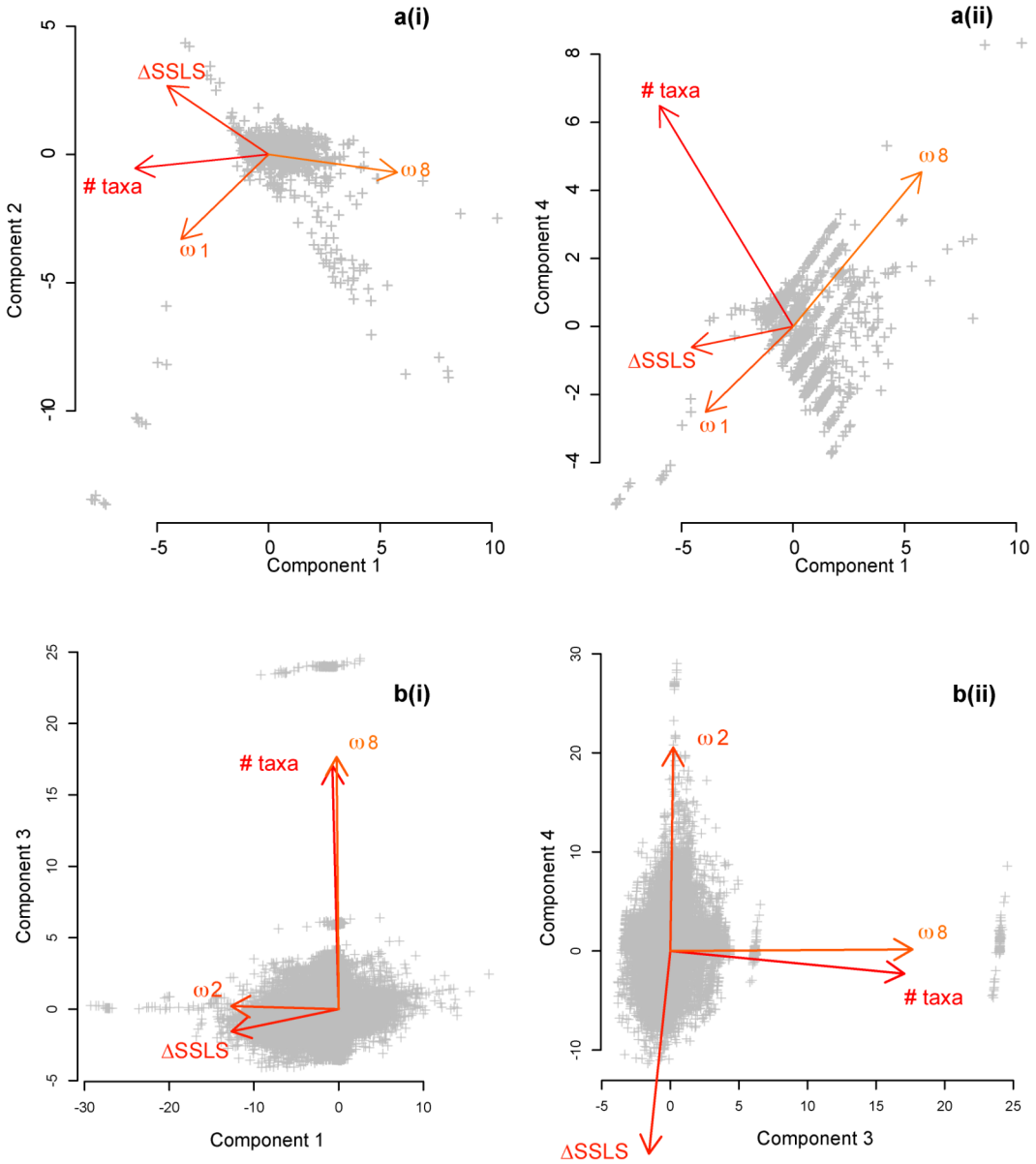

Figure S8: Principal component analysis of sitewise data. Component loadings for numbers of taxa, sitewise ΔSSLS, estimated dN/dS ratio in the convergent clade ('ω1' for H_1 and 'ω2' for H_2) and estimated dN/dS ratio across the whole phylogeny ('ω8') are shown. a) For H_1, components 1 and 2 (i) and 1 and 4 (ii) are shown. b) For H_2, components 1 and 3 (i) and 3 and 4 (ii) are shown.

