

RESEARCH ARTICLE

Open Access



A chromosome-level genome assembly of the Hispid cotton rat (*Sigmodon hispidus*), a model for human pathogenic virus infections

Jingtao Lilue^{1,2,3†}, André Corvelo^{4†}, Jèssica Gómez-Garrido^{5,6†}, Keiko Akagi^{7,8†}, Fengtang Yang^{1,19†}, Gia Green⁹, Bee Ling Ng¹⁰, Beiyuan Fu¹⁰, Uciel Pablo Chorostecki^{11,12}, Sarah C. Warner^{13,20}, Marina Marcet-Houben^{11,12}, Thomas M. Keane¹, James C. Mullikin¹⁴, Tyler Alioto^{5,6}, Toni Gabaldón^{11,12,15,16,17}, Benjamin Hubert^{4,21}, David E. Symer^{7,18,22,23*†} and Stefan Niewiesk^{9*†}

Abstract

Background The cotton rat (*Sigmodon hispidus*), a rodent species native to the Americas, has emerged as a valuable laboratory model of infections by numerous human pathogens including poliovirus and respiratory syncytial virus (RSV).

Results Here we report the first reference assembly of the cotton rat genome organized at a chromosomal level, providing annotation of 24,878 protein-coding genes. Data from PCR-free whole genome sequencing, linked-read sequencing, and RNA sequencing from pooled cotton rat tissues were analyzed to assemble and annotate this novel genome sequence. Spectral karyotyping data using fluorescent probes derived from mouse chromosomes facilitated the assignment of cotton rat orthologs to syntenic chromosomes, comprising 25 autosomes and a sex chromosome in the haploid genome. Comparative phylome analysis revealed both gains and losses of numerous genes including immune defense genes against pathogens. We identified thousands of recently retrotransposed L1, SINE B2, and ERV elements, revealing widespread genomic insertions unique to this species.

Conclusions We anticipate that annotation and characterization of the first chromosome-level cotton rat genome assembly as described here will enable and accelerate ongoing investigations into its host defenses against viral and other pathogens, genome biology, and mammalian evolution.

Keywords Cotton rat, Reference genome assembly, Whole genome sequencing, Linked-read sequencing, Gene annotations, Rodent, RNA-seq, Comparative genomics, Transposable elements, Phylome analysis

[†]Jingtao Lilue, André Corvelo, Jèssica Gómez-Garrido, Keiko Akagi and Fengtang Yang Joint first authors.

[†]David E. Symer and Stefan Niewiesk joint senior authors.

*Correspondence:

David E. Symer
desymer@gmail.com
 Stefan Niewiesk
niewiesk.1@osu.edu

Full list of author information is available at the end of the article



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Background

Cotton rats (*Sigmodon hispidus*) are small rodents extant in the New World. They live in the southern part of the USA, Central America, and the northern part of South America [1–3]. These rodents have been used as laboratory animals since the 1930s, when in pursuit of a non-primate animal model they were infected with poliovirus [4]. The cotton rat subsequently has been found to be susceptible to a large number of human pathogens, and has been used in studies of the pathogenesis of human respiratory viral infections and in testing of vaccines and antivirals [2, 3]. Many human pathogens replicate well in cotton rat but not in house mouse (*Mus musculus*), including RSV, human metapneumovirus, influenza virus, and adenovirus. RSV replicates approximately one 100-fold more in cotton rats, and as a consequence, model studies to test the efficacy of vaccines in humans provide a better predictive value in cotton rat than in mouse [5–7]. The cotton rat is a natural carrier of additional respiratory viruses with tropism for humans, including the Black Creek Canal virus strain of hantavirus. Moreover, other viral pathogens infecting the respiratory tract in humans do not replicate at all in mouse, including parainfluenza virus and measles virus [3].

A comprehensive approach to the analysis of overall immune responses to infection and vaccination in cotton rats requires measurements of gene expression profiles. However, the lack of a cotton rat reference genome has limited such experiments. In recent studies, immune responses of RSV-infected cotton rats were evaluated by comparing RNA expression patterns in infected versus uninfected lungs [8, 9]. As in humans, numerous immune genes related to defense against viral infections were upregulated, while others were downregulated late during infection. Although these results indicate broad similarities with gene expression changes in humans as quantified by microarray [10, 11], additional cotton rat genes appeared novel, as no known orthologs have been identified in other species.

