Database, 2016, 1–10
doi: 10.1093/database/baw063
Original article

Original article

PGD: a pangolin genome hub for the research community

Tze King Tan[1,2], Ka Yun Tan[1,3], Ranjeev Hari[1,2], Aini Mohamed Yusoff[1,2], Guat Jah Wong[1], Cheuk Chuen Siow[1], Naresh V.R. Mutha[1], Mike Rayko[4], Aleksey Komissarov[4], Pavel Dobrynin[4], Ksenia Krasheninnikova[4], Gaik Tamazian[4], Ian C. Paterson[2,5], Wesley C. Warren[6], Warren E. Johnson[7], Stephen J. O'Brien[4,8] and Siew Woh Choo[1,2,9,*]

[1]Genome Informatics Research Laboratory, Centre for Research in Biotechnology for Agriculture (CEBAR), High Impact Research Building, University of Malaya, 50603 Kuala Lumpur, Malaysia, [2]Department of Oral and Craniofacial Sciences, Faculty of Dentistry, University of Malaya, 50603 Kuala Lumpur, Malaysia, [3]Institute of Biology Sciences, Faculty of Science, University of Malaya, 50603 Kuala Lumpur Malaysia, [4]Theodosius Dobzhansky Center for Genome Bioinformatics, Saint Petersburg State University, St. Petersburg 199004, Russia, [5]Oral Cancer Research and Coordinating Centre, Faculty of Dentistry, University of Malaya, 50603 Kuala Lumpur, Malaysia, [6]McDonnell Genome Institute, Washington University, St Louis, MO 63108, USA, [7]Smithsonian Conservation Biology Institute, Front Royal, Virginia 22630, USA, [8]Oceanographic Center, Nova Southeastern University, Ft Lauderdale, FL, 33004, USA and [9]Genome Solutions Sdn Bhd, Suite 8, Innovation Incubator UM, Level 5, Research Management & Innovation Complex, University of Malaya, 50603 Kuala Lumpur, Malaysia

*Corresponding author: Tel: +603-79676463; Email: l.choo@genomesolutions.com.my

Citation details: Tan,T.K., Tan,K.Y., Hari,R. *et al.* PGD: a pangolin genome hub for the research community. *Database* (2016) Vol. 2016: article ID baw063; doi:doi:10.1093/database/baw063

Received 24 October 2015; Revised 15 March 2016; Accepted 11 April 2016

Abstract

Pangolins (order Pholidota) are the only mammals covered by scales. We have recently sequenced and analyzed the genomes of two critically endangered Asian pangolin species, namely the Malayan pangolin (*Manis javanica*) and the Chinese pangolin (*Manis pentadactyla*). These complete genome sequences will serve as reference sequences for future research to address issues of species conservation and to advance knowledge in mammalian biology and evolution. To further facilitate the global research effort in pangolin biology, we developed the Pangolin Genome Database (PGD), as a future hub for hosting pangolin genomic and transcriptomic data and annotations, and with useful analysis tools for the research community. Currently, the PGD provides the reference pangolin genome and transcriptome data, gene sequences and functional information, expressed transcripts, pseudogenes, genomic variations, organ-specific expression data and other useful annotations. We anticipate that the PGD will be an invaluable platform for researchers who are interested in pangolin and mammalian research. We will continue updating this hub by including more data, annotation and analysis tools particularly from our research consortium.

Database URL: http://pangolin-genome.um.edu.my

Introduction

Pangolins are ancient creatures whose ancestors are thought to be members of a suborder of Palaeanodonta from some 60 million years ago (1). Pangolins are rare, toothless nocturnal burrowing mammals that are covered with tough, protective keratin scales over its whole body (2–4). These hard scales are used as a protective layer; when threatened, pangolins quickly roll up into a tight ball and the scales act as a shield. Pangolins are listed as endangered as reported in the IUCN red list (5) of species, because of deforestation and hunting for their meat is considered a delicacy and the scales, blood and skin are used in traditional Chinese medicine (6–8).

We have recently sequenced the genomes of two pangolin species, the Malayan pangolin (*Manis javanica*) and the Chinese pangolin (*Manis pentadactyla*) from Malaysia and Taiwan, respectively, using high-throughput next-generation sequencing (NGS) approaches. Using the NGS technology, we have also sequenced the transcriptomes of different organs of the Malayan pangolin. To further facilitate research into pangolin biology, we have implemented a Pangolin Genome Database (PGD) as a central hub for hosting genomic resources and information of pangolins, together with useful analysis tools for the research community. PGD provides interactive and user-friendly web interfaces for users to search, browse, retrieve, visualize and analyze pangolin genomic and transcriptomic data and information. The PGD will serve as a computational platform for researchers to advance research in this field, particularly in the area of conservation of this unique endangered mammalian species and also provide a useful resource for research in mammalian evolution and biology, in general.

Database Sources

PGD currently hosts all data, results and information about pangolins generated by the International Pangolin Research Consortium (IPaRC) (website: http://pangolin.um.edu.my). Currently, we have the reference genome sequences of Malayan pangolin and Chinese pangolin. The Malayan pangolin genome was sequenced using Illumina HiSeq 2000 at BGI, Hong Kong and assembled using SGA-0.10.10 (9) (coverage = 146×). The assembled contigs were scaffolded using SOAPdenovo2 scaffolder and achieved N50 of 204 525 bp, assembled scaffold genome size of 2.5 Gbp (41). For the Malayan pangolin, we have also sequenced the transcriptomes from different organs including cerebrum, cerebellum, liver, heart, lung and thymus. The sequenced Malayan pangolin specimen was provided by the Department of Wildlife and National Parks Malaysia and sequenced and assembled by the Genome Informatics Research Group (GIRG), University of Malaya, Malaysia. The Chinese pangolin specimen was originated from Taiwan and sequenced and assembled by the Warren Research Group from the Genome Institute of Washington University, USA, and analyzed by Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg, Russia. The Chinese pangolin genome (coverage = 56×) sequenced using Illumina platform, assembled using SOAPdenovo v1.0.5 and scaffolded using L_RNA_scaffolder approach achieving N50 of 157 892, assembled scaffold genome size of 2.2 Gbp (41).