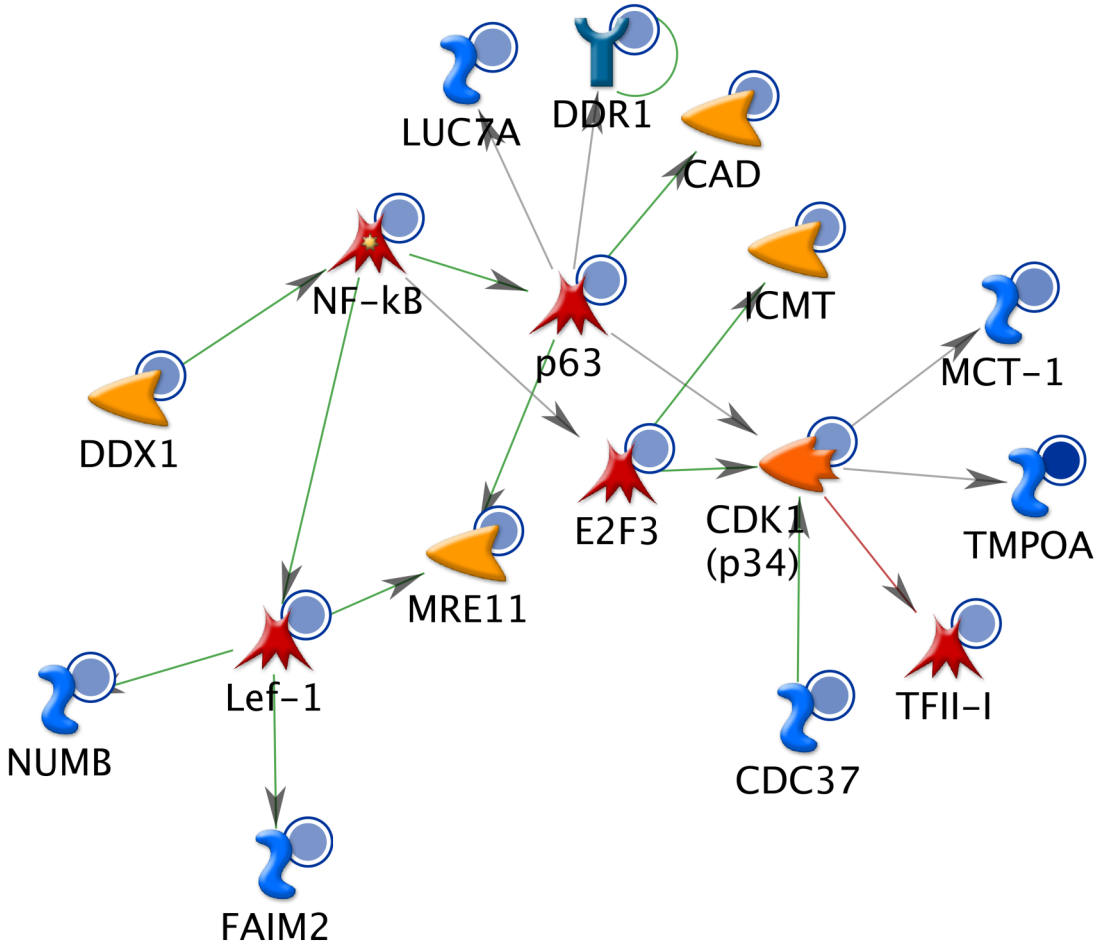

Figure S9: Network of protein-protein interactions described in the literature for 17 proteins encoded by genes taken from the top 5% (n = 117) with strongest signal for H_2 convergence. See Supplementary Methods for details.

