Pangolin genomes and the evolution of mammalian scales and immunity

The International Pangolin Research Consortium (IPaRC)

Siew Woh Choo, Mike Rayko, Tze King Tan, Ranjeev Hari, Aleksey Komissarov, Wei Yee Wee, Andrey A. Yurchenko, Sergey Kliver, Gaik Tamazian, Agostinho Antunes, Richard K. Wilson, Wesley C. Warren, Klaus-Peter Koepfli, Patrick Minx, Ksenia Krasheninnikova, Antoinette Kotze, Desire L. Dalton, Elaine Vermaak, Ian C. Paterson, Pavel Dobrynin, Frankie Thomas Sitam, Jeffrine J. Rovie-Ryan, Warren E. Johnson, Aini Mohamed Yusoff, Shu-Jin Luo, Kayal Vizi Karuppannan, Gang Fang, Deyou Zheng, Mark B. Gerstein, Leonard Lipovich, Stephen J. O'Brien and Guat Jah Wong

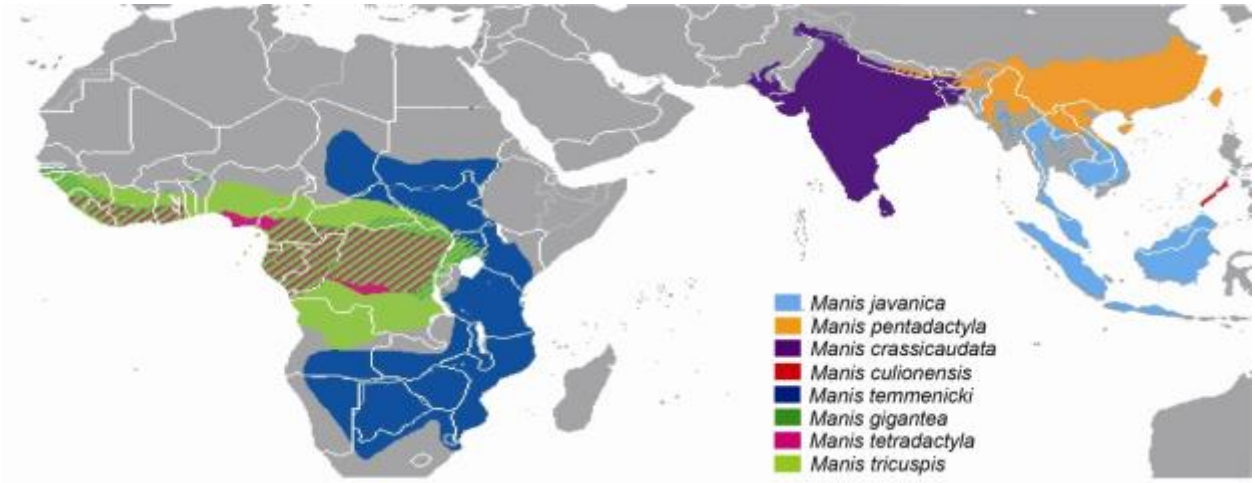
Supplemental Information

1.0 GENOME SEQUENCING, ASSEMBLY AND SCAFFOLDING

1.1 DNA extraction and whole-genome Sequencing

Malayan pangolins are mainly from Southeast Asia, whereas the Chinese pangolins are mainly from China, Taiwan and in some Northern of Southeast Asia (**Supplemental Figure S1.1)**. For Malayan pangolin, we used a female wild Malayan pangolin for the study. The animal was provided by the Department of Wildlife and National Parks (DWNP) Malaysia under a Special Permit No. 003079 (KPM 49) for endangered animals. DNA was extracted using the Qiagen 20/G genomic tip following manufacturer's protocol. Libraries for the Malayan pangolin genome were constructed and the different insert sizes of the libraries were used: 180bp, 500bp, 800bp, 2kb, and 5kb (**Supplemental Table S1.1)**. The libraries were sequenced using Illumina HiSeq 2000 at BGI, Hong Kong.

For Chinese pangolin, the DNA was derived from a single female animal (sample ID=MPE899) collected in the island of Taiwan. The library plan followed the recommendations provided in the SOAPdenovo assembler manual(Luo et al. 2012). Libraries for the Chinese pangolin genome were constructed and the different insert sizes of the libraries were used: 200-300 bp, 3 kb and 8 kb.



Supplemental Figure S1.1. Pangolin distribution and comparisons. Geographical distribution maps of the pangolins. There are eight known pangolin species represented by different colors in the map. The hash mark on the map means overlap of geographic areas where more than one species inhabit the same range. (Source: Modified picture from https://commons.wikimedia.org/wiki/File:Manis_ranges.png)

1.2 Data pre-processing and Next-Generation Sequencing (NGS) libraries reads statistics

For Malayan pangolin the quality of the sequenced reads was assessed by FastQC tool(Andrews). For Malayan pangolin, the sequencing reads were filtered using PRINSEQ(Schmieder and Edwards 2011) with an average quality score above 20. The resulting filtered sequences were then error-corrected using MUSKET 1.1(Liu et al. 2013). The error-corrected reads were subsequently used for *de novo* genome assembly.

For Chinese pangolin, all production data received automated comprehensive quality checks to confirm each sample has met specific metrics related to quality (>20 score minimum) and coverage. Quality control of WGS data was evaluated using the PICARD software package (http://broadinstitute.github.io/picard) module CollectWGSMetrics which provides both depth and breadth of coverage measurements.

Supplemental Table S1.1: An NGS library reads statistics. Number of reads for each library and the estimated sequencing coverage are shown. The sequencing coverage was estimated based on the predicted genome size of Malayan pangolin (2.5Gbp) and Chinese pangolin (2.7Gbp) by *k*-mer analyses (**Supplemental Figure S1.2)**.

Library	Library Type	Total Paired-End (PE) Reads	Sequencing Coverage
Malayan pangolin			
PE 180bp	Paired End	189,915,801	15.19
PE 180bp	Paired End	212,107,889	16.97
PE 500bp	Paired End	160,493,968	12.84
PE 500bp	Paired End	176,434,714	14.11
PE 500bp	Paired End	165,713,431	13.26
PE 800bp	Paired End	115,914,999	9.27
PE 800bp	Paired End	130,704,648	10.46
PE 800bp	Paired End	129,334,465	10.35
MP 2000	Mate-pair	115,649,108	9.25
MP 5000	Mate-pair	115,639,514	9.25
MP 5000	Mate-pair	154,755,684	12.38
MP 5000	Mate-pair	154,736,404	12.38
			Total: 146.07
Chinese pangolin			
PE 206	Paired End	437,447,710	32.40
PE 356	Paired End	296,308,797	21.94
MP 3000	Mate-pair	25,634,681	1.90
MP 8000	Mate-pair	1,172,455	0.08
			Total: 56.32

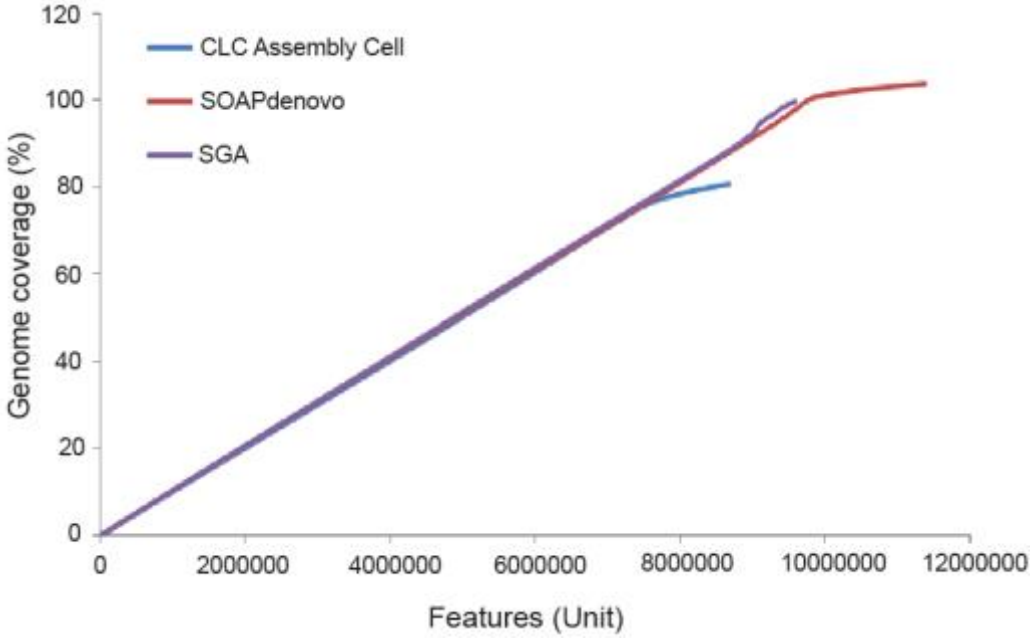
1.3 Genome assembly, scaffolding and gap closing

1.3.1 *Malayan pangolin genome*

The genome assembly was performed using CLC Assembly Cell 4.10, SGA-0.10.10(Simpson and Durbin 2012) and SOAPdenovo2(Luo et al. 2012). The three generated genome assemblies

were compared based on N50 metric using QUAST(Gurevich et al. 2013). After mapping the reads back to the assemblies, the Feature Response Curve (FRC) was produced using the FRCbam program(Vezzi et al. 2012). The sequencing reads were mapped back to the genomes and cumulative feature density were plotted for each genome assembled using CLC, SOAPdenovo2 and SGA-0.10.10(Luo et al. 2012; Simpson and Durbin 2012). The higher the number of features per cumulative nucleotide size, the assembly was regarded to be better. In order to further validate the assembly, the assembled RNA-seq transcripts from Trinity(Grabherr et al. 2011) were mapped to these assemblies. The assembly with best FRC curve and high RNA-seq mappings were regarded as the best assembly and used for the scaffolding step. Using the scaffolder module in SOAPdenovo2, the contigs were scaffolded. The scaffolded genome assemblies were used for subsequent validation analysis.

To choose the optimal assembly from the pangolin assemblies generated from the three different assemblers, we compared them using Feature Response Curve, where the steepest curve denotes the best assembly while covering the estimated genome size of 2.5Gbp (**Supplemental Figure S1.2**). The CLC assembler-generated assembly (N50: 13,423) fell short for the genome coverage metric while SOAPdenovo2-generated assembly (N50: 4,836) showed good coverage yet the SGA-generated assembly (N50: 17,568) contained slightly less genomic features hence a slightly steeper curve. Taken all together, we concluded that the SGA-generated assembly was the best and used it for subsequent analyses.



Supplemental Figure S1.2. Feature response curves. Comparison across three different assemblies showed cumulative features per genomic coverage plot of the three different assemblies of Malayan pangolin.

The selected final SGA-generated assembly was scaffold using SOAPdenovo2 scaffolder achieved better contiguity statistics where the final scaffold N50 is 204,525 (**Supplemental Table S1.2)**. The resulted scaffolds which had many ambiguous gap positions with Ns were gap-closed and this step improved the contig N50 to 18,812 bp.

Supplemental Table S1.2. Summary genome assembly statistics for Malayan pangolin.

	Paired-end insert size	Estimated coverage (X)	N50 (bp)	Total length (bp)
Initial contig (>1k)	180bp, 500bp, 800bp	102.45	17,568	2,047,445,145
Final scaffold (>1k)	180bp, 500bp, 800bp, 2000bp, and 5000bp	145.66	204,525	2,549,959,554
Final contig (after gap-closing)			18,812	2,108,780,390

1.3.2 *Chinese pangolin genome*

Total assembled sequence coverage of Illumina instrument reads was approximately 56X (using a genome size estimate of 2.7Gbp. The first draft assembly was performed with SOAPdenovo v1.0.5 and was referred to as *M. pentadactyla* 1.0. In the *M. pentadactyla* 1.0 assembly, small scaffold gaps were closed with Illumina read mapping and local assembly. Contaminating contig and trimmed vector in the form of X's and ambiguous bases as N's in the sequence were removed. The National Center for Biotechnology Information (NCBI) requires that all contigs with genomic size less than 200bp to be removed. Removing these small contigs was the laststep in preparation for submitting the final 1.1.1 assembly. The *M. pentadactyla* 1.1.1 assembly is made up of a total of 92,722 scaffolds with an N50 scaffold length of 118,853bp (N50 contig length=28,718bp). Including gaps, the total assembly spans about 2.2Gbp (**Supplemental Table S1.3**).

To improve assembly and increase scaffold N50 length, the resulting Chinese pangolin assembly was scaffolded using L_RNA_scaffolder approach, with the support of Malayan pangolin transcriptomic data(Xue et al. 2013). It has reduced the number of scaffolds to 87,621 with N50 of 157,892bp.

Supplemental Table S1.3. Summary genome assembly statistics for Chinese pangolin.

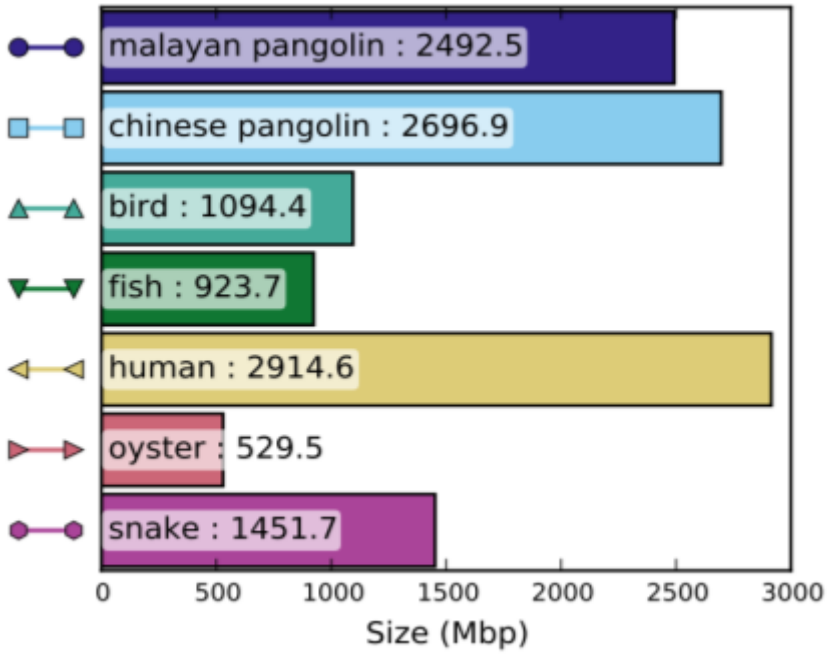
Assembly	**Contigs**	**Scaffolds**	**L_RNA scaffolding**
# contigs/scaffold (>= 0 bp)	230,930	92,772	87,621
# contigs/scaffold (>= 1000 bp)		38,256	33,682
Total length (>= 0 bp)	1,999,057,008	2,204,732,179	2,205,289,822
Largest contig/scaffold	292,755	1,317,973	1,402,852
N50	28,718	118,853	157,892
# N's per 100 kbp	0	9385.12	9407.54

1.4 Genome size estimation

To estimate the genome size, *k*-mer was counted using the tool called Preqc from the String Graph Assembler (SGA) assembly program(Simpson and Durbin 2012). The tool counts and plots 31-mer histogram from 20,000 reads and estimates the genome size by identifying the peak of the Poisson distribution. The tool also accounts for bias for heterozygosity and performs correction as required. The estimation is based on the principle as follows, where the mean number of times a unique genomic *k*-mer appears in the reads, λ_k is as follows:

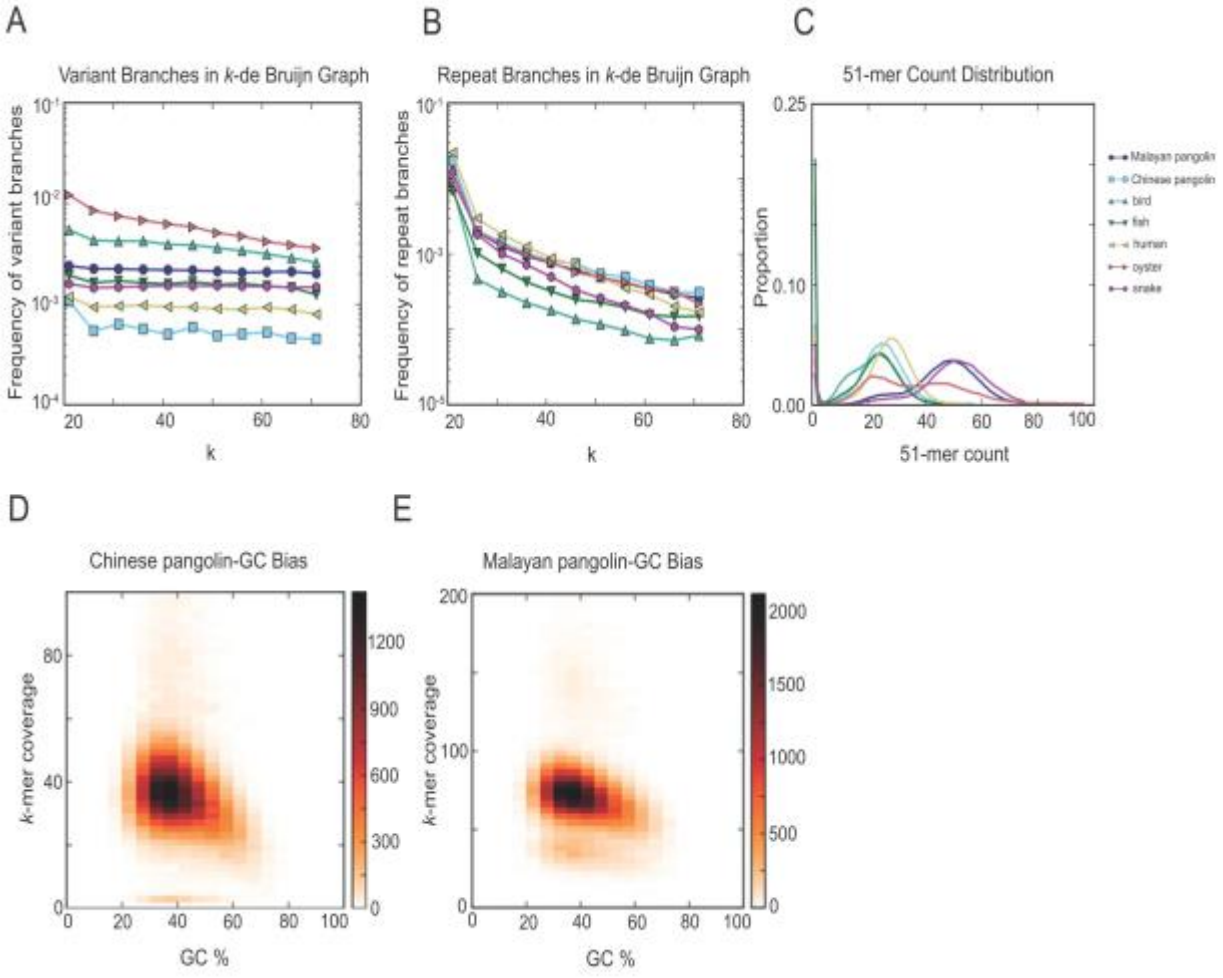
$$\lambda_k = \frac{n(l-k+1)}{G}$$

Where n is the number of reads, l is the read length and G being the genome size. By pooling all the short-insert reads for estimation, the genome size of Malayan pangolin and Chinese pangolin was estimated to be 2,492,544,425 bp and 2,696,930,760 bp, respectively (**Supplemental Figure S1.3)**.