Genome data and functional annotation

We annotated the entire genome of both pangolin species using a well-established MAKER pipeline, which was designed for the annotation of emerging model genomes using evidence from *ab initio* gene prediction, sequence homology approach and the transcriptomic data from Malayan pangolin (10). MAKER server is a well-established annotation pipeline for eukaryotic genomes annotation (10). First, repetitive sequences in the pangolin genomes were identified and masked using RepeatMasker (11) and RepeatModeller (12). After the repeat masking process, the RNA-Seq data from Malayan pangolin were mapped to the genomes as first evidence and the canine cDNA sequences from Ensembl (13) used as a reference for protein homology. This automated pipeline integrates the evidence from its gene prediction algorithm and produces high-quality gene models in subsequent runs. The gene models were *de novo* predicted by Augustus, GENSCAN and SNAP supported by MAKER annotation pipeline (10) using default parameters and canine cDNA sequences as reference, integrates with the homologs and RNA-Seq evidence to generate the final set of gene models. Using this stringent procedure, we identified 23 446 genes in the Malayan pangolin genome and 20 298 genes in the Chinese pangolin genome. These genes were further annotated using BLAST2GO, allowing better understanding of the function of each gene. BLAST2GO uses BLAST algorithm on the annotated protein sequence query to find homologs (14). Then, a straight

Table 1. Summary statistics of two pangolin genome and transcriptome datasets in PGD

Genome	Malayan Pangolin	Chinese Pangolin
Number of scaffolds	81,732	87,621
Estimated coverage (X)	146	56
Estimated Genome size	2,549,959,554 bp	2,205,289,822 bp
N50 (bp)	204,525	157,892
# of protein-coding genes	23,446	20,298
# of annotated genes	21,451 (91%)	19,287 (95%)
# of pseudogenes	4660	2416
# of transcripts	89,751	NA

Assembly statistic of pangolins genome. Adapted from Pangolin genomes and the evolution of mammalian scales and immunity. by Choo et al., 2016.

forward mapping is performed to retrieve the gene ontology (GO) term with the obtained hits. In order to annotate the functional domains/signatures, different biological databases were used such as GO modulate, Go Slim (15), Enzyme Code annotation with KEGG (16) maps and also InterPro (17) annotation with the default parameters. BLAST2GO assigned functions to 21 451 (91%) genes of Malayan pangolin and 19 287 (95%) genes of Chinese pangolin (Table 1).

Transcriptome data

PGD stores the sequences of expressed genes from different organs, namely the cerebellum, cerebrum, liver, heart, kidney, lung, thymus and spleen of the Malayan pangolin. Briefly, these transcriptomes were sequenced using an Illumina HiSeq 100 bp Paired End sequencing strategy. To generate a representative catalog of pangolin genes, the sequencing reads from all organs were pooled and *de novo* assembled using three different approaches: SOAPdenovo (18), Velvet (19) and Trinity (20) software. To generate a high-quality set of pangolin genes, we only accepted common assembled transcripts or genes predicted by the three different assemblers. The assembled data were filtered and clustered based on their similarity, and the longest transcripts were selected as unigene. This resulted in a set of 89 751 unigenes in the Malayan pangolin transcriptome (Table 1).

Pseudogene annotation data

To the best of our knowledge, the Malayan and Chinese pangolin genomes are the first whole-genome sequencing efforts for the Pholidota order; therefore, information regarding pseudogenes are still not available. Here, PGD offers information about the putative pangolin pseudogenes. To predict pseudogenes in the pangolin genomes, we searched the whole-genome sequences using MAKER-generated protein-coding genes (parent genes) as query sequences through the whole-genome using Pseudopipe pipeline (21). Pseudopipe is a well-established pipeline for whole-genome pseudogene screening (22–24). Protein sequences derived from MAKER annotation were used as queries to BLASTALL (25) through the whole pangolin genomes. After all homologous BLAST hits were reported, Pseudopipe removed all hits that belongs to its parent genes, and merged small gaps between two short hits and output as potential pseudogene candidates. These candidate pseudogenes were then screened using the tFasty (26) tool to report insertion, deletion, premature stop codon and frame shift mutation events that may disrupt the normal functions of the genes in order to generate the final pseudogene data set. Stringent thresholds were used to filter all the false-positive candidates (1E−10 e-value, 70% parent gene coverage, 40% gene identity). All pseudogenes were reported as duplicated or processed pseudogenes based on the nature of the reported hits. There are 4660 pseudogenes reported in Malayan pangolin and 2416 pseudogenes (Table 1) in the Chinese pangolin genome after applying the stringent filtering criteria.

PGD implementation

The PGD hub was developed based on the four-tier web application architecture (client workstation, web server, application server and database server) implemented under the Linux system using various common software packages including Apache, MySQL, PHP and Perl. The website was designed in PHP-HTML5 using Codelgniter and Twitter Bootstrap as the back-end and front-end frameworks, respectively (Figure 1). In addition, the website has been separated into logic, presentation and application data into three interconnected parts following the Model-view-controller framework. For the client-side, this was achieved with jQuery, which is a feature-rich JavaScript library that may enhance user interaction with the web pages through AJAX (Asynchronous JavaScript and XML) that is used to transfer data between the client side and the server side. The genomic annotation data that are generated from in-house scripts and published software are stored in the MySQL database management system.

PGD contains four main features/tabs: (i) ‘Browse’ tab for users to browse all pangolin genome annotation and transcripts information; (ii) ‘Tools’ tab that contains a list of analysis tools for users to analyze pangolin genomic data; (iii) ‘Genome Browser’ tab that incorporates dual genome browser, UCSC genome browser and JBrowse for visualizing pangolin genome sequences and annotation and (iv) ‘Download’ tab that allows users to download all pangolin genome/transcript data and annotations. In addition,

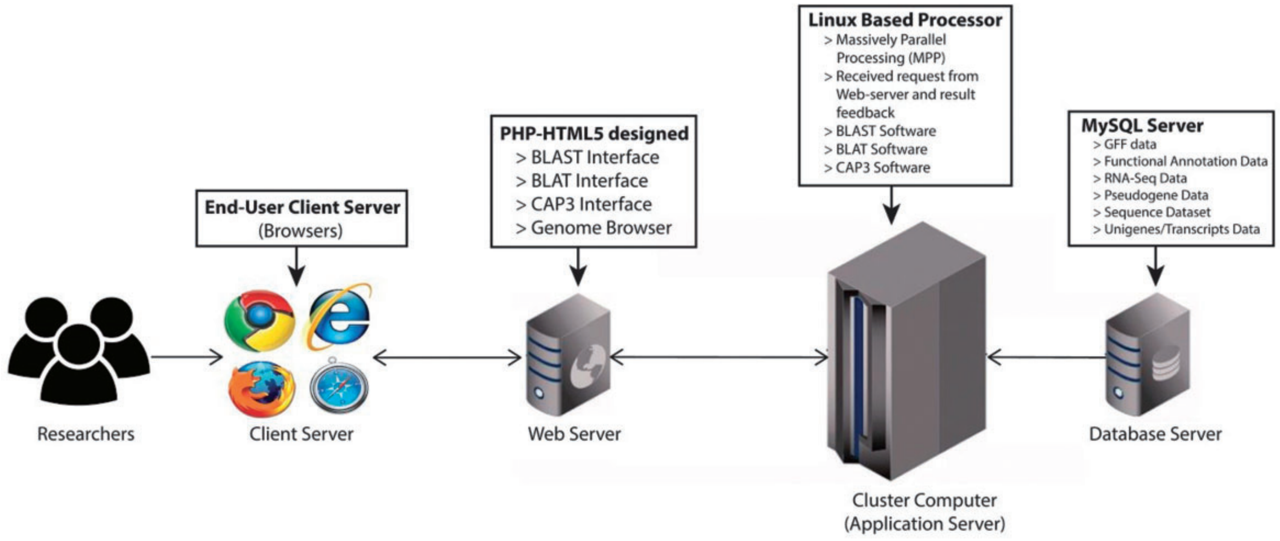


Figure 1. PGD four-tier web application architecture. (client workstation, web server, application server and database server).