Supplementary 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.[13,23-25]) 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[13]; (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. parnelli* & *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[31]) 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_o) 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[32] tool. The resulting assemblies consisted of between 210,080 and 315,526 genomic sequences (depending on species), with an average depth of coverage of 17x to 18x. 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[33] 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[34]. The completeness/contiguity of the gene representation was evaluated using the CEGMA (Core Eukaryotic Genes Mapping Approach) pipeline[35,36] 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[37] incorporating MAFFT[38], 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[39] 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 a*l (2013)[20].

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[40]) 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*[19]; *KCNQ4* (*KQT-4*)[17]; *Pcdh15*, *Cdh23* and *Otoferlin*[18]; and *Pjvk* and *Tmc1*[16]. 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[41], 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[42,43] 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[44], 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[42,43]) and a modified build of version 4.4b of PAML[45]; 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[46] (see Figure S1(i)).

(ii) De novo phylogeny estimation
We generated a separate *de novo* phylogeny for each locus using RAxML 7.6.2[42,43] 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. ΔSSLS 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[45] under the WAG + γ 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[47] showed congruence in relative ΔSSLS estimates for each locus under different models of substitution (available on request). We also repeated our complete ΔSSLS analyses separately under JONES and DAYHOFF; the Pearson's correlation coefficients between ΔSSLS 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 (ΔSSLS)
We used the mean ΔSSLS 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, ΔSSLS, was then calculated:

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

Where ΔSSLS for the i^{th} site is given by the difference in log-likelihood units between the log-likelihood of the i^{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 ΔSSLS, while sites with a better fit to the convergent topologies H_1 or H_2 will have a negative ΔSSLS (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 ΔSSLS show net sitewise signal for convergence (Supplementary Figure S2d). Boxplots showing the distribution of mean ΔSSLS by hypothesis are shown in Figure S3; the top 5% of loci by mean ΔSSLS for H_1 and H_2 are shown in Tables S2 and S3, respectively (*n*=805,053). As expected, mean ΔSSLS for the H_1 and H_2 hypotheses are positive, indicating that a majority of sites in the dataset are not convergent. Equally, mean ΔSSLS 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 ΔSSLS 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 ΔSSLS observations were within ±0.5 lnL units of the mean. However, some sites displayed large ΔSSLS observations: absolute log-likelihood ΔSSLS of >1 lnL 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 ΔSSLS 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 ΔSSLS may influence mean locus ΔSSLS disproportionately in short loci.

(vi) Simulation of expected sitewise ΔSSLS 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[48], 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, ΔSSLS and tree length: ENSG00000008515, ENSG00000095906, ENSG00000121900, ENSG00000167671, ENSG00000170476 and ENSG00000173627. We calculated their sitewise ΔSSLS values in the pipeline and then compared a single randomly-selected replicate's ΔSSLS distribution to that of a variable number of replicates' pooled ΔSSLS 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 ΔSSLS values simulated from a Normal distribution parametised on the alignments' ΔSSLS distribution's means and variances gave a similar result (not shown).

(vii) Determining the significance of ΔSSLS signals.

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

$$U = 1 - cdf\left(\Delta SSLS_{H0\text{-}Ha\,|\,j}\right) \qquad \text{(Equation S2)}$$

(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, ΔSSLS 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_r, and calculated the controlled sitewise U, U_c as:

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

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 ΔSSLS 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 ΔSSLS comparisons.

We filtered loci with corrected unexpectedness $U_c \leq 0$ 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 ω 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 ω 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[45].