Supplemental Figure S1.3. Genome size prediction of the Malayan and Chinese pangolins compared to other organisms using 31-mer counting. The datasets of other organisms used in this analysis were from previous study(Bradnam et al. 2013). Bird=*Melopsittacus undulatus*; Fish=*Maylandia zebra*; Oyster=*Crassostrea gigas*; and *Snake=Boa constrictor constrictor*.

1.5 NGS reads based characterisation



Supplemental Figure S1.4. NGS reads based characterisation of pangolin assemblies. (a) Frequency of variants branches in the *k*-de Bruijn graph. The Malayan pangolin genome is expected to be highly heterozygous like the bird genome and is expected to have higher heterozygosity compared to the Chinese pangolin which possesses lower variant branches. (b) Frequency of repeats branches in the *k*-de Bruijn graph. Like oyster genome, high number of repeats content is to be expected in the Malayan and Chinese pangolin. (c) 51 *k*-mer distribution. The data for Malayan pangolin followed a bimodal curve while the Chinese pangolin appears unimodal. The second peak is the true genomic *k*-mers of which when abundant, aids to better assemble the genome. Therefore, the Malayan pangolin reads appear to have abundant *k*-mers required for proper assembly. (d) GC content plot of the Chinese pangolin genome. (e) GC content plot of the Malayan pangolin genome.

2.0 EVALUATION OF GENOME ASSEMBLIES

2.1 Mapping Malayan pangolin transcripts to assembled genomes

We have 89,754 consensus transcripts/UniGenes generated by combining transcriptomic fragments from three different assemblers: Trinity(Grabherr et al. 2011), SOAPdenovo(Luo et al. 2012) and Velvet(Zerbino and Birney 2008) in our Malayan pangolin transcriptome project (manuscript in preparation). Briefly we sequenced the transcriptomes of eight pangolin organs including cerebellum, cerebrum, lung, heart, kidney, liver, spleen and thymus using Illumina HiSeq technology platform (2x100bp strategy). We *de novo* assembled the pooled sequencing reads from the eight samples using three different assemblers independently. The assembled transcriptomic fragments/transcripts were clustered using CD-Hit-EST program(Huang et al. 2010) with a clustering threshold of 98% sequence identity. The longest sequence representatives in each clustered transcripts were selected and classified as the UniGenes. We selected the 89,754 consensus UniGenes (the transcript clusters that have transcripts generated from the three different assemblers) for this analysis.

Both pangolin assemblies were mapped using the set of 89,754 consensus transcripts (as well as the set of 1,035,201 transcripts generated by Trinity alone(Grabherr et al. 2011)) using GMAP software(Wu and Watanabe 2005). The summary of the percentage of mapped consensus UniGenes and the Trinity-generated transcripts were calculated (**Supplemental Table S2.1**). In general, at least 93% of the transcripts were mapped to both pangolin assemblies for both sets of pangolin transcripts that we used, indicating the high quality of our assemblies for genome annotation.

Supplemental Table S2.1. **Summary statistics of the mapped consensus UniGenes and Trinity-generated transcripts.**

89,754 Consensus Unigenes		
	Malayan pangolin	**Chinese pangolin**
Unmapped transcripts	282 (0.31%)	4654 (5.19 %)
High quality mapped transcripts (alignments with 92% identity and 80% coverage)	75321 (83.91%)	58125 (64.76 %)

Mapped transcripts	14151 (15.76%)	26975 (30.05%)
1,035,201 Trinity-generated Transcripts		
	Malayan pangolin	**Chinese pangolin**
Unmapped transcripts	8419 (00.81%)	58326 (5.63 %)
High quality mapped transcripts (alignments with 92% identity and 80% coverage)	907603 (87.67%)	712245 (68.80 %)
Mapped transcripts	119179 (11.51%)	264630 (25.56%)

2.2 CEGMA analysis

We evaluated the genome assembly quality of Malayan pangolin with the Core Eukaryotic Genes Mapping Approach (CEGMA) pipeline(Parra et al. 2007). CEGMA is a computational method that relies on a defined set of ultra-conserved eukaryotic protein families for building a highly reliable set of gene annotations. The gene space completeness of the pangolin assemblies was given by the CEGMA pipeline identified and the complete hits and partial hits were taken into account. The gene space completeness statistics showed that our draft genome is a good candidate for genome annotation and subsequent analysis as it has high gene space completeness level similar to other eukaryotic genome projects. We used a set of 248 core ultra-conserved genes typical in a CEGMA analysis. Our analyses indicated 91% ultra-conserved eukaryotic genes present in the Malayan pangolin genome and 58% were considered complete genes. The "complete genes" denotes high confident alignments (with an internal threshold) using an ENSEMBL method of BLAST, Genewise and GeneID searches, which also includes Hidden Markov model (HMM)-profiled protein sequences to increase its reliability in predicting the ortholog's gene structures. For the Chinese pangolin assembly, we found 88% ultra-conserved eukaryotic genes present in this genome, 55% of which were considered complete genes.

2.3 Pangolin genome comparison

We also aligned the Chinese pangolin genome sequence with the Malayan pangolin genome sequence using the MUMmer software(Delcher et al. 2003). We found 91% of the Chinese

pangolin assembly covered 78% of the Malayan pangolin assembly with 92% sequence identity. The low genome similarity supports our view that both pangolin species are highly divergent.

3.0 GENOME ANNOTATION

3.1 Protein-coding gene annotation

To predict genes in the pangolin genomes, we used the MAKER annotation pipeline based on several sources of evidence: (i) *ab initio* gene prediction, (ii) transcriptomic data from Malayan pangolin and (iii) protein evidence from *Canis familiaris* reference genome and transcriptome(Cantarel et al. 2008). We identified 23,446 and 20,298 protein-coding genes in the Malayan and Chinese pangolin genomes, respectively (**Supplemental Table S3.1)**. We attempted to assign orthology between the annotated genes of two pangolin species. Of the 23,446 and 20,298 predicted genes for Malayan and Chinese pangolins, we found 12,599 orthologous groups, as well as 10,772 and 7,523 orphans, respectively. We believe that the difference between the gene counts is mainly due to the high level of diversification and divergence within pangolins likely from evolutionary adaptations to different environments, especially the Malayan pangolin was originated from megadiverse Malaysia expected to have high heterozygosity rate compared to the Chinese pangolin originated from an isolated island population. The Malayan pangolin genome has 170,236 exons, whereas the Chinese pangolin genome has 147,455 exons. The total exon length accounted for approximately 1.81% and 1.34% of the Malayan and Chinese pangolin genomes, respectively. All protein-coding genes were annotated using different databases (**Supplemental Information S3.1.1**).

Supplemental Table S3.1. Summary of genome annotation for both pangolin genomes.

	Malayan pangolin	Chinese pangolin
Number of genes	23,446	20,298
Number of exons	170,236	147,455
Number of five_prime_UTR	10,925	8,535
Number of three_prime_UTR	10,919	8,817
Total exons length (bp)	45,284,578	36,081,291

3.1.1 Multiple functional assignments of genes

For functional annotation of pangolin MAKER-generated protein-coding genes, we used InterProScan 5 pipeline(Jones et al. 2014). The number of genes in two pangolins with homologs of functional assignment from various databases including InterPro, GO, KEGG, Swissprot and TrEMBL (**Supplemental Table S3.2**). Encouragingly, all MAKER-generated protein-coding genes had functional assignments from at least one database.

Supplemental Table S3.2. Functional annotation of protein-coding genes.

	Malayan pangolin		Chinese pangolin	
	Number of genes	Percentage of genes (%)	Number of genes	Percentage of genes (%)
Total	23,448	100	20,298	100
InterPro	23,446	100	20,298	100
GO	14,996	64.0	13,632	67.2
KEGG	899	3.8	1,195	5.9
Swissprot	17,408	74.2	19,038	93.8
TrEMBL	21,474	91.6	19,238	94.8

3.2 Pseudogene identification

Pseudogene screening for both Malayan pangolin and Chinese pangolin genomes were performed using Pseudopipe pipeline(Zhang et al. 2006). Maker-annotated protein sequences were used to identify all possible pseudogenes present in pangolin genomes. All protein sequences were BLASTed against the genome sequences. The e-value cut-off ($\leq 1 \times 10{-}4$) was used to identify significant homologous hits. The BLAST hits were then partitioned according to the scaffold ID and strand direction. BLAST hits that had overlap >30bps with the functional genes were discarded. The partitioned BLAST hits were then categorized into different sets based on the match of the hits on the similar or different query protein. Each disjoint set was further merged into single super-hit or pseudo-exon. The pseudo-exon was further extended to both directions for 30 nucleotides to achieve optimal alignment by using the tfasty program from the fasta suite that refines the BLAST result for identification of frameshift, stop codons, indels and calculates accurate sequence similarity(Pearson and Lipman 1988). To get high quality pseudogene prediction, the output results were further filtered based on the sequence similarity cut-off (>40%), BLAST e-value < $1x10^{-10}$ and the predicted pseudogenes must cover at least 70% of the parent genes. The pseudogenes were classified into processed pseudogene (retrotransposed pseudogene) and duplicated pseudogenes. Processed pseudogene were pseudogenes lack of intron, possessed small flanking direct repeats and a 3' polyadenine tail, whereas the duplicated pseudogenes has multiple exons.

Genome-wide screening of pseudogenes revealed a total of 3,316 pseudogene in the Malayan pangolin, which includes 2,649 processed pseudogenes (79.8%) and 667 duplicated pseudogenes (20.1%) (**Supplemental Table S3.3)**. On the other hand, a total of 1,577 pseudogenes were found in the Chinese pangolin genome including 1,227 processed pseudogenes (77.8%) and 350 duplicated pseudogenes (22.1%). The number of pseudogenes in both pangolin species are lower that the number of predicted pseudogenes in human and dog.

Supplemental Table S3.3. Pseudogene identification across different mammalian species. Number of processed and duplicated genes are shown for each species. The numbers in bracket are the percentage of pseudogenes. The human and dog pseudogene datasets were from www.pseudogene.org.

Pseudogene Type	Malayan pangolin	Chinese pangolin	Dog	Human
Processed	3637 (78.05%)	1859 (76.95%)	6075 (86.23%)	8522 (82.25%)
Duplicated	1023 (21.95%)	557 (23.05%)	970 (13.77%)	1839 (17.75%)
Total	4610	2416	7045	10361

3.3 Non-coding genes

Non-coding RNA genes were annotated using RFAM sequence database and Infernal engine (v.1.1.1), according to Ensembl recommendations (http://www.ensembl.org/info/genome/genebuild/ncrna.html). Genomic sequences were aligned against RFAM database using BLASTN with e=10-5. The BLAST hits were clustered and used to seed Infernal searches with the corresponding RFAM covariance models. The resulting BLAST hits were used as supporting evidence for ncRNA genes confirmed by Infernal.

A total of 1,594 and 1,506 miRNAs from the Malayan and Chinese pangolin genomes, respectively were identified using two complementary approaches *ab initio* with HHMMiR(Kadri et al. 2009) and MiRPara(Wu et al. 2011), by similarity to known miRNA genes in miRBase(Griffiths-Jones 2006; Griffiths-Jones et al. 2008). The miRNA sequences accounted <1% of the pangolin genomes with the TE-related mir-9256a-1 and MIR396c being the most abundant families (**Supplemental Table S3.4**). The summary statistics of other types of non-coding genes including miscRNAs, ribosomal RNAs (rRNAs) and small nuclear RNAs (snRNAs) are also shown in Supplementary Table 3.4.

For Transfer RNA (tRNA) gene prediction, we excluded the predicted tRNA genes from the RFAM results and annotated them using tRNAscan-SE with default parameters(Lowe and Eddy 1997). Predictions with a cove score less than twenty were filtered out. We found 267 and 280

predicted tRNA, these numbers are lower than in Carnivora genomes (3,039 in cat genome, and 905 in dog)(Chan and Lowe 2009). It can be explained by carnivora specific SINE elements based on tRNA absent in pangolins(Vassetzky and Kramerov 2002).

Supplemental Table S3.4. Summary statistics on non-coding RNA annotation. No=Number of non-coding genes; bp=total number of bases of the non-coding genes.

Type of non-coding gene		**Malayan pangolin**	**Chinese pangolin**
miRNA	№	1069	1014
	bp	82371	78762
miscRNA	№	682	552
	bp	86278	73945
rRNA	№	68	64
	bp	37730	30024
snRNA	№	1048	864
	bp	130169	104188
snoRNA	№	460	424
	bp	47775	44312
Total loci		313232	256323
Total length (bp)		24186818	19263171

Supplemental Table S3.5. Ten abundant families of miRNA in pangolins genomes.

Family	**Rank**	**#CP**	**Rank**	**#MP**	**Comments**
mir-9256a-1	1	775	1	840	TE-related
MIR396c	2	458	2	362	Plants
mir-8485	6	291	3	359	Neurexins
mir-9256a-2	7	238	4	297	TE-releated

MIR160	3	354	5	286	Plants
MIR169f	5	302	6	201	Plants
MIR6025e	4	320	7	192	Plants
mir-466q	10	103	8	169	Mammals
MIR171d	14	45	9	105	Plants
MIR156e	8	168	10	99	Plants
MIR396e	9	114	11	81	Plants

3.4 Segmental duplications

Segmental duplication (SD) sets for Malayan pangolin and Chinese pangolin were constructed using a modified WGAC method(Gokcumen et al. 2013). We used LAST(Kielbasa et al. 2011) to perform alignment of the draft genome against itself. Matches that lie in regions of high-copy repeats annotated by RepeatMasker were filtered. These filtered alignments were then extended using the clasp program. Subsequently, the resulting chained alignments were filtered if less than 1,000 bp. The remaining chains were globally aligned using either stretcher or needle from the European Molecular Biology Open Software Suite (EMBOSS) package(Rice et al. 2000), depending on the chain size (e.g. when the product of the sequence lengths was greater than 100 Mb stretcher was used, otherwise needle was used). Alignments of smaller than 90% identity, or a gap percentage larger than 30% were discarded.

We found a total of 21,843 duplicated fragments spanning 36.28 Mb (1.45%) in the Malayan pangolin genome and 32,992 fragments spanning 57.03 Mb (2.11%) for Chinese pangolin (Supplementary Table 3.6). The percentage of the genome coverage was calculated by assuming the genome size of Malayan pangolin and Chinese pangolin are 2.5Gb and 2.7Gb, respectively.

Supplemental Table S3.6. Summary of segmental duplications in the pangolin genomes. Different cut-off sizes were used to summarize detected segmental duplications in the pangolin genomes.

SD Size	Malayan pangolin			Chinese pangolin		
	Number	Median	Genome Cov (Mb)	Number	Median	Genome Cov (Mb)
> 1KB	21,843	1,215	36.28	32,992	1,477	57.03
> 5KB	439	7,214	7.85	159	10,840	21.37
> 10KB	135	29,014	5.78	88	15,592	16.65
> 50KB	45	73081	3.78	0	0	0

4.0 SPECIATION TIME AND DIVERGENCE TIME

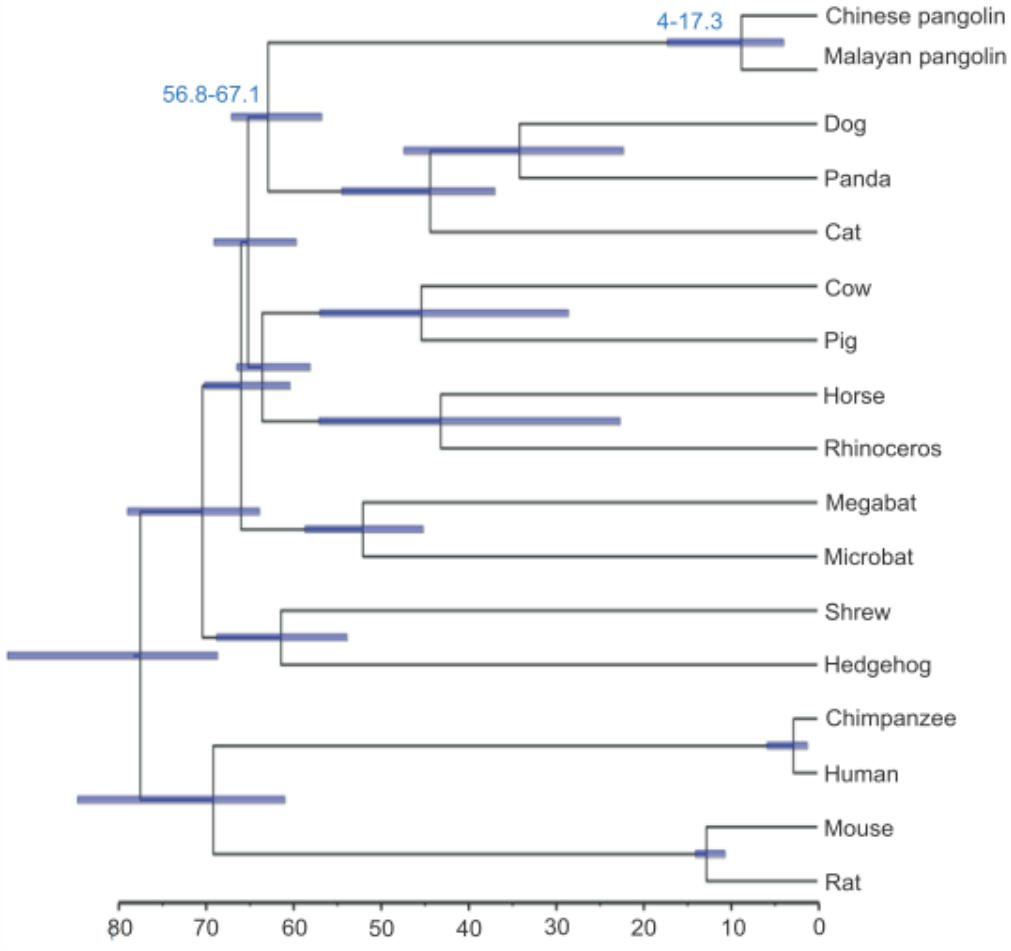
To estimate divergence time of Pholidota from Carnivora lineage, MCMCTREE software as part of PAML4.8. package(Yang 2007) was used. Single copy gene families were used to construct a phylogenetic tree for both pangolin species and other mammalian species. Four-fold degenerate sites were extracted from the alignment of 1423 1:1 orthologs of 17 species resulting in 107,351 sites. Totally 7 reliable calibration points from the Fossil Calibration Database(Ksepka et al. 2015) were used as priors (**Supplemental Table S4.1**).