We anticipated that development of an annotated cotton rat reference genome would facilitate future investigations into the complex interplay between infectious pathogens and host immune responses. In addition, a high-quality, annotated reference assembly would enable comprehensive comparisons of immune responses across species such as between cotton rat and human, for example through the development of agnostic assays of cotton rat gene expression including RNA-seq, single-cell RNA-seq and proteomics. To develop such an annotated cotton rat genome assembly, we sequenced cotton rat DNA and pooled RNA isolated from several tissues. We undertook a cost-effective approach by combining whole genome sequencing (WGS), linked-read sequencing to

obtain long-range genomic data [12, 13], RNA sequencing (RNA-seq), and spectral karyotyping data. Our analysis defined the cotton rat genome length and chromosomes, annotated genes including immune genes and their orthologs, identified active, novel retrotransposon families, and assessed phylogeny. The resulting draft reference genome assembly of the cotton rat has highlighted a few dozen candidate genes for their potential roles in host defenses against various viruses. We anticipate that this new, annotated cotton rat genome assembly will facilitate future research on pathogenic infections, host immunity, genome biology, mammalian evolution, and potentially additional research fields.

Results

Genome assembly

Genomic DNA was extracted from a male cotton rat. To prepare a linked-read library (10× Genomics Chromium), high molecular weight DNA fragments were partitioned and barcoded. This linked-read library was sequenced (Illumina 2×150 nucleotide [nt] paired-end reads) at 72× depth of coverage. In addition, a PCR-free whole genome sequencing (WGS) library with a mean insert length of ~350 nt was prepared and sequenced at 38× depth of coverage (Illumina). A second PCR-free WGS library was made with a mean insert length of ~550 nt and sequenced at 40× depth of coverage.

To assemble a high-quality draft of the cotton rat genome, we combined the linked-read sequencing data and WGS data from the two PCR-free libraries. First, error-corrected linked reads were assembled using Supernova v2.0.1 [14], resulting in a draft pseudo-haploid representation of the genome (Table 1). We used ARKS v1.0.2 [15] to increase scaffold lengths of the draft further, resulting in a ~2.50 Gb assembly. This had an improved scaffold N50 length of 4.95 megabasepairs (Mb) [12]. Assembly gaps were closed by using error-corrected PCR-free short-read data in two complementary strategies (see [Methods](#)). This led to a significant increase in contig N50 lengths which improved from the original 62.8 kilobasepairs (kb) to 360.5 kb in the resulting assembly, which we term HispidOSU (Table 1). This assembly has been deposited at National Center for Biotechnology Information (NCBI) under GenBank accession number GCA_041721565.1 [16]. The HispidOSU assembly [16] shows a high level of sequence completeness, with 99.28% of PCR-free library reads mapping to it, and 98.69% aligned in proper pairs. This high level of completeness also is corroborated by a high degree of observed gene-space completeness, with 242 out of 248 (97.6%) core eukaryotic genes and 8919 out of 9226 (96.7%) mammalian ortholog groups identified in their

Table 1 *S. hispidus* genome assembly statistics

Assembly metric	Supernova v2.0.1	Supernova v2.0.1 + ARKS	HispidOSU Supernova v2.0.1 + ARKS + gap closure + redundancy removal GenBank: GCA_041721565.1	SigHis_v1_BIUU GenBank: GCA_004025045.1
Length (bp)	2,514,336,970	2,514,367,750	2,504,610,684	2,730,600,022
Scaffolds (N)	23,680	20,602	17,569	883,546
Longest scaffold (bp)	7,913,074	24,396,235	24,371,482	1,032,205
Scaffold N50 length (bp)	1,572,529	4,921,260	4,950,377	101,373
Scaffold N90 length (bp)	209,288	746,302	784,357	1886
Contigs (N)	87,086	87,086	31,279	897,290
Longest contig (bp)	520,409	520,409	2,485,181	783,669
Contig N50 length (bp)	62,801	62,801	360,531	67,983
Contig N90 length (bp)	16,655	16,655	78,531	1879
Gaps (N)	63,415	66,492	13,710	13,744
Gapped length (bp)	29,394,360	29,425,140	23,026,440	1,374,400

complete forms, as determined by CEGMA [17] and BUSCO [18], respectively.