PGD also has implemented quick search function, allowing users to rapidly search for genes of interest stored in the MySQL database. Furthermore, users can perform sequence searches against the pangolin genome sequences using their sequences of interest and the incorporated web-based BLAT and BLAST tools. Our web architecture allows users to access and submit their tasks using the front-end provided securely with the web server via the Internet. We normalized and optimized the database schema in order to reduce the data redundancy of pangolin's genomic data. Therefore, the database was well designed in such a way that MySQL would be performing optimally, without the loss of data integrity. As such, user can submit tasks, search, browse and retrieve the genomic data through PGD in a robust and efficient manner.

Graphical User Interface Design

Overview

The homepage of PGD contains general descriptions of pangolins in the main panel and manually compiled information about pangolins such as latest news and conferences, blogs and published articles in the right-side panel. On the top of the PGD homepage, several option tabs are provided, allowing users to access different features of PGD. For instance, the Browse tab allows users to quickly browse the detailed annotation data of Malayan and Chinese pangolins, as well as transcriptomic data for the Malayan pangolin. In addition, the Tools tab allows access to analysis tools such as BLAST (25) and BLAT (27) packages that we have incorporated into PGD. Users also can visualize the genome and genomic features using the unique real-time keyword search feature for fast and smooth searching of genes of interest (Figure 2).

Browsing pangolin genomic data in PGD

By clicking on the Browse tab on the top of the PGD homepage, users can access all gene information of the both pangolin species (Figure 3). All genes will be displayed after users click on the tab. Users can access the gene details page containing the gene annotation and functional information of a particular gene by clicking on the 'Details' button associated with the gene of interest. This gene details page will display all information about the gene of interest including sequence name, scaffold ID, start and stop position of the genes, putative functions, gene sequences and BLAST2GO alignment results such as 'Top Hit Species', 'Min E-value', annotated 'GOs' information, 'InterPro Scan' domain information. JBrowse is also incorporated in this page, allowing users to visualize and explore the locus and neighboring regions of the gene within the pangolin genome.

Keyword and sequence searches

PGD contains a vast amount of pangolin genomic, transcriptomic data and annotations. Therefore, it is necessary to provide intuitive web-based graphical user interface (GUI) allowing users to rapidly search a large volume of data. We have implemented a keyword search system in PGD to allow users to search genes of interest in a spontaneous manner. Alternatively, users can search by sequences. PGD has built-in BLAST (25) and BLAT (27) tools, which allows users to search or compare a query sequence against

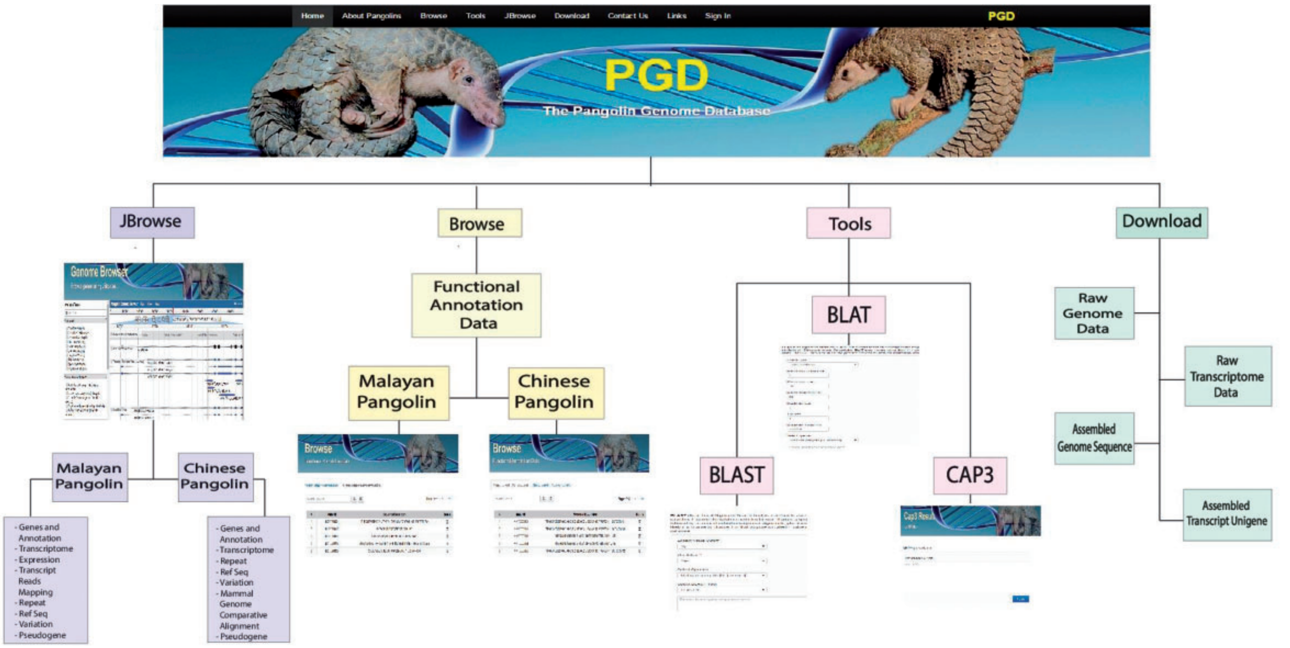


Figure 2. Schematic structure of the PGD.

the pangolin genome and transcriptomic data. Using the BLAST tools, users can perform nucleotide searches to each annotated sequence (BLASTN), whole-genome nucleotide searches (BLAST Whole Genome), protein searches of each annotated sequence (BLASTP), and nucleotide searches of each annotated protein sequence (BLASTX). Users can choose the pangolin species of interest (either Malayan pangolin or Chinese pangolin) and also set the desired cutoffs (e.g. expect value and enable to search for low compositional complexity regions) for their sequence searches. Alternatively, users can perform a sequence search using BLAT, which is structured differently from BLAST, to search the similarity in a query sequence but it accepts an exact or nearly exact match to find the hit. Unlike BLAST which is a local alignment software, BLAT is also designed to map transcripts/ESTs that have no introns back to the reference pangolin genomes. In this case, users would be able to examine the location of this transcript and also its gene structure within the genome. Using our provided web-based BLAST, users are able to set the parameters such as (i) minimum number of matches, (ii) minimum score, (iii) minimum identity percentage, (iv) maximum gap, (v) tile size and (vi) maximum intron size. The BLAT outputs will be linked with the genome browser, allowing instant visual comparison of each match hit.