Two runs of MCMCTREE with 3,000,000 generations and 300,000 burn-in were conducted and convergence of both runs was checked using Tracer 1.5 software(Rambaut A 2014). Tree was visualised using FigTree software (**Supplemental Figure S4.1)**.

Supplemental Table S4.1. Calibration points for the divergence time estimation.

No.	Taxa	Clades	Minimum bound	Maximum bound	Method	Reference
1	Human - mouse	Archonta-Glires	61,5	100,5	biostratigraphy	Benton and Donoghue, 2007
2	Primates+mouse - dog,	Euarchontaglires -Laurasiatheria	61,6	164,6	fossil	Benton et al. 2015; Fossil

	horse, cow					Calibration Database
3	Cat-dog	Carnivora	37,3	66	fossil	Benton et al. 2015; Fossil Calibration Database
4	Megabat-Microbat	Chiroptera	45	58,9	fossil	Phillips, 2015; Fossil Calibration Database
5	Horse+rhino - pig+cow.	Common ancestor of Cetartiodactyla	52,4	66	fossil	Benton et al. 2015; Fossil Calibration Database
6	Mouse-rat	Mus - Rattus	10,4	14	biostratigraphy	Benton et al. 2015; Fossil Calibration Database
7	Hedgehog-Common shrew	Lypotyphla	61,6	164,6	fossil	Benton et al. 2015; Fossil Calibration Database



Supplemental Figure S4.1. Estimation of speciation and divergence time. Chronogram with 95% intervals of posterior divergence time (in millions of years) distribution of 17 mammalian species calibrated with fossil information. Posterior distributions of divergence times (blue numbers on the nodes) of Pholidota lineage and Chinese - Malayan pangolins are shown. Pangolins diverged from their closest relatives, the Carnivora, ~56.8-67.1 million years ago (MYA) (mean=61.9MYA). Malayan and Chinese pangolin species diverged from each other ~4-17.3 MYA (mean=8.84MYA). Nucleotide divergence between pangolins for 4-fold degenerate sites was 0.42 % per million years. Based on this, we calculated substitution/mutation rate, 1.47×10^{-08} with 95% interval (3.2489×10^{-08}-7.5118×10^{-09} for minimum and maximum divergence time).

5.0 HETEROZYGOSITY

To identify Single Nucleotide Polymophisms (SNPs) and indels in the pangolin genomes, we used the Genome Analysis Toolkit (GATK) v. 3.3.0(McKenna et al. 2010). Read mapping was made by BowTie2(Langmead and Salzberg 2012), and duplicated reads were removed using

Picard tools. Raw SNP calling was performed using HaplotypeCaller in the GATK software package. Apart from HaplotypeCaller, the SNVs were also identified using Samtools 1.1(Li et al. 2009). The results from HaplotypeCaller and Samtools were treated independently. The candidate variants were preprocessed based on maximum number of good quality reads of 130 and minimum mapping quality of 25. Regions of repeats were removed from resulting vcf files. The post-filtered variants were considered high quality variants and only the high quality variants that present from both pipelines were used as final list.

All variants were annotated using SnpEff 4.1 effect prediction tool and database constructed from the annotated genes of both pangolin species(Cingolani et al. 2012). For Malayan pangolin, we found 76.2%, of these variants were located outside of the protein-coding genes, 0.8% inside exons and 14.3% inside introns, and for Chinese pangolin, these numbers were 77.8%, 1.5% and 16.6%, respectively (**Supplemental Table S5.1-5.3)**.

Supplemental Table S5.1. Summary of variant calling in both pangolins.

	Total number of variants	**Effective genome size**	**Variant rate**
Malayan pangolin	3,762,073	2,423,229,922	1.55×10^{-3}
Chinese pangolin	807,706	2,008,608,506	0.4×10^{-3}

Supplemental Table S5.2. Variant effects grouped by impact. High, Low, Moderate and Modifier effects relate to potential impact on gene function.

Variant impact	**Malayan pangolin**		**Chinese pangolin**	
	Count	Percent	Count	Percent
High	657	0.02%	354	0.04%
Low	21,249	0.53%	6,229	0.72%
Moderate	16,229	0.4%	7,999	0.92%
Modifier	3,999,556	99.06%	851,869	98.32%

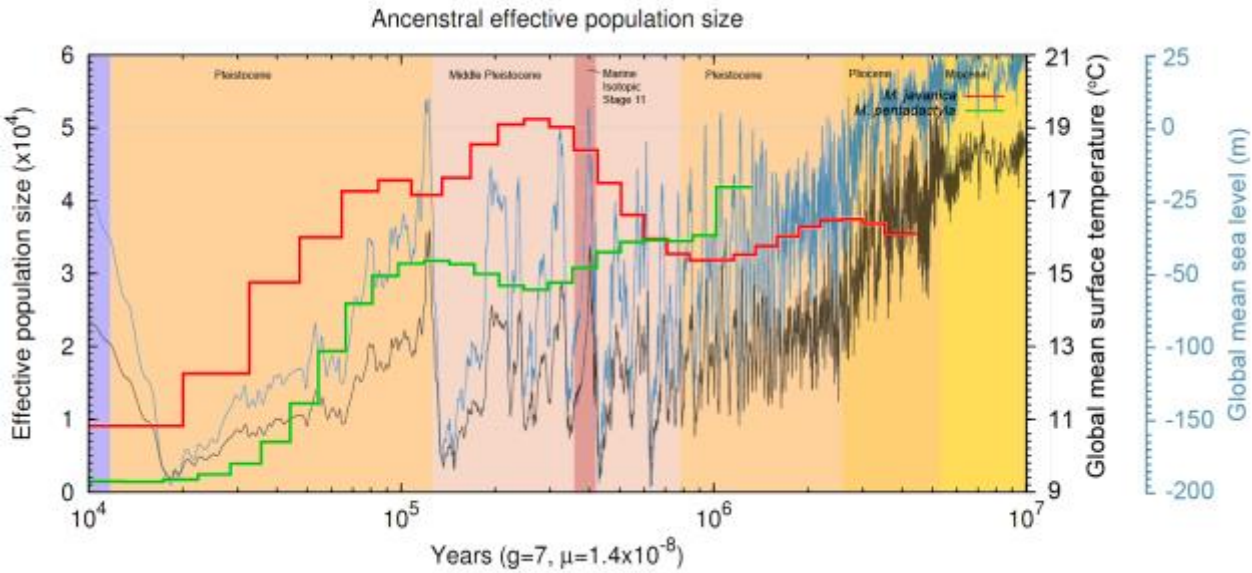
Supplemental Table S5.3. Variant effect grouped by regions.

Variant region	Malayan pangolin		Chinese pangolin	
	Count	Percent	Count	Percent
Downstream	160,033	3.96%	37,144	4.29%
Exon	33,823	0.84%	13,483	1.56%
Intergenic	3,077,204	76.22%	674,720	77.87%
Intron	578,977	14.34%	100,657	11.62%
Splice site acceptor	108	0.003%	45	0.005%
Splice site donor	169	0.004%	69	0.008%
Splice site region	2,986	0.07%	761	0.08%
Upstream	155,706	3.85%	34,611	3.99%

6.0 PANGOLIN POPULATION HISTORY ESTIMATION

Utilizing the Pairwise Sequential Markovian Coalescent (PSMC) model, we inferred the historical effective population size (N_e) over time for both the pangolin genomes (**Supplemental Figure S6.1)**. In order to scale N_e, we used generation time, g of 7 years and per site per generation mutation rate of 1.4×10^{-8} calculated based on neutral theory as per computed phylogenetic tree and divergence time in Supplementary Figure 4.1. The ancestral Chinese pangolin coalesced at least around 1 to 2 million years ago (MYA) and showed inverted population size trends to Malayan pangolin until 100 thousand years ago (KYA) after which there was a dramatic decline in both the species. The Marine Isotopic Stage 11 (420-360 KYA) is known as one of the warmest interglacial event in the past 500 KYA, and pangolin N_e were approaching the maxima and minima for Malayan pangolin and Chinese pangolin, respectively during this period in Middle Pleistocene (728-126 KYA). The paleoclimatic events during Middle Pleistocene particularly warming and sea-level rising above current levels could have provided better genetic fitness for Malayan pangolin and thus we observed the rise in their population size compared to Chinese pangolin. It is also interesting to note that the decline in

pangolin N_e from 100 KYA to 10 KYA coincides with the Late Pleistocene extinction events of other land mammals in other parts of the world(Martin 1989).



Supplemental Figure S6.1. Estimated population size history for both Malayan and Chinese pangolins. The X-axis represents time in years. The first y-axis on the left shows the effective population size scaled to 4μN. On the right, the two colour coded y-axis shows the global mean surface temperature (black) and sea-level data (blue). The global mean sea level is expressed in relation to current sea-level. The paleoclimatic data was obtained from a previous study by Hansen and co-workers (2013).

7.0 VALIDATION OF PSEUDOGENIZED GENES

To validate mutations found in pseudogenized genes, we examined the mutated regions of these genes using PCR and Sanger sequencing. For each gene, we performed the Sanger sequencing using DNA extracted from seven unrelated adult Malayan pangolins and one fetus from one of the adult pangolins that we used in this validation. Primer sequences used in the validation are shown in **Supplemental Table S7.1**. Two sets of primers were used to examine two different mutated regions in the *ENAM* gene. The first primer pair (5'-AGCAAGTTCAAAGGGTTTCTCAGC-3' and 3'- TTCAGCTTGTTCATCAGAATTTGG-5') and second primer pair (5'-CCTTATTTCAGTAACTCCCAAGCT-3' and 3'-CTTGCATTCTTAGTCTGGGTATCTT-5') were used to validate the frameshift mutation and premature stop codon found in the *ENAM* gene. *FBFSP2* amplified with forward primer (5'-GAGTGCCCAGAGTCTATGTAGGGATGG-3') and reverse primer (3'-

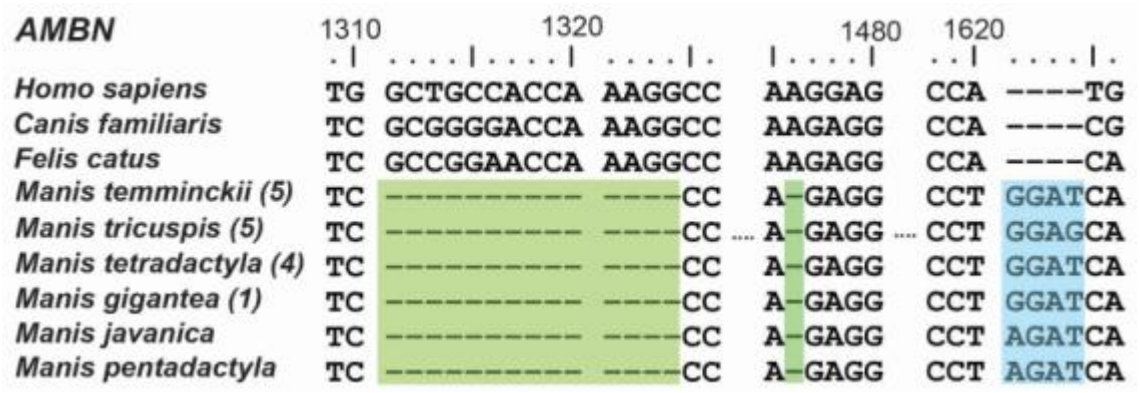
TTCCTGCTCATCGTCCTTCCCAGAG-5') targeting 437bps of the mutation site. While *GUCA1C* was amplified with forward primer (5'-CAGCTGTAAGAGATTTGGAGTAGC-3') and reverse primer (3'-GTTTAATGACTCACTTACCTACAAGC-5') targeting 476bps of the mutation site. The *IFNE* gene was amplified using forward primer (5'-GAGGAAATGTCCCATGAACACTAGG-3') and reverse primer (5'-AGTTCTCTCTCCCATCCACCTACCC-3'). All the primers were used for PCR using the following described protocol. The total reaction volume of 25 µL contained 60ng purified gDNA, 0.3 pmol of each primer, deoxynucleotides triphosphates (dNTP, 400 µM each), 0.5 U Taq DNA polymerase and supplied buffer were used. The PCR was performed as follow: 1 cycle (94 ^{0}C for 2 minutes) for initial denaturation; 30 cycles (98 ^{0}C for 10 sec; 65 ^{0}C for 30 sec; 68 ^{0}C for 30 sec) for annealing and extension for DNA amplification. The PCR products were purified by standard methods and directly sequenced with the same primers using BigDye© Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems).

For the Sanger validation of the genes in African species, we only can use the same sets of primers designed based on the genome of Malayan pangolin since we do not have the genome sequences of African species. Therefore, experimental validation did not work for some genes or genomic regions due to the differences between the sequences of Asian and African species. However, we only showed the results for the successfully validated genes and genomic regions in this section.

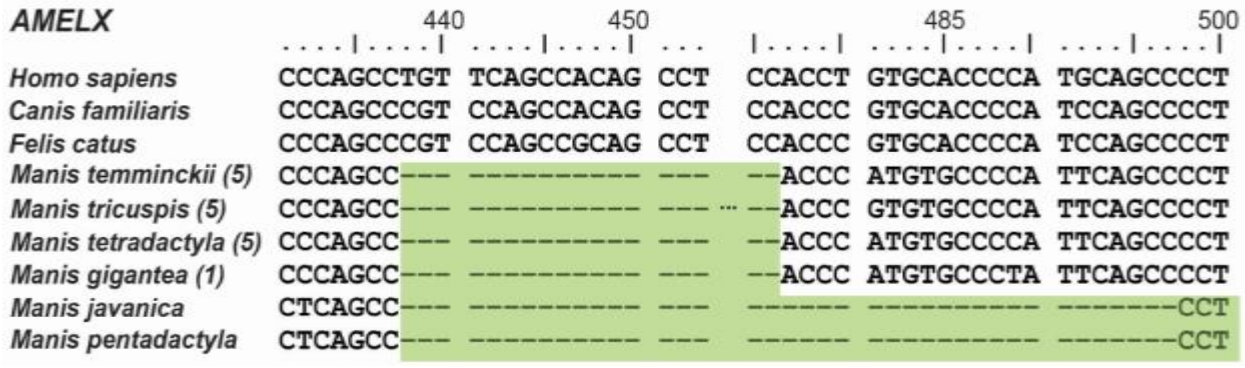
Supplemental Table S7.1. **Primer sequences used for the validation of gene mutations.** The mutations were validated using Sanger sequencing.

Gene	Primers	Primer Length (bp)	Annealing Temperature
ENAM_1	5'-AGCAAGTTCAAAGGGTTTCTCAGC-3' 3'-TTCAGCTTGTTCATCAGAATTTGG-5'	464	62 °C
ENAM_2	5'- CCTTATTTCAGTAACTCCCAAGCT-3' 3'-CTTGCATTCTTAGTCTGGGTATCTT-5'	348	62 °C
AMELX	5' CAAAGAAAGCATTGCTACTTCTCC 3' 3'-GGGTCTAGAGTTTCAGTAACCAGAG-5'	552	62 °C
AMBN	5'-GTCTCTCACTTTGTTACGGTTTCT-3' 3'-CAAGTGTTTGGAAATAATATAGGTCCC-5'	792	62 °C
BFSP2	5'-GAGTGCCCAGAGTCTATGTAGGGATGG-3'	437	68 °C

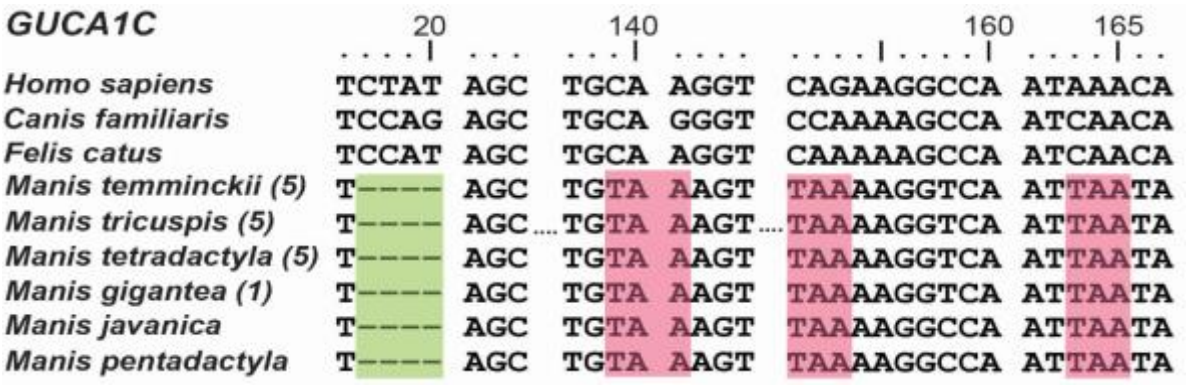
	3'-TTCCTGCTCATCGTCCTTCCCAGAG-5'		
GUCA1C	5'-CAGCTGTAAGAGATTTGGAGTAGC-3' 3'-GTTTAATGACTCACTTACCTACAAGC-5'	476	62 °C
IFNE	5’-GAGGAAATGTCCCATGAACACTAGG-3’ 3’-GGGTAGGTGGATGGGAGAGAGAACT-5’	878	65 °C
KRT75	5’-GAGCATCTAGGGCACAATGC -3’ 3’-GCATTCCCTGAGCTGACAGA -5’	370	66 °C
KRT36_1	5’-CAGTTAGCATGGGTAGCAGGTG -3’ 3’-CTCTGCTTCAACCTCCCTGAG -5’	399	63 °C
KRT36_2	5’-AGAGATGCGGTGTCAGTATG -3’ 3’-CATCCAGAAAGAAAGAGGAGAG -5’	354	60 °C



Supplemental Figure S7.1. Sanger validation of *AMBN* genes in African pangolin species. The numbers in the brackets indicated the number of samples we examined.



Supplemental Figure S7.2. Sanger validation of *AMELX* genes in African pangolin species. The numbers in the brackets indicated the number of samples we examined.