Another independent genome sequence dataset for *S. hispidus* recently was deposited and released by the NCBI, i.e., SigHis_v1_BIUU (abbreviated as BIUU; GenBank accession number GCA_004025045.1) [19]. This earlier BIUU assembly was based exclusively on short-read WGS data. To compare various features of this previous dataset with our HispidOSU assembly [16], we tabulated several of its reported parameters, including total sequenced genome length, number of scaffolds, scaffold N50 length, and number of gaps (Table 1). Its depth of sequencing coverage genomewide was 43.2×, substantially less than the cumulative coverage of ~150× reached in our study. This BIUU assembly has an approximately 50-fold higher count of single-contig scaffolds, and its scaffold N50 lengths are approximately 40-fold lower than in our assembly (Table 1). Its cumulative gap length is less than that in our genome assembly, but this finding is likely attributable to an arbitrary assumption of only 100 nt per gap in that assembly. To our knowledge, chromosomes have not been assigned in the BIUU assembly.

Cross-species chromosome painting and assembly

To assign draft chromosome models for the *S. hispidus* assembly, we performed comparative chromosome painting of metaphasespreads, using probes derived from flow-sorted autosomes and the X chromosome of the mouse, *Mus musculus* (Additional file 1: Fig. S1). *S. hispidus* chromosomes were identified and numbered based on G-banded metaphases [20]. As expected, *S. hispidus* shows extensive synteny when compared with the mouse chromosomes over long genomic distances (Additional

file 1: Fig. S2) [21]. Using the mouse-derived chromosome paints, a strong, continuous signal was observed across most portions of *S. hispidus* chromosomes. However, some small regions of the chromosomes were ambiguous, either displaying multiple probes derived from multiple mouse chromosomes (e.g. genomic repeats) or no probe signal. This result confirmed the accumulation of considerable chromosome-level differences between the two rodent species, despite very high overall levels of orthology and synteny that were observed.

We assembled the de novo draft cotton rat genome sequence scaffolds into 26 pseudo-chromosomes, based on chromosome paints data and on their synteny and orthologous sequence similarities with mouse. The set of pseudo-chromosome sequences contains 803 scaffolds, with a total length of 2.359 M basepairs covering approximately 92.1% of the draft haploid de novo cotton rat assembly. These data indicate that the karyotype of *S. hispidus* is similar to an ancestral karyotype of the genus *Sigmodon*, which harbors 2n=52 chromosomes (i.e., counting both autosomes and sex chromosomes). This finding also corroborates a previous report indicating that the cotton rat genome has accumulated few or no changes in chromosome counts or composition when compared with five other *Sigmodon* species (*S. hirsutus*, *S. leucotis*, *S. ochrognathus*, *S. peruanus*, and *S. toltecus*) [22].

The orthologous sequences and synteny among distinct rodent species were corroborated further upon comparisons with previously reported chromosome assignments in a New World mouse species, *Onychomys torridus* (*O. torridus*, also known as the Southern grasshopper mouse), and in the white-footed mouse (*Peromyscus leucopus*) [23]. These findings were based

on similar methods (Additional file 1: Fig. S3, Additional file 1: Fig. S4, Fig. 1). The results demonstrate a high degree of interspecific relatedness when comparing cotton rat vs. these independent rodent species.

Cotton rat genes

Combining outputs from three ab initio gene-calling programs (see Methods), we identified 24,878 protein-coding genes across the cotton rat genome assembly. From these genes, we conservatively predicted expression of a total of 29,010 transcripts (i.e., 1.17 transcripts per gene), while ignoring most other potential isoforms arising