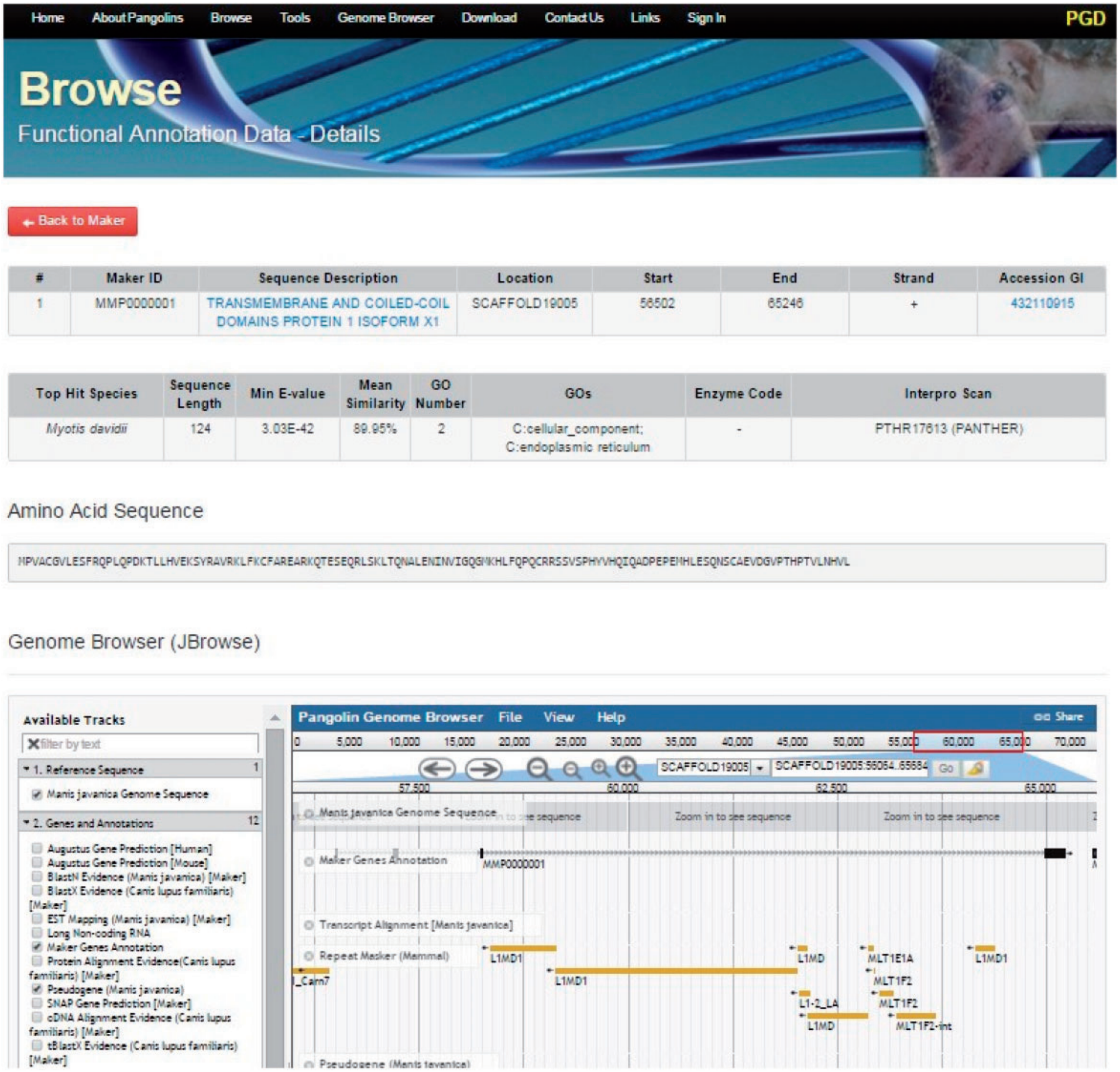
Interactive Pangolin Genome Data Visualization

A fully dynamic real-time genome browser, JBrowse, is incorporated into PGD, allowing users to instantly browse and visualize pangolin genome sequence and annotation data on the fly. JBrowse provides a few useful features: (i) supports fast and smooth genome navigation, (ii) utilizes multiple types of data format, (iii) provides real-time genome browsing with light server resource required, (iv) provides high-speed visualizing the data, sequences and genome annotation results.

Using the pangolin genome browser (JBrowse) (Figure 4), users can choose whether to visualize the pangolin data in Malayan pangolin genome or Chinese pangolin genome. Currently more annotations (e.g. transcriptomic data) are available for the Malayan pangolin than the Chinese pangolin, because many research projects are ongoing under the IPaRC for the Malayan pangolin species. Once users choose which genome to explore, they can see all relevant annotation tracks in the genome browser. For instance, we have a basic track as the reference genome sequence of the selected genome. We also have different annotation tracks that can be triggered on/off for easier visualization and analysis. These tracks are classified into different categories such as ‘Genes and Annotations’, ‘Repeat’, ‘Transcriptome’, ‘Transcript Reads Mapping’, ‘Organ Specific Expression’ and ‘Variation’ which represent our annotations from different analyses.

Genes and annotations tracks

This category will show the results of MAKER gene models including the evidence-based and *de novo* gene



#	Maker ID	Sequence Description	Location	Start	End	Strand	Accession GI
1	MMP0000001	TRANSMEMBRANE AND COILED-COIL DOMAINS PROTEIN 1 ISOFORM X1	SCAFFOLD19005	56502	65246	+	432110915

Top Hit Species	Sequence Length	Min E-value	Mean Similarity	GO Number	GOs	Enzyme Code	Interpro Scan
Myotis davidii	124	3.03E-42	89.95%	2	C:cellular_component; C:endoplasmic reticulum	-	PTHR17613 (PANTHER)

Amino Acid Sequence

```
MPVACGVLESFRQPLQPDKTLLHVEKSYRAVRKLFKCFAREARKQTESEQRLSKLTQNALENINVIGQGMKHLFQPQCRRSSVSPHYVHQIQADPEPEMHLESQNSCAEVDGVPTHPTVLNHVL
```

Genome Browser (JBrowse)