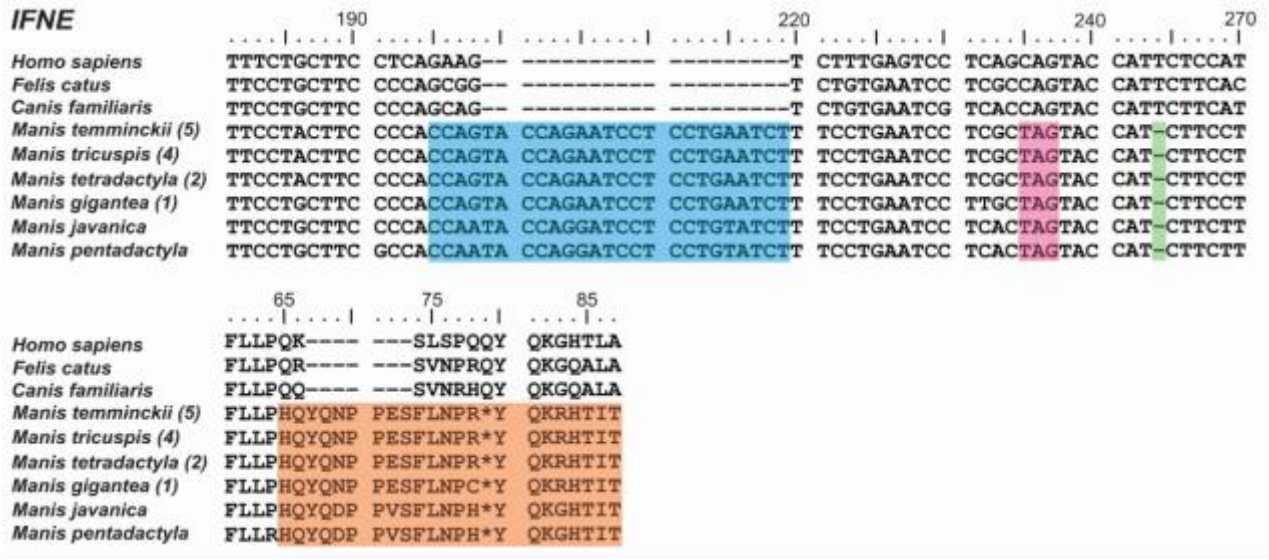


506

507 **Supplemental Figure S7.3. Sanger validation of *GUCA1C* genes in African pangolin species.** The
508 numbers in the brackets indicated the number of samples we examined.

509

510



511

512 **Supplemental Figure S7.4. Sanger validation of *IFNE* genes in African pangolin species.** (Top) *IFNE*
513 nucleotide sequences. (Bottom) *IFNE* protein sequences. The numbers in the brackets indicated the
514 number of samples we examined.

515

516

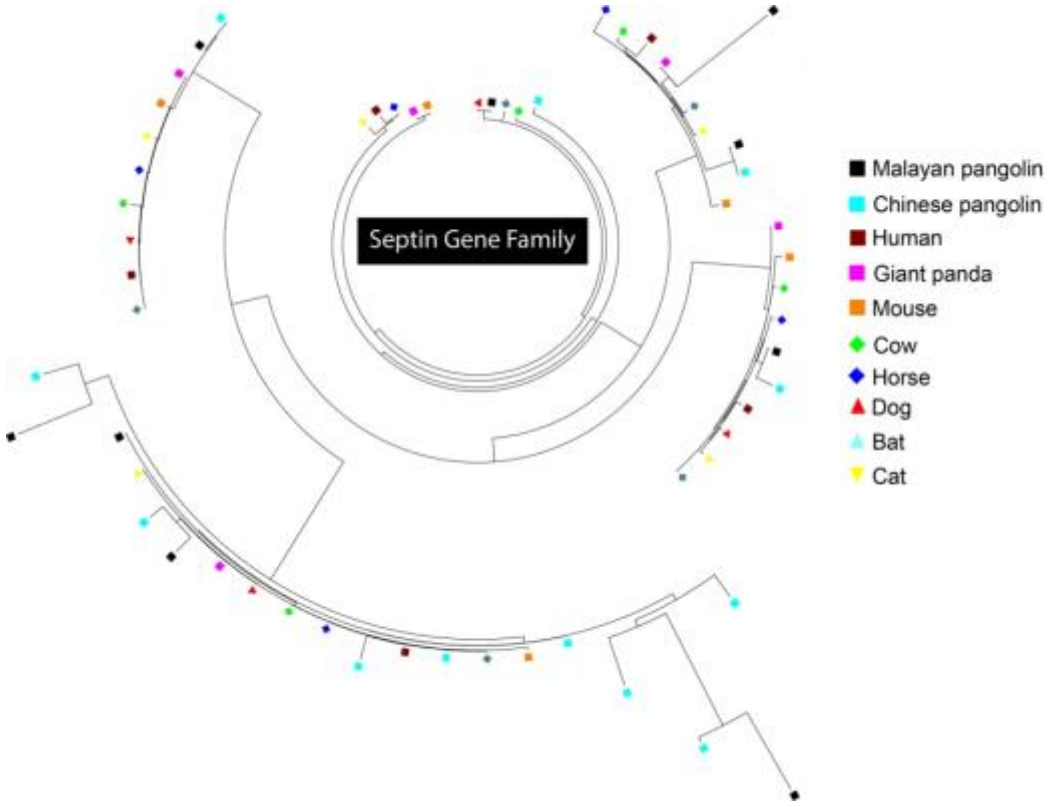
Supplemental Figure S7.5. The keratin-related proteins in African pangolin species. (Left) KRT36 protein sequence alignment. At least two critical amino acid changes that we observed in the Malayan and Chinese pangolins were identical in three African species that we examined. In *M. temminckii*, the C->D amino acid change likely affect the biological function of KRT36 (score=-8.461) as predicted by PROVEAN (Choi and Chan 2015). (Right) Two critical amino acid changes that we observed in the Malayan and Chinese pangolins were identical in *M. tricuspis* and *M. temminckii*, but not *M. tetradactyla*. In *M. tetradactyla*, the changes of V->T and M->deletion are also likely to affect the function of KRT75 as predicted by PROVEAN with scores of -3.459 and -12.790, respectively. No data available for *M. gigantea*.

8.0 GENE FAMILY EXPANSION AND CONTRACTION ANALYSIS

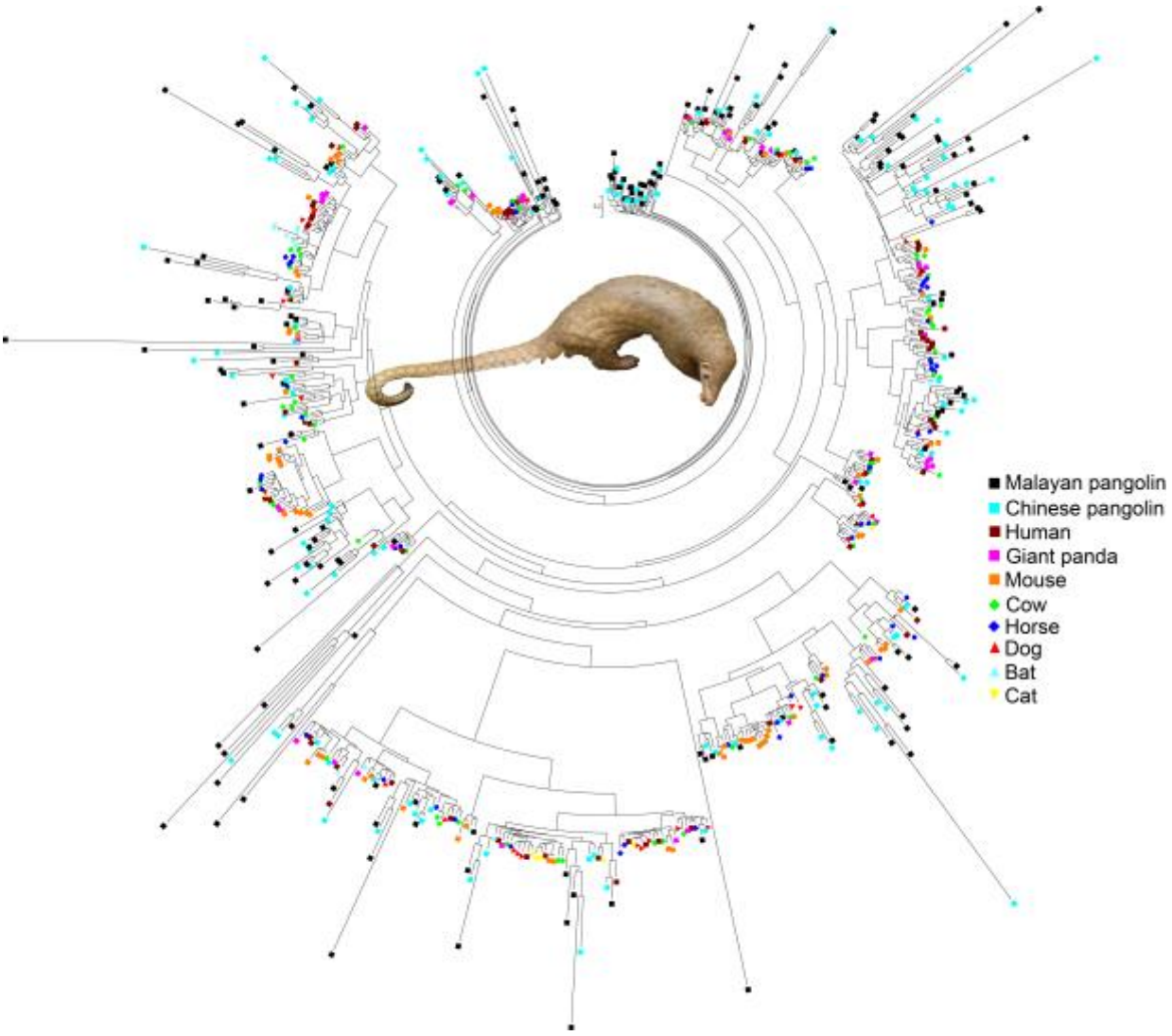
Correct assignment of genes to families is highly important for this analysis. Firstly we performed test for 40 mammalian species present in TreeFam database(Li et al. 2006). Proteins of this species assigned to families annotated in TreeFam database were extracted and mapped to hmm profiles of known families present in this database (http://www.treefam.org/static/download/treefam9.hmm3.tar.gz) by HMMER3 (hmmscan with default parameters)(Johnson et al. 2010). Each protein was assigned to top-hit family if it was significant. Surprisingly, twenty-one proteins were not mapped to any family, and six more proteins showed only not significant hits. 8,885 of 10,306 (86.2%) families were correctly assembled for all species, and 9,213 (89.4%) families were assembled with errors in less than 5% species. This set of families was used in following expansion/contraction analysis. Assignment pangolin genes to families was performed in the same manner. Expansion and contraction analysis of gene families was performed by CAFE v3.1(De Bie et al. 2006). Eight reference species were used: *Ailuropoda melanoleuca* (giant panda), *Bos taurus* (cow), *Canis familiaris* (dog), *Equus caballus* (horse), *Felis catus* (cat), *Homo sapiens* (human), *Mus musculus* (mouse), *Pteropus vampyrus* (megabat). One of CAFE assumptions is that MRCA of all analyzed species had at least one protein in each family. Because of this reason we removed families present in less than 3 species, therefore 8,438 families were retained in final dataset.

GO term enrichment analysis of all significantly 147 expanded gene families in pangolins was performed using Blast2GO (**Supplemental Figure S8.3)**(Conesa et al. 2005). Among the significantly enriched GO terms are translation (174 genes;GO:0006412;FDR p-value=1.13x10^{-}

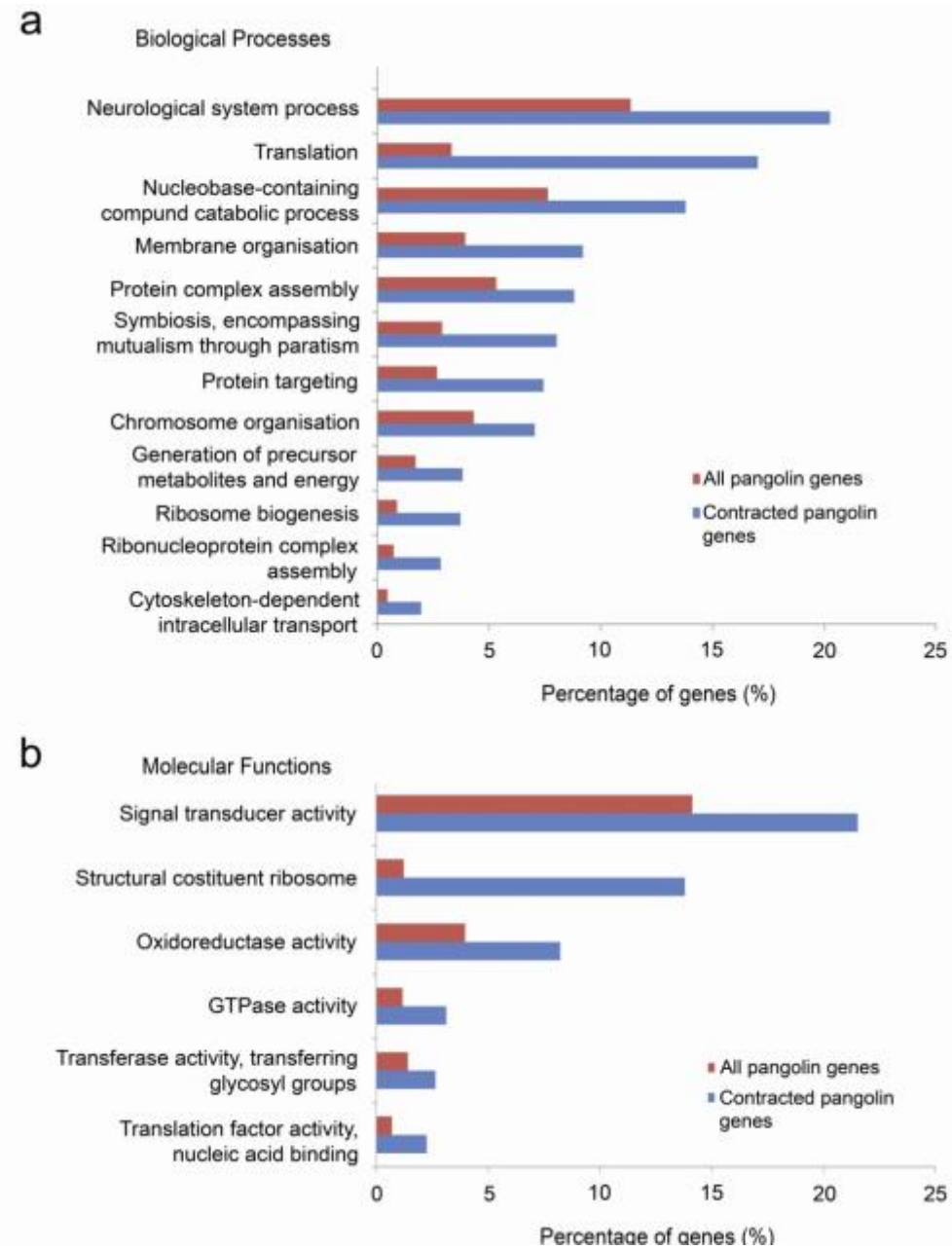
548 [60]) and ribosome biogenesis (38 genes;GO:0042254;FDR p-value=1.02x10^{-10}). The significant
549 expansion of ribosome-associated gene families may be an advantageous selection for global up-
550 regulation of protein synthesis.
551
552



553
554 **Supplemental Figure S8.1. Expansion of the Septin gene family in the pangolin lineage.** Septin was
555 expanded in the Malayan pangolin (9 genes; black square) and Chinese pangolin (12 genes; light blue
556 square).



Supplemental Figure S8.2. Expansion of the olfactory gene (OR) gene family 2 in the pangolin lineage. The protein sequences of OR family 2 used in this analysis were derived from expanded orthologous gene families in the pangolin lineage. The OR genes in the Malayan pangolin and Chinese pangolin are represented by black square and light blue square, respectively. The olfactory receptor genes in both pangolin species were significantly expanded compared to other mammalian species including human, giant panda and mouse.



Supplemental Figure S8.3. Functional enrichment of 147 significantly expanded gene families. Only Gene Ontology (GO) terms with a significant enrichment (Fisher Exact Test; FDR p-value<0.05) are shown. (a) Significantly over-represented biological processes. (b) Significantly over-represented molecular functions.

9.0 POSITIVE SELECTION ANALYSIS

To find the signatures of positive selection in the pangolin lineage, we used 10 mammal species: horse, dog, cat, panda, cow, mouse, human, megabat and the two pangolin species. In total, we

found 8,498 1:1 orthologs between the species using Poff/proteinorhto tool(Lechner et al. 2014). Orthologs were aligned using Prank F+ codon model and filtered out using Gblocks software(Castresana 2000). The final set consisted of 8,250 genes. Codeml software as part of PAML 4.8 package(Yang 2007) was used to test branch-site model of positive selection for pangolin lineage. Totally 8250 1:1 orthologs were tested. For each ortholog alignment codon model of molecular evolution was fitted with pangolin lineage as foreground branch and other lineages as background branches. Null model of neutral evolution of foreground branch with fixed omega = 1 (model =2, NSsites=2) was tested using Likelihood-Ratio test (with $p < 0.05$ significance level) against alternative model with estimated omega value (model =2, NSsites=2). There are 427 genes identified under positive selection in the pangolin lineage (Supplementary Table 9.1).

Supplemental Table S9.1. List of genes under positive selection. Totally 427 genes with raw p-values <0.05 according to standard Likelihood Ratio Tests.