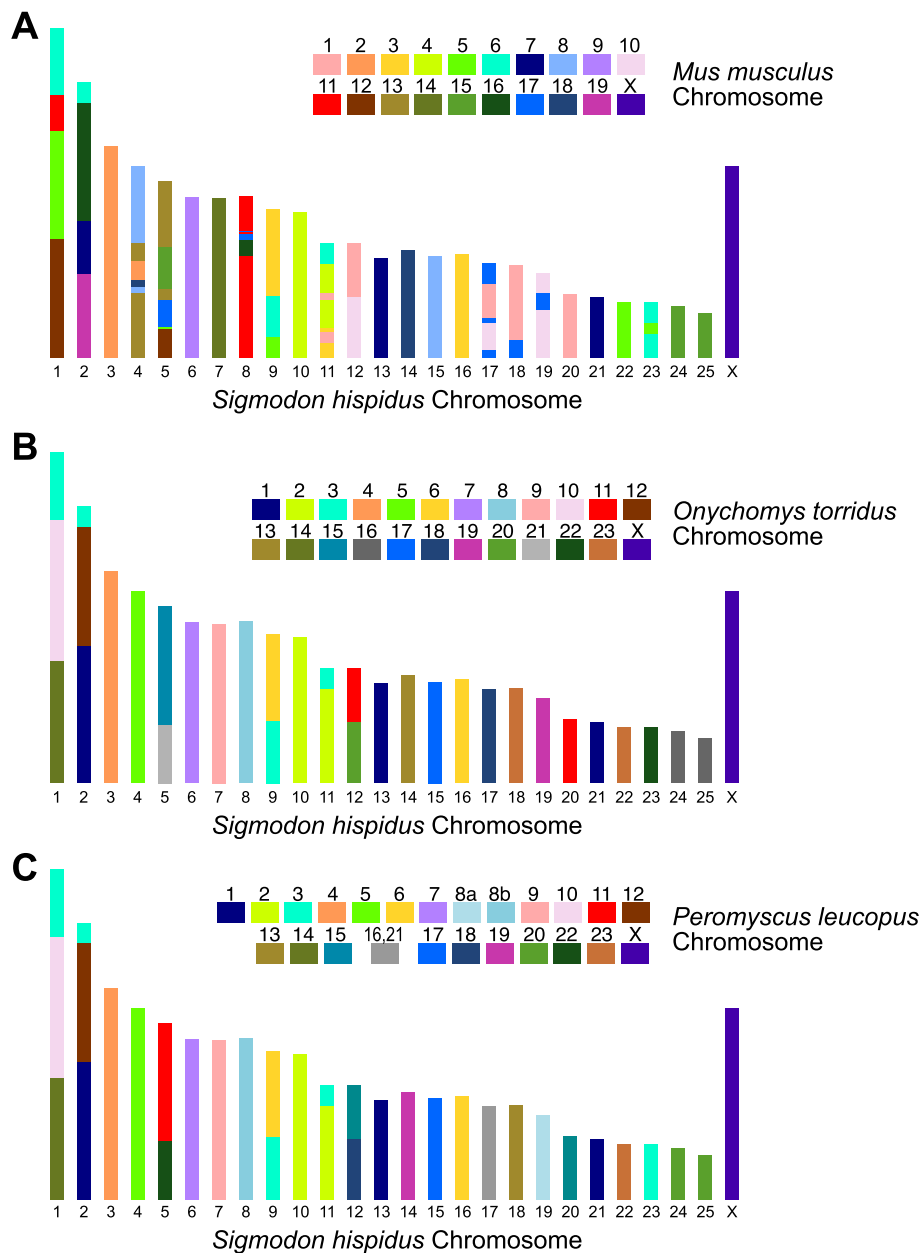


Fig. 1 Comparison of pseudo-chromosome assemblies of cotton rat (*S. hispidus*), determined experimentally with mouse chromosome paints or upon alignments with *O. torridus* and *P. leucopus* orthologous segments. Schematics represent karyotypes from metaphase spreads of cotton rat, **A** painted with probes derived from *M. musculus* chromosomes (key) or annotated from alignments with syntenic segments from annotated pseudo-chromosomes of **B** *O. torridus* (probed with mouse chromosome paints, cf. Methods) and **C** *P. leucopus* [23]. Color codes on *S. hispidus* chromosomes are based on *Mus musculus*

from alternative splicing. In turn, these transcripts are predicted to encode 28,403 unique protein products. We assigned functional labels to 90.8% of the predicted gene products. On average, each annotated gene contains 9.46 exons, with 75% of transcripts predicted as multi-exonic (Table 2). In addition, we predicted 40,481 non-coding transcripts, of which 23,669 and 16,812 appear to be expressed from long and short non-coding RNA genes, respectively.