The codon models M7 (the null model, with F3x4 frequencies, beta-distributed ω and 10 site categories) and M8 (sitewise selection, also beta-distributed ω with an additional ω category representing positive selection[49] 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[50]. We then considered the sitewise ω 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 ω estimates (and their mean for each locus) derived from this method were compared to clade-specific ω estimates (see below, values available on request) and incorporated into the principal component analysis for site data (see below).

Clade-specific ω estimation and derived sitewise ω

Large ΔSSLS 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 ΔSSLS values represent phylogenetic signal this could be due to neutral drift or diversifying (adaptive) selection, in which case ΔSSLS should be proportional to sitewise ω. To test this we fitted the clade-specific Clade Model C (and null model M1a)[51,52] and estimated ω on the H_1 and H_2 topologies for the hypothesised clade of echolocating taxa in each case. In this model, three separate ω ratios were estimated in the given convergent clade; 'category 0', denoted ω_0, for sites under purifying selection where $0 < \omega < 1$; 'category 1', denoted ω_1, for sites evolving neutrally, where $\omega \equiv 1$; and 'category 2', denoted ω_2, for sites under diversifying selection, in which ω free to take values > 1. In this model, the three ω ratios are estimated by ML and the Bayes empirical Bayes (BEB) posterior probabilities that each site falls into category ω_0, ω_1 or ω_2 calculated. For each site, the ω estimated in each category is then weighted by the BEB posterior probabilities and summed to give an estimate of sitewise ω (also see refs. 16 and 19). In subsequent analyses we treated sites with a BEB posterior for the divergent site category 2 (ω_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[50]. We considered the sitewise clade-specific ω 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, ω in the fruit bats was not estimated directly, but instead was averaged over all non-echolocating taxa in the phylogeny. As a result, similar ω 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[53]. 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 ΔSSLS 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 (ω), can be inflated by potential sequencing, annotation, and alignment errors[54], 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 ω 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 ΔSSLS. 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 ΔSSLS (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 ΔSSLS, number of taxa present, total amino acid count, count of amino acids with significant support (see 'Simulation of expected sitewise ΔSSLS distribution', above), mean ω (dN/dS) for sites in the divergent site class in the convergent clade (see 'Estimation of ω (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 (ΔSSLS), number of taxa present, estimated ω (dN/dS; see 'Estimation of ω (dN/dS ratio)', above) and estimated ω in the convergent clade (see 'Estimation of ω (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 ΔSSLS sites were positively correlated with estimated sitewise ω in the clade of echolocating mammals in both H_1 and H_2. A positive correlation was previously found between sitewise ω and ΔSSLS for the *Prestin* gene[19], i.e. sites under stronger positive selection also showed larger ΔSSLS 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:

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

Where $|\Delta SSLS|$ is the absolute value of the site-specific log-likelihood support; α the line constant; β the regression coefficient; ω_2 is the clade-specific estimated sitewise ω (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 ω estimation and derived sitewise ω', 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 ω and ΔSSLS, and also showed support for convergence based on a negative mean ΔSSLS, but were not categorised as convergent in Figure 2 because their ΔSSLS did not exceed the threshold we established above. i.e.,

the value of the empirical cumulative density function (eCDF) of mean locus-wise ΔSSLS 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 ΔSSLS value at ω=2 using the fitted linear relationship in each locus. We predicted two trajectories for each locus: an upper estimate of ΔSSLS 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 ΔSSLS estimate passed the ΔSSLS threshold and 'possibly convergent' if the upper estimate failed the ΔSSLS threshold, but the central estimate passed it.

'Convergent' (Table S6 and S7 for H_1 and H_2, respectively) or 'possibly convergent' predictions (top 50 loci are shown in Tables S8 and S9 for H_1 and H_2, 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_1 (one-tailed T = -2.11; $P \leq 0.019$ with ~159 d.f.) and significantly stronger for hearing loci compared to the background set in H_2 (one-tailed T = -2.23; $P \leq 0.016$ 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 ΔSSLS and by the number of sites per locus with significant sitewise ΔSSLS, under the H_1 (bat-bat) and H_2 (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 ΔSSLS 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 ΔSSLS 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_1) and S11 (H_2).