ENSEMBL transcript ID	Chinese pangolin gene ID	Malayan pangolin gene ID	Gene symbol	Description	p-value
ENST00000612742	MCP0013542	MMP0007444	*METTL8*	methyltransferase like 8	4.53E-19
ENST00000396625	MCP0014112	MMP0015713	*COL4A4*	collagen, type IV, alpha 4	8.76E-19
ENST00000424952	MCP0014989	MMP0001974	*ZDHHC3*	zinc finger, DHH C-type containing 3	5.96E-11
ENST00000620295	MCP0011698	MMP0006416	*KIF1B*	kinesin family member 1B	1.58E-10
ENST00000404158	MCP0013765	MMP0006992	*GATAD2A*	GATA zinc finger domain containing 2A	2.43E-10
ENST00000262134	MCP0002597	MMP0013384	*LPCAT2*	lysophosphatidylcholine acyltransferase 2	6.84E-10
ENST00000438552	MCP0013319	MMP0005369	*SNRPB*	small nuclear ribonucleoprotein polypeptides B and B1	8.28E-10
ENST00000319327	MCP0005590	MMP0004890	*SERINC4*	serine incorporator 4	9.61E-09
ENST00000465301	MCP0003852	MMP0011385	*RGAG1*	retrotransposon gag domain containing 1	2.11E-07
ENST00000256997	MCP0003917	MMP0002374	*ACP2*	acid phosphatase 2, lysosomal	4.65E-07
ENST00000170150	MCP0006225	MMP0004578	*BPIFB2*	BPI fold containing family B, member 2	5.64E-07
ENST00000262288	MCP0018898	MMP0015215	*SCPEP1*	serine carboxypeptidase 1	6.44E-07
ENST00000399518	MCP0004479	MMP0016031	*PLA2G4E*	phospholipase A2, group IVE	8.73E-07
ENST00000417439	MCP0006067	MMP0003610	*LTF*	lactotransferrin	1.80E-06
ENST00000220812	MCP0008149	MMP0010347	*DKK4*	dickkopf WNT signaling pathway inhibitor 4	2.16E-06
ENST00000311772	MCP0015226	MMP0017399	*PPP1R8*	protein phosphatase 1, regulatory subunit	2.67E-06

				8	
ENST00000556816	MCP0000554	MMP0002391	*ISCA2*	iron-sulfur cluster assembly 2	4.03E-06
ENST00000297268	MCP0005262	MMP0004567	*COL1A2*	collagen, type I, alpha 2	7.80E-06
ENST00000218867	MCP0010675	MMP0004325	*SGCG*	sarcoglycan, gamma (35kDa dystrophin-associated glycoprotein)	7.88E-06
ENST00000350051	MCP0008180	MMP0019042	*BIRC5*	baculoviral IAP repeat containing 5	1.14E-05
ENST00000307921	MCP0013631	MMP0009333	*ADAT1*	adenosine deaminase, tRNA-specific 1	1.25E-05
ENST00000252245	MCP0002974	MMP0022581	*KRT75*	keratin 75, type II	1.33E-05
ENST00000039989	MCP0006888	MMP0011614	*TTC17*	tetratricopeptide repeat domain 17	1.50E-05
ENST00000399080	MCP0016897	MMP0011951	*RAD51AP2*	RAD51 associated protein 2	1.62E-05
ENST00000372754	MCP0009233	MMP0007886	*MATN4*	matrilin 4	1.93E-05
ENST00000435159	MCP0011655	MMP0001856	*TMEM132C*	transmembrane protein 132C	2.75E-05
ENST00000430500	MCP0007111	MMP0011114	*SLC22A8*	solute carrier family 22 (organic anion transporter), member 8	2.82E-05
ENST00000330794	MCP0003498	MMP0002794	*TMEM173*	transmembrane protein 173	2.92E-05
ENST00000302555	MCP0003040	MMP0003207	*GP2*	glycoprotein 2 (zymogen granule membrane)	3.36E-05
ENST00000230568	MCP0016099	MMP0019937	*LY86*	lymphocyte antigen 86	3.48E-05
ENST00000229243	MCP0001799	MMP0002450	*ACRBP*	acrosin binding protein	4.09E-05
ENST00000318602	MCP0015520	MMP0022433	*A2M*	alpha-2-macroglobulin	4.39E-05
ENST00000355661	MCP0002865	MMP0018609	*PLEKHA7*	pleckstrin homology domain containing, family A member 7	4.65E-05
ENST00000217169	MCP0011196	MMP0023102	*BIRC7*	baculoviral IAP repeat containing 7	4.77E-05
ENST00000398189	MCP0011417	MMP0007405	*APOF*	apolipoprotein F	4.90E-05
ENST00000367843	MCP0012903	MMP0002856	*DCAF6*	DDB1 and CUL4 associated factor 6	5.00E-05
ENST00000407315	MCP0003536	MMP0017719	*THAP4*	THAP domain containing 4	5.27E-05
ENST00000527372	MCP0019410	MMP0019410	*MYO18A*	myosin XVIIIA	6.33E-05
ENST00000305949	MCP0004057	MMP0021541	*MTBP*	MDM2 binding protein	6.84E-05
ENST00000621805	MCP0006948	MMP0015704	*RAB18*	RAB18, member RAS oncogene family	7.88E-05
ENST00000265081	MCP0014405	MMP0008354	*MSH3*	mutS homolog 3	9.12E-05
ENST00000377122	MCP0012659	MMP0012306	*NEBL*	nebulette	0.0001
ENST00000298694	MCP0010292	MMP0010258	*ARHGEF40*	Rho guanine nucleotide exchange factor (GEF) 40	0.0001
ENST00000286428	MCP0002071	MMP0019114	*VBP1*	von Hippel-Lindau binding protein 1	0.0001
ENST00000367975	MCP0011264	MMP0021735	*SDHC*	succinate dehydrogenase complex, subunit C, integral membrane protein, 15kDa	0.0001
ENST00000260810	MCP0016712	MMP0022386	*TOPBP1*	topoisomerase (DNA) II binding protein 1	0.0001
ENST00000306058	MCP0016146	MMP0001730	*ASXL1*	additional sex combs like transcriptional	0.0001

				regulator 1	
ENST00000394518	MCP0006774	MMP0016921	*AKAP13*	A kinase (PRKA) anchor protein 13	0.0002
ENST00000394931	MCP0015501	MMP0006535	*BMPR1B*	bone morphogenetic protein receptor, type IB	0.0002
ENST00000598162	MCP0018058	MMP0004019	*BCAT2*	branched chain amino-acid transaminase 2, mitochondrial	0.0002
ENST00000535274	MCP0012621	MMP0021425	*B3GNT4*	UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 4	0.0002
ENST00000380010	MCP0012645	MMP0004724	*SLC19A1*	solute carrier family 19 (folate transporter), member 1	0.0002
ENST00000389758	MCP0017570	MMP0014493	*MROH2A*	maestro heat-like repeat family member 2A	0.0002
ENST00000360351	MCP0018405	MMP0012150	*MAP2*	microtubule-associated protein 2	0.0002
ENST00000332578	MCP0008778	MMP0014322	*TMPRSS9*	transmembrane protease, serine 9	0.0002
ENST00000261652	MCP0001724	MMP0007814	*TNFRSF13B*	tumor necrosis factor receptor superfamily, member 13B	0.0002
ENST00000532211	MCP0002130	MMP0006949	*PIH1D2*	PIH1 domain containing 2	0.0003
ENST00000493960	MCP0006786	MMP0016372	*FAM208A*	family with sequence similarity 208, member A	0.0003
ENST00000265992	MCP0019338	MMP0001436	*CCNJ*	cyclin J	0.0003
ENST00000400897	MCP0009359	MMP0017922	*MASP2*	mannan-binding lectin serine peptidase 2	0.0003
ENST00000371253	MCP0001163	MMP0015162	*ADGRF1*	adhesion G protein-coupled receptor F1	0.0004
ENST00000404938	MCP0020085	MMP0019810	*ABCB5*	ATP-binding cassette, sub-family B (MDR/TAP), member 5	0.0005
ENST00000361373	MCP0019329	MMP0001401	*LDB3*	LIM domain binding 3	0.0005
ENST00000338005	MCP0000142	MMP0008507	*CCDC183*	coiled-coil domain containing 183	0.0005
ENST00000390683	MCP0006516	MMP0006681	*RAD51B*	RAD51 paralog B	0.0005
ENST00000370378	MCP0007955	MMP0008286	*KIAA1107*	KIAA1107	0.0005
ENST00000222286	MCP0001801	MMP0002446	*GAPDHS*	glyceraldehyde-3-phosphate dehydrogenase, spermatogenic	0.0006
ENST00000295500	MCP0018223	MMP0014289	*GPR155*	G protein-coupled receptor 155	0.0006
ENST00000357590	MCP0018531	MMP0009444	*SLC37A4*	solute carrier family 37 (glucose-6-phosphate transporter), member 4	0.0006
ENST00000270115	MCP0012832	MMP0023303	*LRRC56*	leucine rich repeat containing 56	0.0007
ENST00000436874	MCP0004214	MMP0002213	*VEZT*	vezatin, adherens junctions transmembrane protein	0.0007
ENST00000353364	MCP0018742	MMP0014729	*USP15*	ubiquitin specific peptidase 15	0.0007
ENST00000486257	MCP0011128	MMP0017391	*TIGIT*	T cell immunoreceptor with Ig and ITIM domains	0.0007
ENST00000371421	MCP0011485	MMP0018057	*ARRDC1*	arrestin domain containing 1	0.0008

ENST00000374075	MCP0008783	MMP0003826	*AKNA*	AT-hook transcription factor	0.0008
ENST00000381527	MCP0000833	MMP0021053	*CDK8*	cyclin-dependent kinase 8	0.0008
ENST00000529629	MCP0017036	MMP0011201	*CAPN5*	calpain 5	0.0008
ENST00000301391	MCP0015374	MMP0013367	*CYB5D2*	cytochrome b5 domain containing 2	0.0009
ENST00000325888	MCP0015344	MMP0001388	*FLNC*	filamin C, gamma	0.0010
ENST00000346736	MCP0004001	MMP0016910	*C19orf57*	chromosome 19 open reading frame 57	0.0010
ENST00000302787	MCP0003872	MMP0005668	*LETM1*	leucine zipper-EF-hand containing transmembrane protein 1	0.0010
ENST00000360299	MCP0013647	MMP0008458	*RAB5B*	RAB5B, member RAS oncogene family	0.0010
ENST00000311601	MCP0018930	MMP0009659	*SH3PXD2B*	SH3 and PX domains 2B	0.0010
ENST00000414749	MCP0010652	MMP0012397	*MLXIPL*	MLX interacting protein-like	0.0010
ENST00000427836	MCP0007865	MMP0007255	*PLEKHM3*	pleckstrin homology domain containing, family M, member 3	0.0011
ENST00000314583	MCP0001012	MMP0006283	*HCLS1*	hematopoietic cell-specific Lyn substrate 1	0.0012
ENST00000251643	MCP0012401	MMP0022286	*KRT12*	keratin 12, type I	0.0012
ENST00000261847	MCP0000241	MMP0014830	*SECISBP2L*	SECIS binding protein 2-like	0.0012
ENST00000373095	MCP0001404	MMP0017002	*FAM102A*	family with sequence similarity 102, member A	0.0012
ENST00000257974	MCP0013193	MMP0017552	*KRT82*	keratin 82, type II	0.0012
ENST00000250018	MCP0010835	MMP0010087	*TPH1*	tryptophan hydroxylase 1	0.0013
ENST00000483725	MCP0006513	MMP0017318	*KIAA0408*	KIAA0408	0.0014
ENST00000394479	MCP0003938	MMP0014594	*REL*	v-rel avian reticuloendotheliosis viral oncogene homolog	0.0014
ENST00000289359	MCP0005854	MMP0008461	*MITD1*	MIT, microtubule interacting and transport, domain containing 1	0.0014
ENST00000375077	MCP0017522	MMP0015837	*CORO2A*	coronin, actin binding protein, 2A	0.0015
ENST00000249776	MCP0002618	MMP0013567	*KNSTRN*	kinetochore-localized astrin/SPAG5 binding protein	0.0015
ENST00000371605	MCP0017168	MMP0005817	*ABCA2*	ATP-binding cassette, sub-family A (ABC1), member 2	0.0016
ENST00000324366	MCP0017142	MMP0001601	*PRMT5*	protein arginine methyltransferase 5	0.0016
ENST00000372236	MCP0004577	MMP0018467	*POLH*	polymerase (DNA directed), eta	0.0016
ENST00000228887	MCP0017874	MMP0014991	*GPRC5D*	G protein-coupled receptor, class C, group 5, member D	0.0017
ENST00000294618	MCP0018019	MMP0003901	*DOCK6*	dedicator of cytokinesis 6	0.0018
ENST00000269025	MCP0000730	MMP0007037	*LRRC46*	leucine rich repeat containing 46	0.0020
ENST00000373368	MCP0006150	MMP0004180	*SPP2*	secreted phosphoprotein 2, 24kDa	0.0020
ENST00000391588	MCP0012404	MMP0014927	*KRTAP3-1*	keratin associated protein 3-1	0.0022
ENST00000549091	MCP0003001	MMP0006493	*WDR90*	WD repeat domain 90	0.0023

ENST00000381821	MCP0002417	MMP0012852	*TEX33*	testis expressed 33	0.0023
ENST00000295897	MCP0000306	MMP0019724	*ALB*	albumin	0.0023
ENST00000361756	MCP0018455	MMP0017079	*RNF121*	ring finger protein 121	0.0023
ENST00000264661	MCP0012356	MMP0006932	*KCNH4*	potassium channel, voltage gated eag related subfamily H, member 4	0.0024
ENST00000278742	MCP0016769	MMP0018597	*ST14*	suppression of tumorigenicity 14 (colon carcinoma)	0.0026
ENST00000355898	MCP0015864	MMP0018417	*ZNF507*	zinc finger protein 507	0.0027
ENST00000251481	MCP0018685	MMP0005848	*SULT1C2*	sulfotransferase family, cytosolic, 1C, member 2	0.0027
ENST00000265838	MCP0006597	MMP0020696	*ACAT1*	acetyl-CoA acetyltransferase 1	0.0028
ENST00000328278	MCP0018196	MMP0011129	*LRRC14B*	leucine rich repeat containing 14B	0.0028
ENST00000611114	MCP0007699	MMP0004874	*ZNF804B*	zinc finger protein 804B	0.0028
ENST00000431016	MCP0013950	MMP0008195	*PCYT1A*	phosphate cytidylyltransferase 1, choline, alpha	0.0029
ENST00000616727	MCP0009477	MMP0020481	*MUC13*	mucin 13, cell surface associated	0.0029
ENST00000268766	MCP0019423	MMP0018895	*NEK8*	NIMA-related kinase 8	0.0029
ENST00000306534	MCP0018265	MMP0009051	*ROBO4*	roundabout, axon guidance receptor, homolog 4 (Drosophila)	0.0030
ENST00000295408	MCP0007017	MMP0010239	*MERTK*	MER proto-oncogene, tyrosine kinase	0.0030
ENST00000394810	MCP0005684	MMP0005627	*SYNPO2L*	synaptopodin 2-like	0.0032
ENST00000324444	MCP0013626	MMP0013848	*SYNE4*	spectrin repeat containing, nuclear envelope family member 4	0.0032
ENST00000431282	MCP0007907	MMP0014887	*APOBR*	apolipoprotein B receptor	0.0032
ENST00000543976	MCP0014448	MMP0022930	*TMF1*	TATA element modulatory factor 1	0.0033
ENST00000215886	MCP0018753	MMP0012866	*LGALS2*	lectin, galactoside-binding, soluble, 2	0.0033
ENST00000506113	MCP0010371	MMP0021546	*ABLIM3*	actin binding LIM protein family, member 3	0.0033
ENST00000354666	MCP0017215	MMP0008867	*ELOVL2*	ELOVL fatty acid elongase 2	0.0033
ENST00000354042	MCP0003926	MMP0004701	*SLC13A4*	solute carrier family 13 (sodium/sulfate symporter), member 4	0.0034
ENST00000356575	MCP0016969	MMP0011779	*MEGF6*	multiple EGF-like-domains 6	0.0034
ENST00000244709	MCP0009903	MMP0013921	*TREM1*	triggering receptor expressed on myeloid cells 1	0.0034
ENST00000314400	MCP0015810	MMP0019471	*C3orf17*	chromosome 3 open reading frame 17	0.0035
ENST00000305544	MCP0004733	MMP0013542	*LAMB2*	laminin, beta 2 (laminin S)	0.0035
ENST00000260563	MCP0018133	MMP0011335	*RTCA*	RNA 3'-terminal phosphate cyclase	0.0035
ENST00000322344	MCP0004077	MMP0013763	*PNKP*	polynucleotide kinase 3'-phosphatase	0.0036
ENST00000313961	MCP0014297	MMP0017195	*RGS5*	regulator of G-protein signaling 5	0.0037
ENST00000328963	MCP0019554	MMP0016407	*P2RX7*	purinergic receptor P2X, ligand gated ion	0.0038

				channel, 7	
ENST00000439706	MCP0017942	MMP0007807	*SLC38A1*	solute carrier family 38, member 1	0.0038
ENST00000452135	MCP0016083	MMP0006968	*MAPK9*	mitogen-activated protein kinase 9	0.0038
ENST00000324894	MCP0003934	MMP0001358	*GTPBP3*	GTP binding protein 3 (mitochondrial)	0.0040
ENST00000226021	MCP0000027	MMP0021116	*CACNG1*	calcium channel, voltage-dependent, gamma subunit 1	0.0041
ENST00000380752	MCP0019069	MMP0003494	*SLC7A1*	solute carrier family 7 (cationic amino acid transporter, y+ system), member 1	0.0043
ENST00000342203	MCP0000920	MMP0003716	*SSX2IP*	synovial sarcoma, X breakpoint 2 interacting protein	0.0043
ENST00000263681	MCP0001224	MMP0016227	*POLD3*	polymerase (DNA-directed), delta 3, accessory subunit	0.0043
ENST00000539925	MCP0003183	MMP0002442	*LTBR*	lymphotoxin beta receptor (TNFR superfamily, member 3)	0.0045
ENST00000343629	MCP0010938	MMP0000384	*TLDC1*	TBC/LysM-associated domain containing 1	0.0045
ENST00000415318	MCP0018547	MMP0009442	*CCDC153*	coiled-coil domain containing 153	0.0045
ENST00000394511	MCP0011838	MMP0000988	*UGT8*	UDP glycosyltransferase 8	0.0046
ENST00000296684	MCP0017370	MMP0000981	*NDUFS4*	NADH dehydrogenase (ubiquinone) Fe-S protein 4, 18kDa (NADH-coenzyme Q reductase)	0.0046
ENST00000407426	MCP0014869	MMP0002002	*WDR43*	WD repeat domain 43	0.0047
ENST00000542230	MCP0007969	MMP0011163	*TMEM50B*	transmembrane protein 50B	0.0048
ENST00000274811	MCP0003067	MMP0003667	*RNF44*	ring finger protein 44	0.0048
ENST00000334418	MCP0018528	MMP0009454	*CCDC84*	coiled-coil domain containing 84	0.0048
ENST00000355303	MCP0013532	MMP0002716	*ANO1*	anoctamin 1, calcium activated chloride channel	0.0049
ENST00000264852	MCP0009814	MMP0019635	*SIDT1*	SID1 transmembrane family, member 1	0.0050
ENST00000309863	MCP0005702	MMP0007961	*GCC2*	GRIP and coiled-coil domain containing 2	0.0050
ENST00000215912	MCP0003785	MMP0000371	*PIK3IP1*	phosphoinositide-3-kinase interacting protein 1	0.0054
ENST00000545068	MCP0019773	MMP0005122	*FOXJ3*	forkhead box J3	0.0055
ENST00000330550	MCP0016133	MMP0015659	*SLC22A16*	solute carrier family 22 (organic cation/carnitine transporter), member 16	0.0055
ENST00000240364	MCP0005737	MMP0018438	*FAM117A*	family with sequence similarity 117, member A	0.0055
ENST00000300456	MCP0004129	MMP0016727	*SLC27A4*	solute carrier family 27 (fatty acid transporter), member 4	0.0055
ENST00000374566	MCP0018084	MMP0015061	*EPB41L4B*	erythrocyte membrane protein band 4.1 like 4B	0.0056