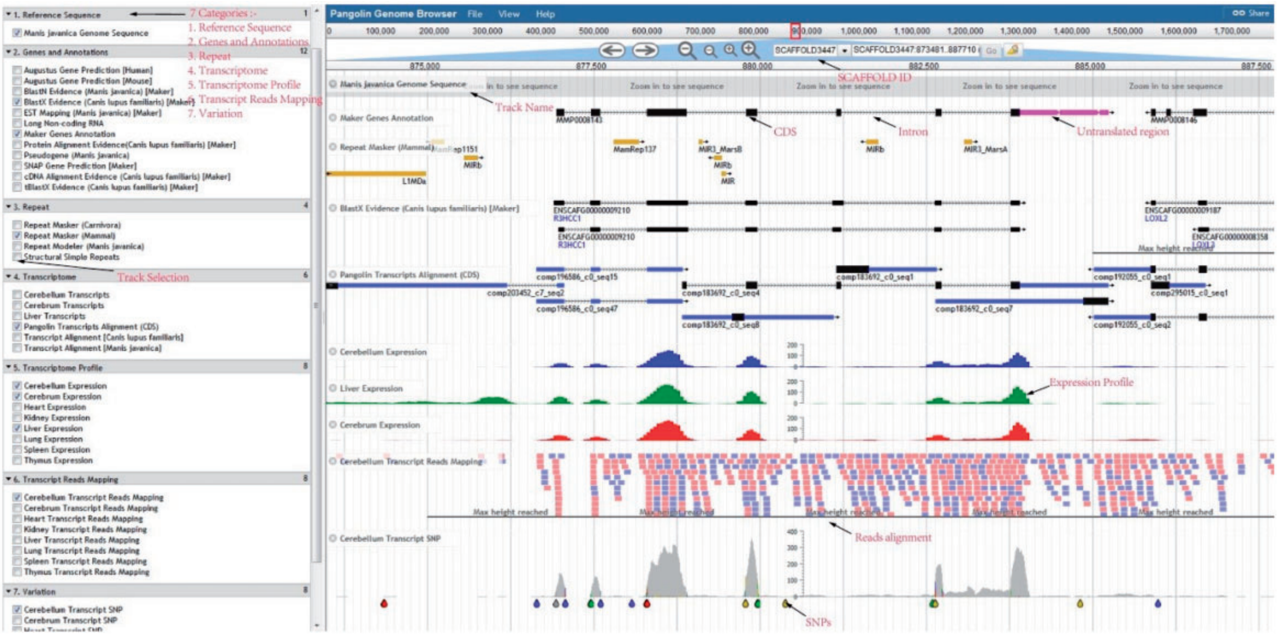
Available Tracks
filter by text
1. Reference Sequence 1
☑ Manis javanica Genome Sequence
2. Genes and Annotations 12
☐ Augustus Gene Prediction [Human]
☐ Augustus Gene Prediction [Mouse]
☐ BlastN Evidence (Manis javanica) [Maker]
☐ BlastX Evidence (Canis lupus familiaris) [Maker]
☐ EST Mapping (Manis javanica) [Maker]
☐ Long Non-coding RNA
☑ Maker Genes Annotation
☐ Protein Alignment Evidence(Canis lupus familiaris) [Maker]
☑ Pseudogene (Manis javanica)
☐ SNAP Gene Prediction [Maker]
☐ cDNA Alignment Evidence (Canis lupus familiaris) [Maker]
☐ tBlastX Evidence (Canis lupus familiaris) [Maker]
Pangolin Genome Browser
File
View
Help
Share
0
5,000
10,000
15,000
20,000
25,000
30,000
35,000
40,000
45,000
50,000
55,000
60,000
65,000
70,000
SCAFFOLD19005
SCAFFOLD19005:56054..65684
Go
57,500
60,000
62,500
65,000
Manis javanica Genome Sequence
Zoom in to see sequence
Maker Genes Annotation
MMP0000001
Transcript Alignment [Manis javanica]
Repeat Masker (Mammal)
L1MD1
L1MD
MLT1E1A
L1MD1
_Carn7
L1MD1
MLT1F2
L1-2_LA
MLT1F2
L1MD
MLT1F2-int
Pseudogene (Manis javanica)

Figure 3. A screenshot of gene details page. This page will display information of a gene including its sequences and functional annotation.

predictions, together with the predicted pseudogenes. In the pseudogene track, users can visualize the predicted pseudogenes in the selected pangolin genome. These pseudogenes are labeled with 'PSSD' and 'DUP' as categorized by PseudoPipe pipeline (21). The pseudogene tracks report all useful features found in the pseudogenes, including parent gene coverage 'Frac', identity 'Ident', insertion 'Ins', deletion 'Del', frameshift 'Shift' and stop codon 'Stop'. This information will provide a better understanding of pseudogene attributes found in both pangolins.

Repeat tracks

The repetitive element information is provided by different software which we display in three different tracks: RepeatMasker, Repeat Modeler and Structural Simple Repeat tracks. RepeatMasker track shows the repetitive elements screened in the pangolin genomes using two different repeat libraries: the carnivore repeat library and mammal repeat library. We used SSRIT (28) for scanning the simple sequence repeat or structural sequence repeat present in both pangolin genome. The SSRIT reports the



▾ 1. Reference Sequence 1
- ☑ Manis Javanica Genome Sequence

▾ 2. Genes and Annotations 12
- ☐ Augustus Gene Prediction [Human]
- ☐ Augustus Gene Prediction [Mouse]
- ☐ BlastN Evidence (Manis javanica) [Maker]
- ☑ BlastX Evidence (Canis lupus familiaris) [Maker]
- ☐ EST Mapping (Manis javanica) [Maker]
- ☐ Long Non-coding RNA
- ☑ Maker Genes Annotation
- ☐ Protein Alignment Evidence(Canis lupus familiaris) [Maker]
- ☐ Pseudogene (Manis javanica)
- ☐ SNAP Gene Prediction [Maker]
- ☐ cDNA Alignment Evidence (Canis lupus familiaris) [Maker]
- ☐ tBlastX Evidence (Canis lupus familiaris) [Maker]

▾ 3. Repeat 4
- ☐ Repeat Masker (Carnivora)
- ☑ Repeat Masker (Mammal)
- ☐ Repeat Modeler (Manis javanica)
- ☐ Structural Simple Repeats

▾ 4. Transcriptome 6
- ☐ Cerebellum Transcripts
- ☐ Cerebrum Transcripts
- ☐ Liver Transcripts
- ☑ Pangolin Transcripts Alignment (CDS)
- ☐ Transcript Alignment [Canis lupus familiaris]
- ☐ Transcript Alignment [Manis javanica]

▾ 5. Transcriptome Profile 8
- ☑ Cerebellum Expression
- ☑ Cerebrum Expression
- ☐ Heart Expression
- ☐ Kidney Expression
- ☑ Liver Expression
- ☐ Lung Expression
- ☐ Spleen Expression
- ☐ Thymus Expression

▾ 6. Transcript Reads Mapping 8
- ☑ Cerebellum Transcript Reads Mapping
- ☐ Cerebrum Transcript Reads Mapping
- ☐ Heart Transcript Reads Mapping
- ☐ Kidney Transcript Reads Mapping
- ☐ Liver Transcript Reads Mapping
- ☐ Lung Transcript Reads Mapping
- ☐ Spleen Transcript Reads Mapping
- ☐ Thymus Transcript Reads Mapping

▾ 7. Variation 8
- ☑ Cerebellum Transcript SNP
- ☐ Cerebrum Transcript SNP

SCAFFOLD3447 ▾ SCAFFOLD3447:873481..887710 Go

Figure 4. Pangolin genome browser. Users can turn on and off the annotation tracks in the left panel.

entire simple sequence repeats and allows user to study microsatellites in pangolin genomes.