Functional enrichment analysis

Having screened these genomes for convergence, we compiled lists of the most convergent (strongest ΔSSLS 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_1 and H_2) was carried out using Fisher's exact test and Benjamini-Hochberg[50] 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[55], KEGG pathways[56], common interactors according to the HPRD database[57], microRNA targets according to Targetscan[58] and genes associated to human diseases according to the Genetic Association Database[59].

At a Benjamini-Hochberg[50] FDR of 30%, and retaining only functional categories represented by at least 5 genes in our H1 and H2 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_1 genes, pertaining to alcohol metabolic processes. For H_2 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_1 and H_2).

Several loci with the most strongest signals for H_2 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_2 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_1 and H_2 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_2 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[30]. Similarly, recent work by Terrinoni *et al.* (2013)[29] suggests that *p63* is involved in cochlear development. Networks extracted from H_1 and random tree query lists were much smaller, with a maximum of six loci present.

Supplementary References

31 Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol*. **29**:1969-1973 (2009).

32 Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966-1967 (2009).

33 She, R. *et al.* genBlastG: using BLAST searches to build homologous gene models. *Bioinformatics* **27**, 2141-2143 (2011).

34 Kim, E. B. *et al.* Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nature* **479**, 223-227 (2011).

35 Parra, G., Bradnam, K & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, **23**:1061-1067 (2007).

36 Parra, G., Bradnam,K, Ning, Z, Keane, T and Korf, I. Assessing the gene space in draft genomes. *Nucleic Acids Research*, **37**:298-297 (2009).

37 Bininda-Emonds, O. R. transAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. *BMC Bioinformatics* **6**, 156 (2005).

38 Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511-518 (2005).

39 Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540-552 (2000).

40 Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44-57 (2009).

41 Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).

42 Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690 (2006).

43 Stamatakis, A. in *Proceedings of 20th IEEE/ACM International Parallel and Distributed Processing Symposium (IPDPS2006),* High Performance Computational Biology Workshop, Rhodos, Greece, (2006).

44 Hancock, A. M. et al. Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *PNAS* **107**:8924–8930. (2010).

45 Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586-1591 (2007).

46 Junier, T. & Zdobnov, E. M. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics* **26**, 1669-1670 (2010).

47 Ranwez, V. *et al.* OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. *BMC Evol. Biol.* **7**, 241 (2007).

48 Lartillot, N., Lepage, T. & Blanquart, S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*. **25**(17):2286-8 (2009).
49 Yang, Z., Nielsen, R., Goldman, N. & Pedersen, A. M. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**, 431-449 (2000).
50 Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J. Roy. Stat. Soc. B Met.* **57**, 289-300 (1995).
51 Bielawski, J. P. & Yang, Z. H. A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. *J. Mol. Evol.* **59**, 121-132 (2004).
52 Wong, W. S., 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* **168**, 1041-1051 (2004).
53 Fitch, W. M. & Margoliash, E. Construction of phylogenetic trees. *Science* **155**:279-284 (1967).
54 Schneider, A. *et al.* Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome Biol. Evol.* **1**, 114-118 (2009).
55 The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**:25-9 (2000).
56 Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M.; KEGG for integration and interpretation of large-scale molecular datasets. *Nucleic Acids Res.* **40**:D109-D114 (2012).
57 Prasad, T. S. K. *et al.* Human Protein Reference Database - 2009 Update. *Nucleic Acids Research.* **37**:D767-72 (2008).
58 Lewis B. P., Burge C. B. & Bartel D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**:15-20 (2005).
59 Becker, K. G., Barnes, K. C., Bright, T. J. & Wang, S. A. The Genetic Association Database. *Nature Genetics* **36**:431-432 (2004).