ENST00000301740	MCP0005104	MMP0003365	*SRRM2*	serine/arginine repetitive matrix 2	0.0056
ENST00000344846	MCP0003787	MMP0001893	*SYNGR4*	synaptogyrin 4	0.0058
ENST00000319211	MCP0019438	MMP0013162	*F2R*	coagulation factor II (thrombin) receptor	0.0059
ENST00000356090	MCP0001098	MMP0018684	*NCDN*	neurochondrin	0.0059
ENST00000300060	MCP0014718	MMP0010943	*ANPEP*	alanyl (membrane) aminopeptidase	0.0060
ENST00000340413	MCP0004303	MMP0020777	*NUP43*	nucleoporin 43kDa	0.0060
ENST00000245457	MCP0001670	MMP0009897	*PTGER2*	prostaglandin E receptor 2 (subtype EP2), 53kDa	0.0060
ENST00000289416	MCP0013726	MMP0011504	*ACSM3*	acyl-CoA synthetase medium-chain family member 3	0.0062
ENST00000324001	MCP0017442	MMP0020847	*PRX*	periaxin	0.0064
ENST00000323563	MCP0007589	MMP0004153	*MRPS31*	mitochondrial ribosomal protein S31	0.0064
ENST00000402418	MCP0020037	MMP0010176	*SLC25A19*	solute carrier family 25 (mitochondrial thiamine pyrophosphate carrier), member 19	0.0066
ENST00000367500	MCP0018845	MMP0007546	*SWT1*	SWT1 RNA endoribonuclease homolog (S. cerevisiae)	0.0066
ENST00000247138	MCP0017955	MMP0017948	*SLC35A2*	solute carrier family 35 (UDP-galactose transporter), member A2	0.0067
ENST00000292432	MCP0003074	MMP0003660	*HK3*	hexokinase 3 (white cell)	0.0070
ENST00000266943	MCP0004222	MMP0021379	*SLC46A3*	solute carrier family 46, member 3	0.0070
ENST00000517768	MCP0004190	MMP0010549	*MYOZ3*	myozenin 3	0.0072
ENST00000391809	MCP0012341	MMP0013799	*KLK5*	kallikrein-related peptidase 5	0.0073
ENST00000449910	MCP0011050	MMP0010488	*ADAM15*	ADAM metallopeptidase domain 15	0.0074
ENST00000308278	MCP0017882	MMP0004783	*FAM57A*	family with sequence similarity 57, member A	0.0074
ENST00000245663	MCP0000136	MMP0003285	*ZBTB46*	zinc finger and BTB domain containing 46	0.0075
ENST00000330498	MCP0005710	MMP0005291	*SLC5A2*	solute carrier family 5 (sodium/glucose cotransporter), member 2	0.0079
ENST00000217407	MCP0019043	MMP0006507	*LBP*	lipopolysaccharide binding protein	0.0079
ENST00000354753	MCP0016915	MMP0004855	*GPSM1*	G-protein signaling modulator 1	0.0079
ENST00000301645	MCP0014225	MMP0001776	*CYP7A1*	cytochrome P450, family 7, subfamily A, polypeptide 1	0.0080
ENST00000263593	MCP0001387	MMP0009049	*SIAE*	sialic acid acetylesterase	0.0080
ENST00000621632	MCP0001103	MMP0017093	*DDX24*	DEAD (Asp-Glu-Ala-Asp) box helicase 24	0.0081
ENST00000447092	MCP0003849	MMP0023062	*HYAL2*	hyaluronoglucosaminidase 2	0.0082
ENST00000359088	MCP0007332	MMP0017878	*ST6GALNAC1*	ST6 (alpha-N-acetyl-neuraminyl-2,3-beta-galactosyl-1,3)-N-	0.0083

				acetylgalactosaminide alpha-2,6-sialyltransferase 1	
ENST00000233615	MCP0019960	MMP0021807	*WBP1*	WW domain binding protein 1	0.0083
ENST00000361840	MCP0010459	MMP0000525	*SPRYD7*	SPRY domain containing 7	0.0085
ENST00000389722	MCP0013161	MMP0018690	*SPTB*	spectrin, beta, erythrocytic	0.0091
ENST00000411641	MCP0015400	MMP0000222	*AHSG*	alpha-2-HS-glycoprotein	0.0091
ENST00000616898	MCP0003819	MMP0005509	*HEMGN*	hemogen	0.0091
ENST00000370630	MCP0000919	MMP0003715	*CTBS*	chitobiase, di-N-acetyl-	0.0092
ENST00000550527	MCP0014906	MMP0012839	*APAF1*	apoptotic peptidase activating factor 1	0.0092
ENST00000327042	MCP0011082	MMP0012434	*TMEM86B*	transmembrane protein 86B	0.0092
ENST00000373574	MCP0012203	MMP0007019	*WDR38*	WD repeat domain 38	0.0092
ENST00000373921	MCP0015227	MMP0017398	*THEMIS2*	thymocyte selection associated family member 2	0.0092
ENST00000593360	MCP0015953	MMP0001364	*HAUS8*	HAUS augmin-like complex, subunit 8	0.0093
ENST00000325207	MCP0015257	MMP0001745	*RIC8A*	RIC8 guanine nucleotide exchange factor A	0.0097
ENST00000360586	MCP0000997	MMP0019124	*WDHD1*	WD repeat and HMG-box DNA binding protein 1	0.0099
ENST00000358430	MCP0003232	MMP0002103	*TXLNB*	taxilin beta	0.0099
ENST00000337526	MCP0008816	MMP0001505	*RTN4*	reticulon 4	0.0100
ENST00000301919	MCP0011580	MMP0014551	*MSANTD4*	Myb/SANT-like DNA-binding domain containing 4 with coiled-coils	0.0100
ENST00000352105	MCP0017673	MMP0016505	*ACAN*	aggrecan	0.0103
ENST00000209929	MCP0019534	MMP0017754	*FMO2*	flavin containing monooxygenase 2 (non-functional)	0.0110
ENST00000334186	MCP0018062	MMP0004021	*PPFIA3*	protein tyrosine phosphatase, receptor type, f polypeptide (PTPRF), interacting protein (liprin), alpha 3	0.0114
ENST00000223129	MCP0002345	MMP0014069	*RPA3*	replication protein A3, 14kDa	0.0115
ENST00000267082	MCP0009066	MMP0008491	*ITGB7*	integrin, beta 7	0.0116
ENST00000602404	MCP0005556	MMP0004720	*NDUFA6*	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 6, 14kDa	0.0117
ENST00000452691	MCP0004730	MMP0013547	*CCDC36*	coiled-coil domain containing 36	0.0117
ENST00000366844	MCP0007354	MMP0015889	*ENAH*	enabled homolog (Drosophila)	0.0117
ENST00000269373	MCP0013449	MMP0021667	*FN3KRP*	fructosamine 3 kinase related protein	0.0121
ENST00000296161	MCP0016527	MMP0006292	*DTX3L*	deltex 3 like, E3 ubiquitin ligase	0.0122
ENST00000311907	MCP0003090	MMP0018178	*F2*	coagulation factor II (thrombin)	0.0123
ENST00000308304	MCP0009102	MMP0017064	*PROP1*	PROP paired-like homeobox 1	0.0125
ENST00000457542	MCP0006577	MMP0016026	*MAPKBP1*	mitogen-activated protein kinase binding protein 1	0.0125
ENST00000307017	MCP0002694	MMP0011274	*USP38*	ubiquitin specific peptidase 38	0.0127

ENST00000398955	MCP0007453	MMP0013031	*MGARP*	mitochondria-localized glutamic acid-rich protein	0.0130
ENST00000560491	MCP0007085	MMP0020598	*LYSMD2*	LysM, putative peptidoglycan-binding, domain containing 2	0.0130
ENST00000354955	MCP0014950	MMP0002579	*FMOD*	fibromodulin	0.0131
ENST00000370646	MCP0003657	MMP0022557	*HOGA1*	4-hydroxy-2-oxoglutarate aldolase 1	0.0132
ENST00000260970	MCP0004809	MMP0009231	*PPIG*	peptidylprolyl isomerase G (cyclophilin G)	0.0132
ENST00000216223	MCP0002419	MMP0012854	*IL2RB*	interleukin 2 receptor, beta	0.0133
ENST00000379144	MCP0000296	MMP0013153	*PCYT1B*	phosphate cytidylyltransferase 1, choline, beta	0.0134
ENST00000370978	MCP0000772	MMP0021084	*ZNF280C*	zinc finger protein 280C	0.0135
ENST00000368662	MCP0011383	MMP0016855	*TUBE1*	tubulin, epsilon 1	0.0136
ENST00000337508	MCP0010776	MMP0010042	*NRIP2*	nuclear receptor interacting protein 2	0.0137
ENST00000522447	MCP0002806	MMP0011283	*LACTB2*	lactamase, beta 2	0.0139
ENST00000351989	MCP0013788	MMP0020268	*DGCR8*	DGCR8 microprocessor complex subunit	0.0142
ENST00000263383	MCP0008985	MMP0002425	*ILVBL*	ilvB (bacterial acetolactate synthase)-like	0.0142
ENST00000374672	MCP0006831	MMP0010874	*KLF4*	Kruppel-like factor 4 (gut)	0.0144
ENST00000428726	MCP0000978	MMP0023338	*CD44*	CD44 molecule (Indian blood group)	0.0144
ENST00000589872	MCP0016178	MMP0021992	*NBR1*	neighbor of BRCA1 gene 1	0.0148
ENST00000358242	MCP0005976	MMP0007904	*DMTN*	dematin actin binding protein	0.0151
ENST00000378313	MCP0009768	MMP0009843	*C19orf54*	chromosome 19 open reading frame 54	0.0153
ENST00000266754	MCP0019383	MMP0005009	*GAS2L3*	growth arrest-specific 2 like 3	0.0156
ENST00000361632	MCP0004387	MMP0008043	*CSF3R*	colony stimulating factor 3 receptor (granulocyte)	0.0158
ENST00000336431	MCP0011090	MMP0002047	*GGT7*	gamma-glutamyltransferase 7	0.0159
ENST00000356517	MCP0001853	MMP0019384	*AADACL2*	arylacetamide deacetylase-like 2	0.0160
ENST00000342427	MCP0020078	MMP0006212	*ZNF341*	zinc finger protein 341	0.0162
ENST00000378079	MCP0017117	MMP0012259	*FBXO47*	F-box protein 47	0.0163
ENST00000361249	MCP0010977	MMP0013244	*C8A*	complement component 8, alpha polypeptide	0.0165
ENST00000373256	MCP0019426	MMP0011949	*GLP1R*	glucagon-like peptide 1 receptor	0.0165
ENST00000318663	MCP0006710	MMP0003096	*ORAI3*	ORAI calcium release-activated calcium modulator 3	0.0165
ENST00000491431	MCP0011328	MMP0020561	*ZNF786*	zinc finger protein 786	0.0166
ENST00000373795	MCP0002472	MMP0011315	*SRSF4*	serine/arginine-rich splicing factor 4	0.0167
ENST00000368842	MCP0006994	MMP0008481	*LHPP*	phospholysine phosphohistidine inorganic pyrophosphate phosphatase	0.0167

ENST00000338560	MCP0019120	MMP0016525	*TRPV2*	transient receptor potential cation channel, subfamily V, member 2	0.0169
ENST00000599564	MCP0006528	MMP0000806	*GRAMD1A*	GRAM domain containing 1A	0.0170
ENST00000343484	MCP0016587	MMP0003292	*TCEA2*	transcription elongation factor A (SII), 2	0.0171
ENST00000253801	MCP0016167	MMP0021988	*G6PC*	glucose-6-phosphatase, catalytic subunit	0.0171
ENST00000341500	MCP0002795	MMP0008821	*INSR*	insulin receptor	0.0171
ENST00000378357	MCP0015565	MMP0022380	*CA9*	carbonic anhydrase IX	0.0175
ENST00000359128	MCP0000445	MMP0006146	*NLRC3*	NLR family, CARD domain containing 3	0.0175
ENST00000007722	MCP0005741	MMP0000160	*ITGA3*	integrin, alpha 3 (antigen CD49C, alpha 3 subunit of VLA-3 receptor)	0.0177
ENST00000324464	MCP0009769	MMP0002064	*ADCK4*	aarF domain containing kinase 4	0.0179
ENST00000370611	MCP0012259	MMP0003720	*LPAR3*	lysophosphatidic acid receptor 3	0.0180
ENST00000344657	MCP0009992	MMP0007014	*SPHKAP*	SPHK1 interactor, AKAP domain containing	0.0181
ENST00000436063	MCP0002089	MMP0018374	*DNTTIP2*	deoxynucleotidyltransferase, terminal, interacting protein 2	0.0181
ENST00000347842	MCP0000655	MMP0005502	*ASIC4*	acid sensing (proton gated) ion channel family member 4	0.0182
ENST00000320027	MCP0004993	MMP0018827	*HSPH1*	heat shock 105kDa/110kDa protein 1	0.0185
ENST00000521315	MCP0005965	MMP0008119	*SFTPC*	surfactant protein C	0.0187
ENST00000344206	MCP0006746	MMP0023042	*MST1R*	macrophage stimulating 1 receptor	0.0188
ENST00000487519	MCP0006585	MMP0009797	*TFAM*	transcription factor A, mitochondrial	0.0194
ENST00000396091	MCP0013441	MMP0012082	*ANO10*	anoctamin 10	0.0194
ENST00000372838	MCP0004130	MMP0012292	*CERCAM*	cerebral endothelial cell adhesion molecule	0.0195
ENST00000072869	MCP0014531	MMP0010736	*ADCK2*	aarF domain containing kinase 2	0.0195
ENST00000274605	MCP0010862	MMP0016560	*N4BP3*	NEDD4 binding protein 3	0.0195
ENST00000378750	MCP0001807	MMP0006321	*PEX16*	peroxisomal biogenesis factor 16	0.0197
ENST00000262753	MCP0007362	MMP0003307	*POF1B*	premature ovarian failure, 1B	0.0201
ENST00000358807	MCP0008894	MMP0008203	*MICAL1*	microtubule associated monooxygenase, calponin and LIM domain containing 1	0.0204
ENST00000290472	MCP0017819	MMP0016029	*PLA2G4D*	phospholipase A2, group IVD (cytosolic)	0.0204
ENST00000486442	MCP0011737	MMP0018623	*KLHL29*	kelch-like family member 29	0.0207
ENST00000268379	MCP0014411	MMP0019697	*UQCRC2*	ubiquinol-cytochrome c reductase core protein II	0.0207
ENST00000370759	MCP0000461	MMP0021023	*GIPC2*	GIPC PDZ domain containing family, member 2	0.0209
ENST00000504930	MCP0006210	MMP0017961	*POLR3G*	polymerase (RNA) III (DNA directed) polypeptide G (32kD)	0.0209
ENST00000357175	MCP0017012	MMP0000275	*MUM1L1*	melanoma associated antigen (mutated)	0.0213

				1-like 1	
ENST00000225174	MCP0011163	MMP0010105	*PPIF*	peptidylprolyl isomerase F	0.0213
ENST00000281722	MCP0018684	MMP0001866	*RBM46*	RNA binding motif protein 46	0.0214
ENST00000402105	MCP0013036	MMP0000963	*HPS4*	Hermansky-Pudlak syndrome 4	0.0215
ENST00000301908	MCP0005275	MMP0003962	*PNOC*	prepronociceptin	0.0216
ENST00000356443	MCP0016738	MMP0011564	*MYOM1*	myomesin 1	0.0219
ENST00000257909	MCP0015119	MMP0018357	*TROAP*	trophinin associated protein	0.0220
ENST00000342315	MCP0019800	MMP0010055	*OAS2*	2'-5'-oligoadenylate synthetase 2, 69/71kDa	0.0221
ENST00000256190	MCP0005252	MMP0019435	*SBF2*	SET binding factor 2	0.0223
ENST00000328119	MCP0012382	MMP0015051	*KRT36*	keratin 36, type I	0.0224
ENST00000273183	MCP0002165	MMP0015548	*STAC*	SH3 and cysteine rich domain	0.0225
ENST00000370853	MCP0013876	MMP0001106	*MBNL3*	muscleblind-like splicing regulator 3	0.0226
ENST00000551812	MCP0011596	MMP0006972	*BAZ2A*	bromodomain adjacent to zinc finger domain, 2A	0.0226
ENST00000263816	MCP0016283	MMP0000802	*LRP2*	low density lipoprotein receptor-related protein 2	0.0227
ENST00000382751	MCP0010551	MMP0010509	*URB1*	URB1 ribosome biogenesis 1 homolog (S. cerevisiae)	0.0228
ENST00000617259	MCP0012431	MMP0003984	*IL13*	interleukin 13	0.0230
ENST00000253686	MCP0001504	MMP0007745	*MRPS25*	mitochondrial ribosomal protein S25	0.0231
ENST00000378043	MCP0007191	MMP0016825	*BEST1*	bestrophin 1	0.0232
ENST00000287295	MCP0018314	MMP0021081	*AIFM1*	apoptosis-inducing factor, mitochondrion-associated, 1	0.0235
ENST00000228936	MCP0012592	MMP0023387	*ART4*	ADP-ribosyltransferase 4 (Dombrock blood group)	0.0238
ENST00000216274	MCP0014040	MMP0022997	*RIPK3*	receptor-interacting serine-threonine kinase 3	0.0240
ENST00000320005	MCP0002024	MMP0018307	*CNGB3*	cyclic nucleotide gated channel beta 3	0.0242
ENST00000229729	MCP0014390	MMP0018839	*SLC44A4*	solute carrier family 44, member 4	0.0243
ENST00000294818	MCP0016138	MMP0014854	*LRRC52*	leucine rich repeat containing 52	0.0244
ENST00000318315	MCP0006161	MMP0022978	*C5orf46*	chromosome 5 open reading frame 46	0.0244
ENST00000369478	MCP0009172	MMP0001338	*CD2*	CD2 molecule	0.0245
ENST00000316428	MCP0003115	MMP0002258	*LRRC31*	leucine rich repeat containing 31	0.0248
ENST00000392588	MCP0008467	MMP0016019	*WASF1*	WAS protein family, member 1	0.0250
ENST00000279873	MCP0015156	MMP0010966	*ARID5B*	AT rich interactive domain 5B (MRF1-like)	0.0251
ENST00000301671	MCP0012355	MMP0006933	*GHDC*	GH3 domain containing	0.0253
ENST00000358241	MCP0015861	MMP0011975	*RTP2*	receptor (chemosensory) transporter protein 2	0.0254