Transcriptome tracks

Users can explore and visualize expressed genes in Malayan pangolin through the tracks under this category. The expressed transcripts generated by pooling all reads from the eight pangolin organs and assembled using three different assemblers are available here. These common transcripts were mapped onto the reference genome using GMAP (29). Moreover, users can also explore and examine which transcripts/genes are expressed in a specific pangolin organ such as cerebrum, cerebellum and liver through the provided organ-specific transcripts tracks.

Transcript reads mapping tracks

Users can also examine the expression level of each gene of Malayan pangolin across different organs. Briefly, for each organ or organ-specific transcriptome, we mapped RNA-Seq reads to the Malayan pangolin genome sequence and the reads mapping data of eight different organs (cerebrum, cerebellum, liver, heart, kidney, thymus, spleen and lung) were available in the 'Transcript Reads Mapping' category.

Transcriptome profile tracks

Raw transcript reads mapping results of eight different pangolin organs were normalized using RPKM value and converted into a graphical bigwig format track that visualized the expression profile of each organ. Users are able to examine the expression profiles of a gene of interest across eight different organs simultaneously. Of note, the PGD genome browser also links some features (e.g. genes and repeats) to external resources. Among of these resources are RepeatMasker database for each repeats elements, BLASTX evidence tracks is linked to well-established databases to check each hit on ENSEMBL database, PANTHER database (30), Dog Genome SNP database (31) and UniProt (32) by simply right clicking on the features in the tracks. These external database interlinks may help users understand further information of the features.

Table 2. Genome assembly version for each mammal genome used for multiple sequence alignment

Animal	Scientific name	Genome assembly
Dog	*Canis familiaris*	CanFam3.1
Cat	*Felis catus*	Felis_catus_8.0
Cow	*Bos taurus*	Bos_taurus_3.1
Horse	*Equus caballus*	EquCab_2.0
Human	*Homo sapiens*	GRCh37.p5
Mouse	*Mus musculus*	GRCm38.p4

Multiple mammalian genome structural alignment

We performed the multispecies structural alignment of two pangolin genomes to other mammalian genomes (dog, cat, cow, horse, human and mouse) using the Progressive Cactus software (33). The animal genome data used in the alignment are shown in Table 2.

Prior to performing the structural alignment procedure, repetitive regions of the assemblies were masked with RepeatMasker (34). The phylogenetic tree used to perform the alignment is given in Figure 5.

Pangolin genome conservation tracks

Basewise conservation scores were obtained for Malayan and Chinese pangolin genomes from the multispecies structural alignment using the HAL-phyloP tool from the HAL package (35). HAL-phyloP wraps the original phyloP algorithm (36) and is designed to efficiently process multispecies structural alignments by ProgressiveCactus.

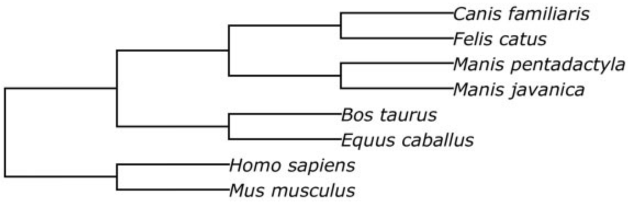
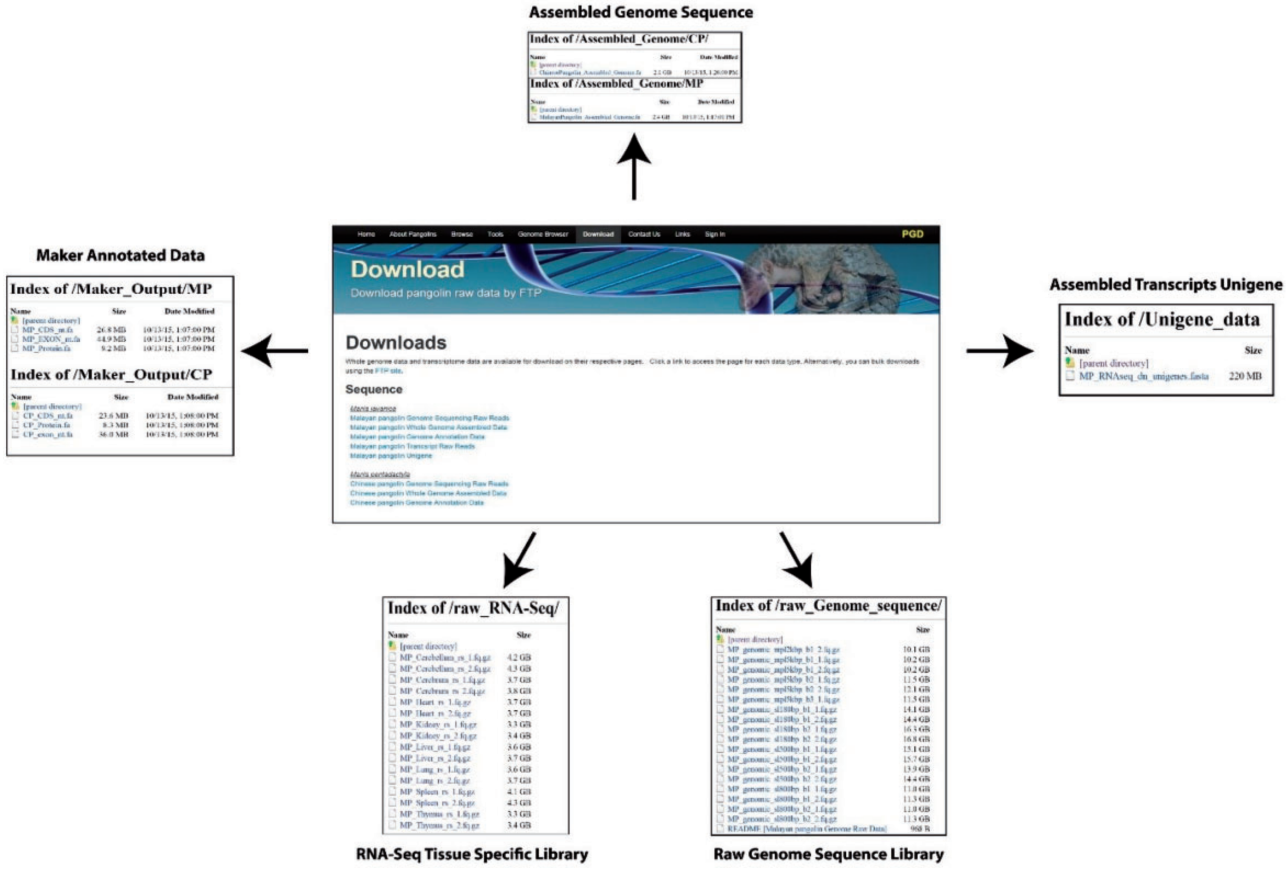


Figure 5. The phylogenetic tree of species involved into the structural alignment.

Coding sequences (CDSs) of Malayan pangolin genes were used to train a neutral evolution model for phyloP. Next, phyloP was launched in the conservation and acceleration ('CONACC') mode to obtain the corresponding *P* values for each position in both pangolin genomes. The common logarithms of the basewise *P* values were reported; positions under acceleration were flagged by making their values negative, whereas positions under conservation had their values positive.