ENST00000252032	MCP0012682	MMP0003973	*C20orf194*	chromosome 20 open reading frame 194	0.0254
ENST00000254928	MCP0019425	MMP0018894	*ERAL1*	Era-like 12S mitochondrial rRNA chaperone 1	0.0256
ENST00000216471	MCP0014469	MMP0019215	*SAMD15*	sterile alpha motif domain containing 15	0.0256
ENST00000223026	MCP0013849	MMP0019748	*HYAL4*	hyaluronoglucosaminidase 4	0.0259
ENST00000255039	MCP0006306	MMP0000774	*HAPLN2*	hyaluronan and proteoglycan link protein 2	0.0260
ENST00000366518	MCP0017619	MMP0021968	*KIF26B*	kinesin family member 26B	0.0262
ENST00000297596	MCP0018348	MMP0016530	*GEM*	GTP binding protein overexpressed in skeletal muscle	0.0265
ENST00000371901	MCP0006938	MMP0003865	*CYP4X1*	cytochrome P450, family 4, subfamily X, polypeptide 1	0.0269
ENST00000367025	MCP0002075	MMP0019880	*TRAF3IP3*	TRAF3 interacting protein 3	0.0269
ENST00000594369	MCP0007755	MMP0022304	*ZNF446*	zinc finger protein 446	0.0271
ENST00000268124	MCP0017667	MMP0007595	*POLG*	polymerase (DNA directed), gamma	0.0274
ENST00000371980	MCP0012607	MMP0003592	*LURAP1*	leucine rich adaptor protein 1	0.0275
ENST00000389194	MCP0003387	MMP0020809	*LTN1*	listerin E3 ubiquitin protein ligase 1	0.0275
ENST00000544455	MCP0016017	MMP0012495	*BRCA2*	breast cancer 2, early onset	0.0278
ENST00000372479	MCP0002921	MMP0017770	*RBM41*	RNA binding motif protein 41	0.0282
ENST00000454584	MCP0012178	MMP0018514	*GAS2*	growth arrest-specific 2	0.0282
ENST00000264499	MCP0009711	MMP0007162	*BBS7*	Bardet-Biedl syndrome 7	0.0283
ENST00000310823	MCP0013431	MMP0011525	*ADAM17*	ADAM metallopeptidase domain 17	0.0285
ENST00000447906	MCP0016536	MMP0011234	*OTUD4*	OTU deubiquitinase 4	0.0286
ENST00000548660	MCP0016668	MMP0013267	*GLT8D2*	glycosyltransferase 8 domain containing 2	0.0291
ENST00000408910	MCP0011260	MMP0012771	*UMODL1*	uromodulin-like 1	0.0291
ENST00000330342	MCP0005412	MMP0009538	*ATP6V0A2*	ATPase, H+ transporting, lysosomal V0 subunit a2	0.0294
ENST00000315939	MCP0014652	MMP0002522	*WNK1*	WNK lysine deficient protein kinase 1	0.0295
ENST00000409652	MCP0009467	MMP0010667	*APOL6*	apolipoprotein L, 6	0.0298
ENST00000552570	MCP0009953	MMP0022593	*TNS2*	tensin 2	0.0299
ENST00000441366	MCP0013779	MMP0015144	*EPB42*	erythrocyte membrane protein band 4.2	0.0300
ENST00000613904	MCP0000800	MMP0022899	*TTI2*	TELO2 interacting protein 2	0.0301
ENST00000531316	MCP0000865	MMP0022729	*C11orf63*	chromosome 11 open reading frame 63	0.0302
ENST00000619684	MCP0013903	MMP0004499	*KIAA1524*	KIAA1524	0.0305
ENST00000263663	MCP0000943	MMP0018392	*TAF1B*	TATA box binding protein (TBP)-associated factor, RNA polymerase I, B, 63kDa	0.0307
ENST00000266971	MCP0013648	MMP0004763	*SUOX*	sulfite oxidase	0.0309
ENST00000202556	MCP0003446	MMP0012094	*PPP1R13B*	protein phosphatase 1, regulatory subunit	0.0311

				13B	
ENST00000254508	MCP0009571	MMP0006747	*NUP210*	nucleoporin 210kDa	0.0312
ENST00000302250	MCP0002176	MMP0009244	*FAM151A*	family with sequence similarity 151, member A	0.0313
ENST00000382492	MCP0019195	MMP0001867	*TAS2R1*	taste receptor, type 2, member 1	0.0317
ENST00000441564	MCP0020157	MMP0018382	*PSD4*	pleckstrin and Sec7 domain containing 4	0.0317
ENST00000223836	MCP0017617	MMP0007936	*AK1*	adenylate kinase 1	0.0318
ENST00000358755	MCP0005575	MMP0017408	*FZD6*	frizzled class receptor 6	0.0320
ENST00000225275	MCP0016854	MMP0005707	*MPO*	myeloperoxidase	0.0322
ENST00000395048	MCP0017309	MMP0003875	*CYP1A1*	cytochrome P450, family 1, subfamily A, polypeptide 1	0.0323
ENST00000395343	MCP0011403	MMP0023107	*DIDO1*	death inducer-obliterator 1	0.0325
ENST00000547303	MCP0013065	MMP0004933	*DDIT3*	DNA-damage-inducible transcript 3	0.0331
ENST00000371941	MCP0005304	MMP0009152	*PREX1*	phosphatidylinositol-3,4,5-trisphosphate-dependent Rac exchange factor 1	0.0334
ENST00000399645	MCP0000449	MMP0006148	*C16orf90*	chromosome 16 open reading frame 90	0.0334
ENST00000271636	MCP0018789	MMP0006678	*CGN*	cingulin	0.0335
ENST00000378910	MCP0003244	MMP0000620	*NPHS1*	nephrosis 1, congenital, Finnish type (nephrin)	0.0336
ENST00000373347	MCP0014149	MMP0011869	*DLGAP3*	discs, large (Drosophila) homolog-associated protein 3	0.0337
ENST00000373362	MCP0014152	MMP0011866	*GJB3*	gap junction protein, beta 3, 31kDa	0.0337
ENST00000339777	MCP0012322	MMP0017512	*LRRC43*	leucine rich repeat containing 43	0.0339
ENST00000340990	MCP0015698	MMP0021217	*ADIPOR1*	adiponectin receptor 1	0.0339
ENST00000378673	MCP0016480	MMP0008644	*GDF9*	growth differentiation factor 9	0.0343
ENST00000228837	MCP0014289	MMP0021398	*FGF6*	fibroblast growth factor 6	0.0343
ENST00000361256	MCP0011630	MMP0008295	*C9orf114*	chromosome 9 open reading frame 114	0.0345
ENST00000216487	MCP0006109	MMP0021879	*RIN3*	Ras and Rab interactor 3	0.0348
ENST00000320634	MCP0011619	MMP0004344	*FAIM2*	Fas apoptotic inhibitory molecule 2	0.0350
ENST00000374954	MCP0013496	MMP0001124	*ASIP*	agouti signaling protein	0.0353
ENST00000573584	MCP0010752	MMP0006662	*NUP88*	nucleoporin 88kDa	0.0353
ENST00000409991	MCP0018538	MMP0009448	*NLRX1*	NLR family member X1	0.0354
ENST00000340611	MCP0001629	MMP0023234	*BRAT1*	BRCA1-associated ATM activator 1	0.0355
ENST00000331456	MCP0006745	MMP0023043	*TRAIP*	TRAF interacting protein	0.0356
ENST00000401399	MCP0019651	MMP0023154	*NFASC*	neurofascin	0.0358
ENST00000370017	MCP0012422	MMP0018767	*FNDC7*	fibronectin type III domain containing 7	0.0358
ENST00000263314	MCP0016354	MMP0005860	*P2RX3*	purinergic receptor P2X, ligand gated ion channel, 3	0.0359
ENST00000398712	MCP0008237	MMP0020956	*SHARPIN*	SHANK-associated RH domain interactor	0.0359
ENST00000551568	MCP0015111	MMP0009618	*CPM*	carboxypeptidase M	0.0360
ENST00000357089	MCP0004560	MMP0005830	*UBXN11*	UBX domain protein 11	0.0361

ENST00000380379	MCP0005453	MMP0017402	*BPHL*	biphenyl hydrolase-like (serine hydrolase)	0.0362
ENST00000495893	MCP0018184	MMP0007567	*PHC3*	polyhomeotic homolog 3 (Drosophila)	0.0363
ENST00000518444	MCP0006591	MMP0023036	*LARP4*	La ribonucleoprotein domain family, member 4	0.0363
ENST00000258403	MCP0009994	MMP0006068	*SLC19A3*	solute carrier family 19 (thiamine transporter), member 3	0.0366
ENST00000409423	MCP0015797	MMP0004754	*NCAPG2*	non-SMC condensin II complex, subunit G2	0.0370
ENST00000287497	MCP0004385	MMP0004069	*ITGAM*	integrin, alpha M (complement component 3 receptor 3 subunit)	0.0375
ENST00000409655	MCP0019334	MMP0001742	*ATHL1*	ATH1, acid trehalase-like 1 (yeast)	0.0377
ENST00000374479	MCP0001536	MMP0020606	*FUCA1*	fucosidase, alpha-L- 1, tissue	0.0378
ENST00000239440	MCP0014237	MMP0002154	*ARAP3*	ArfGAP with RhoGAP domain, ankyrin repeat and PH domain 3	0.0381
ENST00000355710	MCP0019667	MMP0010315	*RET*	ret proto-oncogene	0.0387
ENST00000376236	MCP0017424	MMP0007618	*APBB1IP*	amyloid beta (A4) precursor protein-binding, family B, member 1 interacting protein	0.0388
ENST00000376042	MCP0014429	MMP0012318	*CCNB3*	cyclin B3	0.0393
ENST00000371139	MCP0014209	MMP0015425	*SH2D1A*	SH2 domain containing 1A	0.0394
ENST00000226067	MCP0017959	MMP0023337	*HLF*	hepatic leukemia factor	0.0394
ENST00000392692	MCP0009724	MMP0009221	*ECT2*	epithelial cell transforming 2	0.0399
ENST00000457717	MCP0015300	MMP0000748	*MTTP*	microsomal triglyceride transfer protein	0.0401
ENST00000372626	MCP0010273	MMP0005152	*TCEAL1*	transcription elongation factor A (SII)-like 1	0.0406
ENST00000315251	MCP0005367	MMP0004513	*CHDH*	choline dehydrogenase	0.0411
ENST00000262415	MCP0010589	MMP0001013	*DHX8*	DEAH (Asp-Glu-Ala-His) box polypeptide 8	0.0411
ENST00000325680	MCP0008831	MMP0001371	*YLPM1*	YLP motif containing 1	0.0414
ENST00000328886	MCP0001924	MMP0010218	*TMIGD1*	transmembrane and immunoglobulin domain containing 1	0.0414
ENST00000449682	MCP0012616	MMP0015381	*MST1*	macrophage stimulating 1	0.0415
ENST00000378387	MCP0015555	MMP0022373	*ARHGEF39*	Rho guanine nucleotide exchange factor (GEF) 39	0.0416
ENST00000263577	MCP0007381	MMP0007092	*CDON*	cell adhesion associated, oncogene regulated	0.0422
ENST00000284288	MCP0001385	MMP0009043	*PANX3*	pannexin 3	0.0426
ENST00000261170	MCP0012594	MMP0011166	*GUCY2C*	guanylate cyclase 2C	0.0431
ENST00000217026	MCP0017044	MMP0018634	*MYBL2*	v-myb avian myeloblastosis viral oncogene homolog-like 2	0.0434

ENST00000261435	MCP0000775	MMP0019092	*N4BP2*	NEDD4 binding protein 2	0.0434
ENST00000427704	MCP0016402	MMP0009616	*PHACTR2*	phosphatase and actin regulator 2	0.0439
ENST00000406927	MCP0013369	MMP0020781	*METTL21A*	methyltransferase like 21A	0.0439
ENST00000555619	MCP0000558	MMP0002393	*NPC2*	Niemann-Pick disease, type C2	0.0442
ENST00000222598	MCP0019547	MMP0016736	*DLX5*	distal-less homeobox 5	0.0442
ENST00000340360	MCP0013583	MMP0011026	*XRRA1*	X-ray radiation resistance associated 1	0.0443
ENST00000392542	MCP0009499	MMP0015985	*RFC5*	replication factor C (activator 1) 5, 36.5kDa	0.0449
ENST00000286096	MCP0010880	MMP0012089	*KDM8*	lysine (K)-specific demethylase 8	0.0450
ENST00000447830	MCP0010287	MMP0008659	*SPATC1*	spermatogenesis and centriole associated 1	0.0450
ENST00000216445	MCP0013018	MMP0019280	*C14orf105*	chromosome 14 open reading frame 105	0.0453
ENST00000265983	MCP0001047	MMP0002130	*HPX*	hemopexin	0.0453
ENST00000257789	MCP0012558	MMP0017456	*ORC3*	origin recognition complex, subunit 3	0.0461
ENST00000403389	MCP0007994	MMP0013200	*OSM*	oncostatin M	0.0464
ENST00000257189	MCP0014019	MMP0017373	*DSG3*	desmoglein 3	0.0467
ENST00000411463	MCP0005155	MMP0008143	*R3HCC1*	R3H domain and coiled-coil containing 1	0.0471
ENST00000297581	MCP0011582	MMP0010726	*DCSTAMP*	dendrocyte expressed seven transmembrane protein	0.0474
ENST00000281282	MCP0010015	MMP0015394	*CGNL1*	cingulin-like 1	0.0475
ENST00000378588	MCP0013522	MMP0014350	*CYBB*	cytochrome b-245, beta polypeptide	0.0475
ENST00000312777	MCP0009217	MMP0002230	*TCHP*	trichoplein, keratin filament binding	0.0479
ENST00000434748	MCP0001658	MMP0022723	*FBRSL1*	fibrosin-like 1	0.0484
ENST00000585618	MCP0004820	MMP0000552	*SEC14L1*	SEC14-like 1 (S. cerevisiae)	0.0485
ENST00000498508	MCP0004308	MMP0022478	*PROX1*	prospero homeobox 1	0.0490
ENST00000325203	MCP0005437	MMP0008689	*ANGPT2*	angiopoietin 2	0.0490
ENST00000389005	MCP0000413	MMP0022658	*C17orf85*	chromosome 17 reading frame 85	0.0493
ENST00000296137	MCP0016953	MMP0015637	*FYCO1*	FYVE and coiled-coil domain containing 1	0.0496

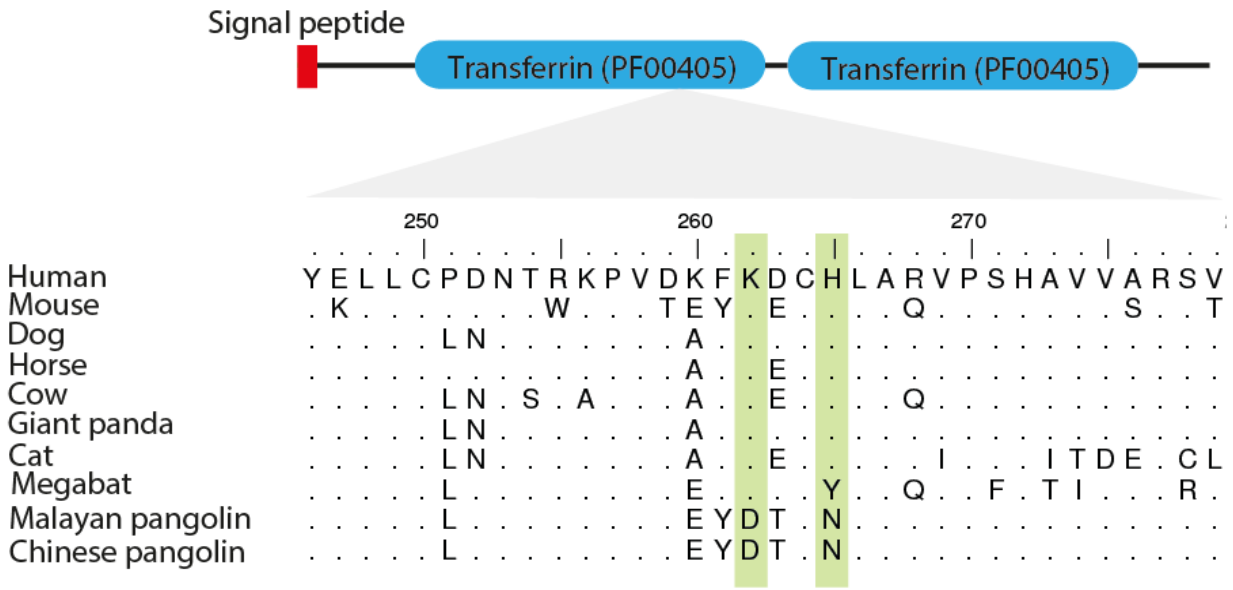
Supplemental Table S9.2. List of the selected genes under positive selection in pangolins with their assigned functions or pathways.

Function	Genes	Function	Genes
Immunity	Hematopoietic cell lineage *CD44, ITGA3, ITGAM, CD2, ANPEP, CSF3R* Cytokine-cytokine receptor interaction *LTBR, IL2RB, TNFRSF13B, IL13, BMPR1B, CSF3R, OSM* Complement and coagulation cascades *F2, F2R, C8A, A2M, MASP2* Phagosome pathway *ATP6V0A2, MPO, RAB5B, ITGAM, CYBB* Ameobiasis pathway *LAMB2,COL1A2, RAB5B, COL4A4, ITGAM, C8A* Cytosolic DNA-sensing pathway *TMEM173, RIPK3, POLR3G* Lysosome pathway *FUCA1, ATP6V0A2, ACP2, NPC2, ABCA2* Others *LTF, LY86, TNFRSF13B, MASP2, TIGIT, AKNA, HCLS1, TREM1, MAPK9, SIAE, ITGB7*	**Inflammation**	*LPCAT2, RAB18, CAPN5, CORO2A, LBP, ADAM17, REL, ITGAM, RIPK3, IL13, P2RX7, PTGER2, IL2RB, MPO,LTF, AHSG, LTBR, TREM1, LBP, F2R, ANGPT2 , OSM, MASP2*
Energy storage and metabolism	*GP2, ASXL1, SLC19A1, SLC27A4, ACSM3, SLC35A2, HK3, CYP7A1, AHSG, CTBS, TPH1, ANPEP*	**Nervous system**	*KIF1B, DOCK6, UGT8, PRX, MAP2, KCNH4, MERTK, LAMB2, PNKP, RTN4*
Mitochondrial metabolism	*ISCA2, SDHC, BCAT2, LETM1, ACAT1, GTPBP3, NDUFS4, FOXJ3, MRPS31, SLC25A19, HOGA1*	**Osteogenesis**	*COL1A2, MSH3, BMPR1B, SPP2, ACAN*
Muscular	*SGCG, NEBL, LDB3, FLNC, NEK8, LGALS2, RTCA, CACNG1, MYOZ3, TXLNB*	**Apoptosis**	*BIRC5, BIRC7, TOPBP1, APAF1*
Hair/Scale	*KRT75, KRT82, KRTAP3-1, SYNE4, KRT36*	**Skin**	*P2RX7, KLK5*
Eye	*KRT12, ADAMTS14*	**Olfaction**	*ARHGEF40, NCDN*
Stress	*USP15, RPA3*	**ECM-receptor interaction pathway**	*LAMB2, ITGB7, COL1A2, CD44, ITGA3, COL4A4*

Regulation of actin cytoskeleton pathway	*ITGB7, F2, ITGA3, WASF1, FGF6, ITGAM, ENAH, F2R*	**Homologous recombination**	*POLD3, RPA3, BRCA2, RAD51B*
Mismatch repair	*RFC5, POLD3, RPA3, MSH3*		

Supplemental Table S9.3. **Positively-selected genes associated with known diseases.**

Function	Genes	Function	Genes
Cancer or viral infection	*KLK5, ABCB5, TOPBP1, SDHC, ST14, RET, MYBL2, HYAL2, KIAA1524, SLC19A1, CYP1A1, SH2D1A, BRCA2, BIRC7, ITGA3, , MST1R, BIRC5, KLF4, MSH3, RAD51B, APAF1, CD44, POLH, DDIT3, CA9*	**Inflammation**	*ADAM17, REL, ITGAM, RIPK3, IL13, P2RX7, PTGER2, IL2RB, MPO, LTF AHSG, LTBR, TREM1, LBP, F2R, ANGPT2, OSM, MASP2*
Bacterial infection	*SLC37A4, TREM1, ITGAM, IL13, P2RX7, LBP, MPO, LTF, MASP2, CYBB*	**Pneumonia**	*LBP, , TREM1, MPO, SFTPC IL13, CYBB, MASP2*
Skin diseases	*KLK5, BRCA2, CYP1A1, MST1R, ST14, THEMIS2, HPS4, IL13, RAD51B, GJB3, ASIP, KRT75, CD44, POLH, DSG3*	**Gastrointestinal diseases**	*CYP1A1, MST1R, BIRC5, KLF4, RET, MSH3, DKK4, GUCY2C, PTGER2, CD44, MST1, ANO1, MUC13, CA9*



Supplemental Figure S9.1. Positively selected *LTF* gene. Two critical amino acid changes which are located in the functionally relevant Transferrin domain predicted by Pfam are highlighted in green. For the LTF, p-value of the detection of positive selection was 1.80×10^{-006}.

10.0 REPEAT ANNOTATION

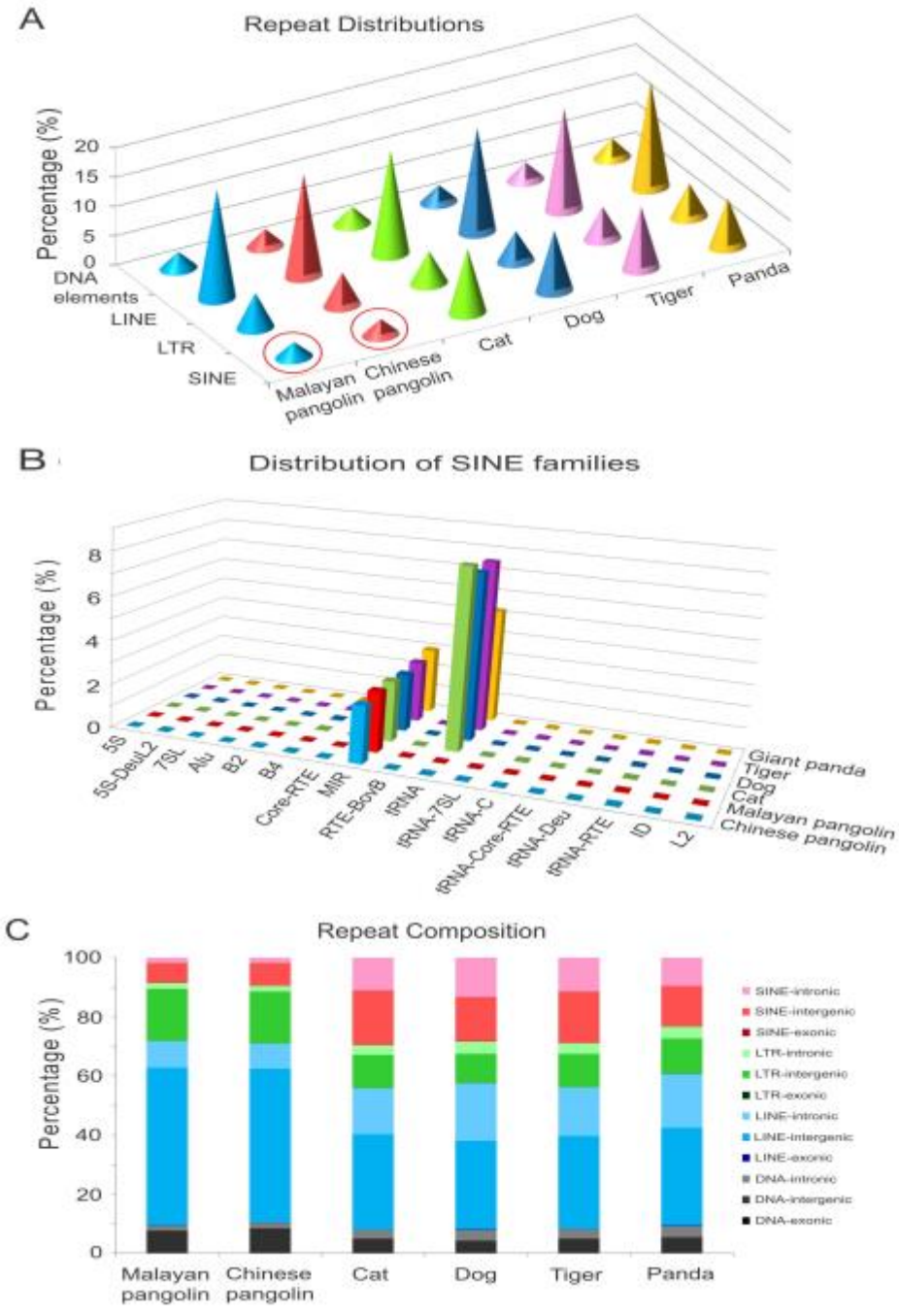
10.1 RepeatMasker

Repetitive elements (RE) were identified using RepeatMasker open-4.0.5(Smit 2013-2015) against the Repbase TE database (version 2014-01-31)(Jurka 2000). The RepeatMasker open-4.0.5 program was run separately with species parameter set to mammal and carnivora. The percentage of each repeat classes were summarised in **Supplemental Table S10.1**. We also analyzed and compared the repeat landscapes (depict the relative abundance of repeat classes in the genome versus the Kimura divergence) of Malayan and Chinese pangolins (**Supplemental Figure S10.2**). Tandem repeats were also predicted and shown in Supplemental Information 10.3.

Supplemental Table S10.1. Summary of repeats in the pangolin genomes. Repeats were searched using RepeatMasker against the Repbase TE database**.**

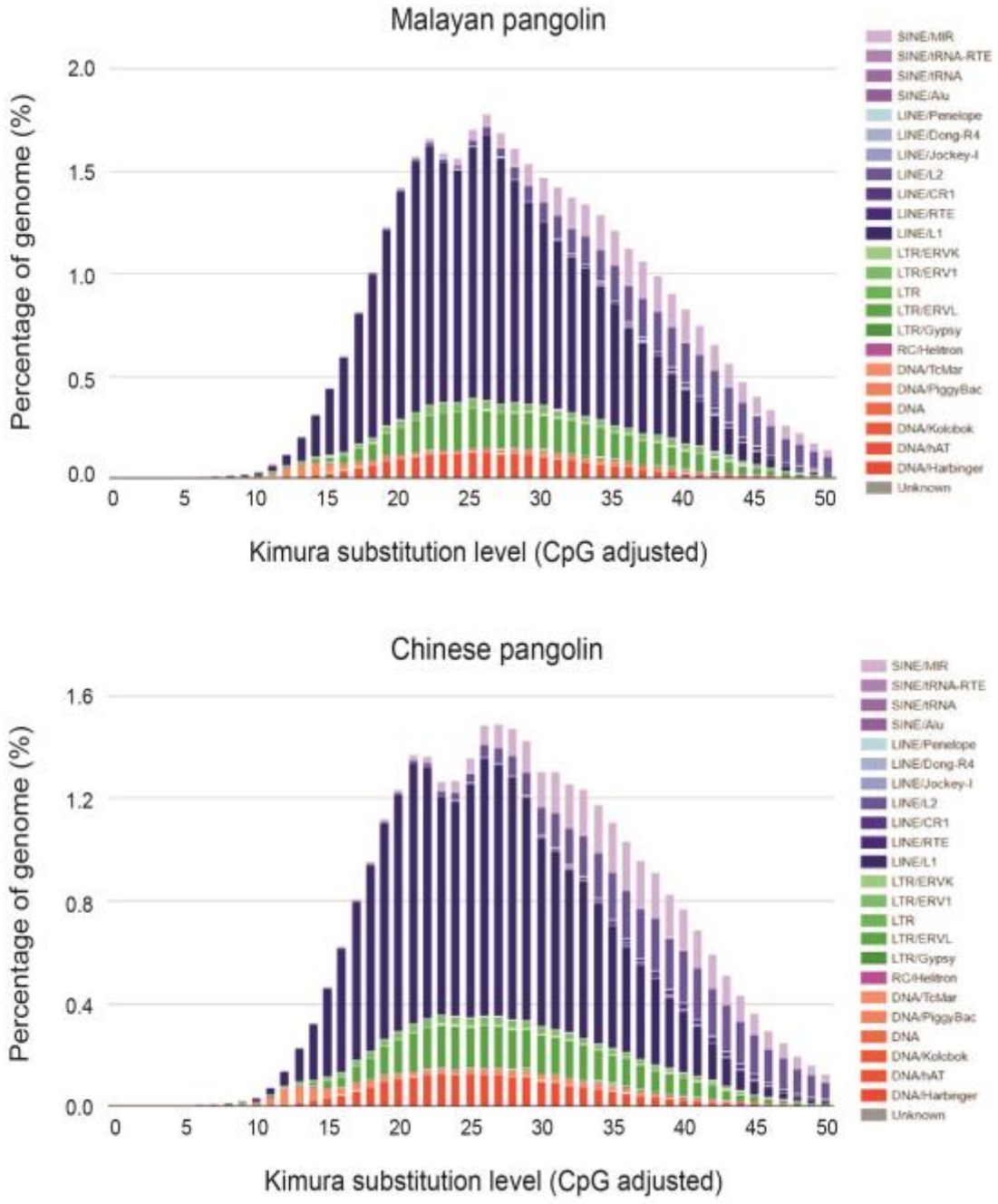
	Malayan pangolin		**Chinese pangolin**	
Library	mammal	carnivora	mammal	carnivora
LINE	18.74%	21.80%	17.36%	21.72%
SINE	2.61%	2.63%	2.79%	2.82%

LTR Element	6.05%	4.93%	5.65%	5.50%
DNA Elements	2.90%	2.76%	3.02%	3.13%



Supplemental Figure S10.1. Repeats analysis. (a) Major repeat elements found in selected mammalian species. The LTR, DNA elements and LINEs, except SINEs in pangolin genomes are comparable to other

mammals. (b) Distribution of SINE families across pangolins and closely related mammalian species. (c) Repeat composition by genome structures.



Supplemental Figure S10.2. Comparison between the repeat landscapes of Malayan and Chinese pangolins. No significant recent repeat family activity unique to either Malayan or Chinese pangolins was observed in this analysis.

Supplemental Table S10.2. Summary of *de novo* repeats in the pangolin genomes. Repeats were searched using RepeatMasker against the consensus repeat library generated using RepeatModeler**.**

	Malayan pangolin	Chinese pangolin
LINE	19.19%	16.37%
SINE	1.41%	1.11%
LTR Element	5.83%	5.07%
DNA Elements	1.86%	1.81%
Unclassified	0.89%	1.22%

10.2 Tandem repeats

TRF output was processed with Trevis software to compute basic statistics for tandem repeats. It is mostly the estimation of assembly quality because tandem repeats is the hardest genome part for assembly. Statistics of the identified tandem repeats are summarised in Supplementary Table 10.3.

Supplemental Table S10.3. Tandem repeat identification for both pangolin species.

	Malayan pangolin	Chinese pangolin
TRs all	922,360	381,505
TRs complex	4,401	2,510
TRs >1kb	1,800	5,701
TRs >3kb	546	1,351
TRs >10kb	81	43

11.0 SHORT NOTES ON PANGOLIN MUSCULOSKELETAL SYSTEM

Pangolins have unique adaptations to their musculoskeletal system. Firstly, pangolins have evolved and adapted for special muscles allowing them to close their ears and nostrils to protect them from the attack of insects and preventing ants and termites from escaping in mouth. Secondly, pangolins have extremely elongated and muscular tongues that can flicks in and out quickly for capturing the ants and termites. Thirdly, the toothless pangolins have strong stomach

muscles with keratinous spines to digest its food through strong muscle contraction. Fourthly, Pangolins can roll into tight balls with their strong muscles, when threatened by predators.

12.0 REFERENCES

Trevis: a tool for search and visualisation of large and complex tandem repeats at <https://github.com/ad3002/Trevis>.

Andrews S. FastQC:A quality control tool for high throughput sequence data at <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>.

Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R et al. 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience* **2**(1): 10.

Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sanchez Alvarado A, Yandell M. 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* **18**(1): 188-196.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular biology and evolution* **17**(4): 540-552.

Chan PP, Lowe TM. 2009. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic acids research* **37**(Database issue): D93-97.

Choi Y, Chan AP. 2015. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* **31**(16): 2745-2747.

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**(2): 80-92.

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**(18): 3674-3676.

De Bie T, Cristianini N, Demuth JP, Hahn MW. 2006. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**(10): 1269-1271.

Delcher AL, Salzberg SL, Phillippy AM. 2003. Using MUMmer to identify similar regions in large sequence sets. *Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]* **Chapter 10**: Unit 10 13.

Gokcumen O, Tischler V, Tica J, Zhu Q, Iskow RC, Lee E, Fritz MH, Langdon A, Stutz AM, Pavlidis P et al. 2013. Primate genome architecture influences structural variation mechanisms and functional consequences. *Proceedings of the National Academy of Sciences of the United States of America* **110**(39): 15764-15769.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* **29**(7): 644-652.

Griffiths-Jones S. 2006. miRBase: the microRNA sequence database. *Methods in molecular biology* **342**: 129-138.

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. 2008. miRBase: tools for microRNA genomics. *Nucleic acids research* **36**(Database issue): D154-158.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**(8): 1072-1075.

Huang Y, Niu B, Gao Y, Fu L, Li W. 2010. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**(5): 680-682.

Johnson LS, Eddy SR, Portugaly E. 2010. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11**: 431.
Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**(9): 1236-1240.
Jurka J. 2000. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet* **16**(9): 418-420.
Kadri S, Hinman V, Benos PV. 2009. HHMMiR: efficient de novo prediction of microRNAs using hierarchical hidden Markov models. *BMC Bioinformatics* **10 Suppl 1**: S35.
Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. 2011. Adaptive seeds tame genomic sequence comparison. *Genome Res* **21**(3): 487-493.
Ksepka DT, Parham JF, Allman JF, Benton MJ, Carrano MT, Cranston KA, Donoghue PC, Head JJ, Hermsen EJ, Irmis RB et al. 2015. The Fossil Calibration Database-A New Resource for Divergence Dating. *Systematic biology*.
Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**(4): 357-359.
Lechner M, Hernandez-Rosales M, Doerr D, Wieseke N, Thevenin A, Stoye J, Hartmann RK, Prohaska SJ, Stadler PF. 2014. Orthology detection combining clustering and synteny for very large datasets. *PloS one* **9**(8): e105015.
Li H, Coghlan A, Ruan J, Coin LJ, Heriche JK, Osmotherly L, Li R, Liu T, Zhang Z, Bolund L et al. 2006. TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic acids research* **34**(Database issue): D572-580.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16): 2078-2079.
Liu Y, Schroder J, Schmidt B. 2013. Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data. *Bioinformatics* **29**(3): 308-315.
Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research* **25**(5): 955-964.
Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**(1): 18.
Martin PS, and Richard G. Klein. 1989. *Quaternary extinctions: a prehistoric revolution*. University of Arizona Press.
McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**(9): 1297-1303.
Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**(9): 1061-1067.
Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America* **85**(8): 2444-2448.
R. A. FigTree at <http://tree.bio.ed.ac.uk/software/figtree>.
Rambaut A SM, Xie D & Drummond AJ. 2014. Tracer v1.6.
Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**(6): 276-277.
Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**(6): 863-864.
Simpson JT, Durbin R. 2012. Efficient de novo assembly of large genomes using compressed data structures. *Genome Res* **22**(3): 549-556.

Smit A, Hubley, R & Green, P. 2013-2015. RepeatMasker Open-4.0.
Vassetzky NS, Kramerov DA. 2002. CAN--a pan-carnivore SINE family. *Mammalian genome : official journal of the International Mammalian Genome Society* **13**(1): 50-57.
Vezzi F, Narzisi G, Mishra B. 2012. Reevaluating assembly evaluations with feature response curves: GAGE and assemblathons. *PloS one* **7**(12): e52210.
Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**(9): 1859-1875.
Wu Y, Wei B, Liu H, Li T, Rayner S. 2011. MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences. *BMC Bioinformatics* **12**: 107.
Xue W, Li JT, Zhu YP, Hou GY, Kong XF, Kuang YY, Sun XW. 2013. L_RNA_scaffolder: scaffolding genomes with transcripts. *BMC Genomics* **14**: 604.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution* **24**(8): 1586-1591.
Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**(5): 821-829.
Zhang Z, Carriero N, Zheng D, Karro J, Harrison PM, Gerstein M. 2006. PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics* **22**(12): 1437-1439.