Assembly hub for UCSC Genome Browser

The UCSC Genome Browser (37) is a well-known and widely used tool for bioinformatics analysis. Alongside with the PGD database, we provide an assembly hub (38) for the UCSC Genome Browser that contains datasets from PGD and provides means for viewing them in the genome browser. The hub also presents the structural alignment between genomes of both pangolin species and other mammals in the form of the snake tracks (39) and the genomic



Assembled Genome Sequence

Index of /Assembled_Genome/CP/

Name	Size	Date Modified
[parent directory]		
ChinesePangolin_Assembled_Genome.fa	2.1 GB	10/13/15, 1:26:00 PM

Index of /Assembled_Genome/MP

Name	Size	Date Modified
[parent directory]		
MalayanPangolin_Assembled_Genome.fa	2.4 GB	10/13/15, 1:17:01 PM

Maker Annotated Data

Index of /Maker_Output/MP

Name	Size	Date Modified
[parent directory]		
MP_CDS_nt.fa	26.8 MB	10/13/15, 1:07:00 PM
MP_EXON_nt.fa	44.9 MB	10/13/15, 1:07:00 PM
MP_Protein.fa	9.2 MB	10/13/15, 1:07:00 PM

Index of /Maker_Output/CP

Name	Size	Date Modified
[parent directory]		
CP_CDS_nt.fa	23.6 MB	10/13/15, 1:08:00 PM
CP_Protein.fa	8.3 MB	10/13/15, 1:08:00 PM
CP_exon_nt.fa	36.0 MB	10/13/15, 1:08:00 PM

Download

Download pangolin raw data by FTP

Downloads

Whole genome data and transcriptome data are available for download on their respective pages. Click a link to access the page for each data type. Alternatively, you can bulk downloads using the FTP site.

Sequence

Manis javanica
Malayan pangolin Genome Sequencing Raw Reads
Malayan pangolin Whole Genome Assembled Data
Malayan pangolin Genome Annotation Data
Malayan pangolin Transcript Raw Reads
Malayan pangolin Unigene

Manis pentadactyla
Chinese pangolin Genome Sequencing Raw Reads
Chinese pangolin Whole Genome Assembled Data
Chinese pangolin Genome Annotation Data

Assembled Transcripts Unigene

Index of /Unigene_data

Name	Size
[parent directory]	
MP_RNAseq_dn_unigenes.fasta	220 MB

Index of /raw_RNA-Seq/

Name	Size
[parent directory]	
MP_Cerebellum_rs_1.fq.gz	4.2 GB
MP_Cerebellum_rs_2.fq.gz	4.3 GB
MP_Cerebrum_rs_1.fq.gz	3.7 GB
MP_Cerebrum_rs_2.fq.gz	3.8 GB
MP_Heart_rs_1.fq.gz	3.7 GB
MP_Heart_rs_2.fq.gz	3.7 GB
MP_Kidney_rs_1.fq.gz	3.3 GB
MP_Kidney_rs_2.fq.gz	3.4 GB
MP_Liver_rs_1.fq.gz	3.6 GB
MP_Liver_rs_2.fq.gz	3.7 GB
MP_Lung_rs_1.fq.gz	3.6 GB
MP_Lung_rs_2.fq.gz	3.7 GB
MP_Spleen_rs_1.fq.gz	4.1 GB
MP_Spleen_rs_2.fq.gz	4.3 GB
MP_Thymus_rs_1.fq.gz	3.3 GB
MP_Thymus_rs_2.fq.gz	3.4 GB

RNA-Seq Tissue Specific Library

Index of /raw_Genome_sequence/

Name	Size
[parent directory]	
MP_genomic_mpl2kbp_b1_2.fq.gz	10.1 GB
MP_genomic_mpl5kbp_b1_1.fq.gz	10.2 GB
MP_genomic_mpl5kbp_b1_2.fq.gz	10.2 GB
MP_genomic_mpl5kbp_b2_1.fq.gz	11.5 GB
MP_genomic_mpl5kbp_b2_2.fq.gz	12.1 GB
MP_genomic_mpl5kbp_b3_1.fq.gz	11.5 GB
MP_genomic_sl180bp_b1_1.fq.gz	14.1 GB
MP_genomic_sl180bp_b1_2.fq.gz	14.4 GB
MP_genomic_sl180bp_b2_1.fq.gz	16.3 GB
MP_genomic_sl180bp_b2_2.fq.gz	16.8 GB
MP_genomic_sl500bp_b1_1.fq.gz	15.1 GB
MP_genomic_sl500bp_b1_2.fq.gz	15.7 GB
MP_genomic_sl500bp_b2_1.fq.gz	13.9 GB
MP_genomic_sl500bp_b2_2.fq.gz	14.4 GB
MP_genomic_sl800bp_b1_1.fq.gz	11.0 GB
MP_genomic_sl800bp_b1_2.fq.gz	11.3 GB
MP_genomic_sl800bp_b2_1.fq.gz	11.0 GB
MP_genomic_sl800bp_b2_2.fq.gz	11.3 GB
README [Malayan pangolin Genome Raw Data]	968 B

Raw Genome Sequence Library

Figure 6. Web interfaces for data download in PGD.

conservation tracks obtained from the alignment with the HAL-phyloP tool.

In addition to presenting the pangolin genome annotation tracks, the hub provides access to other features and tools available in the UCSC Genome Browser, including Table Browser (40) for manipulating the tracks and obtaining nucleotide sequences of the features.

Data download

PGD provides a user-friendly interface for downloading pangolin genome and transcriptomic data and annotations. The raw sequencing reads from different library sizes are available for download. Users also can download the assembled reference genome sequences of the Malayan and Chinese pangolin for downstream analyses. Other available data or annotations are available for download including the MAKER-generated gene sequences (nucleotide CDS, exons and protein sequences), expressed genes from Malayan pangolin transcriptomes and the raw RNA-Seq data.

There are two ways whereas users can download these data and annotations. First, users can download these data through the ‘Download’ page using the provided web interface. Users can select the data or annotations of interest to download. Second, users can download the raw data and annotations through the File Transfer Protocol interface provided in the download page (Figure 6).

Conclusion

We anticipate that the PGD will become an invaluable hub, allowing researchers to access, browse, retrieve and analyze pangolin genomic and expression data and annotations. This hub would facilitate research in pangolin biology, particularly in the conservation of this critically endangered species and will also enhance our understanding of mammalian biology and evolution. We will continue updating PGD by incorporating more data, annotations and analysis tools particularly from our IPaRC consortium as it becomes available. We also welcome researchers to provide suggestions and/or share data for the improvement of the PGD hub.

Availability

PGD is accessible at http://pangolin-genome.um.edu.my. The UCSC Genome Browser hub configuration file is available at http://public.dobzhanskycenter.ru/PangolinHub/hub.txt. Users can download all the raw sequences and datasets used in this paper from the PGD website.

Acknowledgements

We appreciate all members of GIRG for the helpful discussion and valuable comments. We also want to thank Mark Diekhans, Benedict Paten, Joel Armstrong, Glenn Hickey and Brian Raney of the UCSC Genomics Institute for the support in use of ProgressiveCactus tool and the HAL package. Also thank to Tan Shi Yang for providing IT and bioinformatics supports in this project.

Funding

This project was supported by University of Malaya and Ministry of Education, Malaysia under the High Impact Research (HIR) grant UM.C/625/HIR/MOHE/CHAN-08.

References

1. Kondrashov,P. and Agadjanian,A.K. (2012) A nearly complete skeleton of ernanodon (Mammalia, Palaeanodonta) from Mongolia: morphofunctional analysis. *J. Vertebr. Paleontol.*, 32, 983–1001.
2. Kawashima,T., Thorington,R.W., Bohaska,P.W. *et al.* (2015) Anatomy of shoulder girdle muscle modifications and walking adaptation in the scaly Chinese Pangolin (*Manis Pentadactyla* Pentadactyla: Pholidota) compared with the partially Osteoderm-Clad Armadillos (Dasypodidae). *Anat. Rec.*, 298, 1217–1236.
3. Lin,M.F., Chang,C.Y., Yang,C.W. *et al.* (2015) Aspects of digestive anatomy, feed intake and digestion in the Chinese pangolin (*Manis Pentadactyla*) at Taipei zoo. *Zoo Biol.*, 34, 262–270.
4. Chin,S.C., Lien,C.Y., Chan,Y.T. *et al.* (2015) Hematologic and Serum Biochemical Parameters of Apparently Healthy Rescued Formosan Pangolins (*Manis Pentadactyla* Pentadactyla). *J. Zoo Wildlife Med.*, 46, 68–76.
5. IUCN 2015. The IUCN Red List of Threatened Species. Version 2015-4. <http://www.iucnredlist.org>.
6. Guo,Y., Zou,X., Chen,Y. *et al.* (1997) Sustainability of wildlife use in traditional Chinese medicine. *Conserv. China's Biodivers.*, 190–220.
7. Zhou,Z.M., Zhou,Y., Newman,C. *et al.* (2014) Scaling up pangolin protection in China. *Front. Ecol. Environ.*, 12, 97–98.
8. Luo,J., Yan,D., Song,J. *et al.* (2013) A strategy for trade monitoring and substitution of the organs of threatened animals. *Sci. Rep.*, 3,
9. Simpson,J.T. and Durbin,R. (2012) Efficient de novo assembly of large genomes using compressed data structures. *Genome Res.*, 22, 549–556.
10. Cantarel,B.L., Korf,I., Robb,S.M. *et al.* (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.*, 18, 188–196.
11. Shirak,A., Grabherr,M., Di Palma,F. *et al.* (2010) Identification of repetitive elements in the genome of *Oreochromis niloticus*: Tilapia Repeat Masker. *Mar. Biotechnol.*, 12, 121–125.
12. Smith,C.D., Edgar,R.C., Yandell,M.D. *et al.* (2007) Improved repeat identification and masking in Dipterans. *Gene*, 389, 1–9.
13. Cunningham,F., Amode,M.R., Barrell,D. *et al.* (2015) Ensembl 2015. *Nucleic Acid Res.*, 43, D662–D669.

14. Conesa,A., Götz,S., García-Gómez,J.M. *et al.* (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21, 3674–3676.
15. Ashburner,M., Ball,C.A., Blake,J.A. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, 25, 25–29.
16. Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acid Res.*, 28, 27–30.
17. Hunter,S., Apweiler,R., Attwood,T.K. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acid Res.*, 37, D211–D215.
18. Xie,Y.L., Wu,G.X., Tang,J.B. *et al.* (2014) SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, 30, 1660–1666.
19. Zerbino,D.R., Birney,E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, 18, 821–829.
20. Haas,B.J., Papanicolaou,A., Yassour,M. *et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.*, 8, 1494–1512.
21. Zhang,Z.L., Carriero,N., Zheng,D.Y. *et al.* (2006) PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics*, 22, 1437–1439.
22. Sisu,C., Pei,B.K., Leng,J. *et al.* (2014) Comparative analysis of pseudogenes across three phyla. *Proc. Natl Acad. Sci. USA*, 111, 13361–13366.
23. Pei,B.K., Sisu,C., Frankish,A. *et al.* (2012) The GENCODE pseudogene resource. *Genome Biol.*, 13,
24. Karro,J.E., Yan,Y.P., Zheng,D.Y. *et al.* (2007) Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. *Nucleic Acids Res.*, 35, D55–D60.
25. Altschul,S.F., Gish,W., Miller,W. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.* , 215, 403–410.
26. Pearson,W.R. (1999) Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.* 132, 185–219.
27. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, 12, 656–664.
28. Temnykh,S., DeClerck,G., Lukashova,A. *et al.* (2001) Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): Frequency, length variation, transposon associations, and genetic marker potential. *Genome Res.*, 11, 1441–1452.
29. Wu,T.D., Watanabe,C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21, 1859–1875.
30. Thomas,P.D., Campbell,M.J., Kejariwal,A. *et al.* (2003) PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res.*, 13, 2129–2141.
31. Bai,B., Zhao,W.M., Tang,B.X. *et al.* (2015) DoGSD: the dog and wolf genome SNP database. *Nucleic Acid Res.*, 43, D777–D783.
32. Chan,W.M., UniProt Consortium. (2010) The UniProt Knowledgebase (UniProtKB): a freely accessible, comprehensive and expertly curated protein sequence database. *Genet. Res.*, 92, 78–79.
33. Paten,B., Diekhans,M., Earl,D. *et al.* (2011) Cactus graphs for genome comparisons. *J. Computat. Biol.*, 18, 469–481.
34. Smit,A.F.A., Hubley,R., Green, P. (2004). RepeatMasker Open-3.0. 2004. Seattle (WA): Institute for Systems Biology.
35. Hickey,G., Paten,B., Earl,D. *et al.* (2013) HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics.*, 29, 1341–1342.
36. Pollard,K.S., Hubisz,M.J., Rosenbloom,K.R. *et al.* (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, 20, 110–121.
37. Kent,W.J., Sugnet,C.W., Furey,T.S. *et al.* (2002) The human genome browser at UCSC. *Genome Res.*, 12, 996–1006.
38. Raney,B.J., Dreszer,T.R., Barber,G.P. *et al.* (2014) Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*, 30, 1003–1005.
39. Nguyen,N., Hickey,G., Raney,B.J. *et al.* (2014) Comparative assembly hubs: web-accessible browsers for comparative genomics. *Bioinformatics*, 30, 3293–3301.
40. Karolchik,D., Hinrichs,A.S., Furey,T.S. *et al.* (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acid Res.*, 32, D493–D496.
41. Choo,S.W., Rayko,M., Tan,T.K. *et al.* (2016) Pangolin genomes and the evolution of mammalian scales and immunity. *Genome Research.*, gr-203521.