We sought to obtain experimental evidence confirming that a large portion of the predicted cotton rat transcriptome is expressed. In addition, concordance between expressed transcript structures and our ab initio gene predictions needed to be checked. For these reasons, we performed RNA-seq on two pools of total RNAs. RNA was isolated from multiple tissues dissected from two adult male cotton rats, it was pooled to represent roughly similar concentrations across the tissues from each individual, and finally two RNA-seq libraries were prepared and sequenced. One of the cotton rat individuals was untreated and healthy (naive), while the other was injected intraperitoneally with house dust mite (HDM) antigen, followed by intranasal exposure after eight days and subsequent euthanasia four days thereafter. RNA-seq data from each individual were aligned using STAR

v-2.5.3a [24] against HispidOSU [16] (Table 3; Additional file 1: Table S1).

At least one RNA-seq transcript read from either pool could be assigned to each of 21,417 genes, which comprise 86.1% of the predicted protein-coding genes (Table 3; Additional file 1: Fig. S5). These included 911 genes expressed in the naive individual tissue pool, and 583 in the other pool from the individual exposed to house dust mite (HDM; Additional file 1: Fig. S5). To assess if annotations of our reference genome included immune response genes that were expressed in these pools, first, we selected 224 genes identified in our reference assembly (Additional File 1: Table S2) that were annotated by the specific gene ontology (GO) term “immune response” (GO:0006955). Of these, 213 (95%) were found to be expressed in our RNA-seq data (including 208 expressed genes in the naive dataset and 207 in the HDM one, out of the total counts in Table 3). The 11 remaining “immune response” genes lacking transcripts detected in our RNA-seq pools were tabulated as well (Additional file 1: TableS3).

A recent report identified host response transcripts that were differentially expressed upon infection with RSV [8]. We downloaded a list of their cDNA sequences and aligned them against our reference genome assembly. Of 19 differentially expressed genes that also were annotated and confirmed in our reference assembly, we detected expression for 15 (79%) of them in at least one of our two RNA-seq pools (Additional file 1: TableS4) [8].

To investigate structure and expression of key immune gene family members involved in cotton rat responses against RSV and other infectious pathogens, we aligned RNA-seq reads against our reference genome assembly. We focused analysis on clusters of genes in the *IFIT*, *CXCL*, and *GBP* families, which have been shown to play important roles in RSV replication. These gene clusters are located on cotton rat Chrs. 2, Chr. 1, and Chr. 16, respectively. No gaps were identified in any of the genes in these gene clusters, highlighting the high quality of our reference genome assembly (Fig. 2; Table 1). By contrast,

Table 2 Summary of annotated protein-coding genes in cotton rat genome assembly

	Hispid2A annotation
Number of protein-coding genes	24,878
Median gene length in genome (bp)	8808
Number of transcripts	29,005
Number of exons	235,349
Number of coding exons	227,132
Exons/transcript	9.93
Transcripts/gene	1.17
Multi-exonic transcripts	0.75
Gene density (gene/Mb)	9.93

Table 3 Features of RNA-seq read alignments. RNA-seq was conducted on RNA pools from several tissues extracted from naive vs. HDM-treated cotton rat individuals. Results from these two tissue pools were combined

Read counts and features	Treatment naïve	HDM	Combined
RNA-seq read counts	196,659,156	218,393,360	415,052,516
% uniquely mapped	81.69%	85.84%	83.77%
% mapped to multiple loci	5.25%	7.00%	6.13%
% unmapped	12.89%	6.87%	9.88%
expressed genes identified	20,834	20,506	21,417
% of all annotated protein-coding genes	93%	92%	96%

when the same RNA-seq reads were aligned against the previously reported BIUU assembly [19], numerous gaps interrupting the clustered family members were observed (data not shown).

Predicted exons, introns, and splice sites were confirmed by the RNA-seq read alignments. Specifically, we detected expression of alternative splicing isoforms of *Ifit2*, *Ifit1bl1*, *Ifit1*, and *Gbp7*, while no splicing isoforms were identified for other family members in the clusters. Counts of RNA-seq reads aligned against each annotated family member in the three families' chromosomal loci were wide-ranging, but each gene (and predicted exon) was expressed at a detectable level (transcripts per million, TPM > 1). These data were confirmed further by alignment of independent RNA-seq libraries generated from cotton rat individuals exposed to RSV infection [9].

Phylome analysis

To investigate the molecular evolution of the cotton rat genome in the context of other sequenced rodents and additional, more distant mammalian relatives, we first reconstructed its phylome, i.e., a complete collection of gene evolutionary histories [25] comprising the cotton rat genome. To provide an evolutionary framework for our comparisons, we reconstructed the evolutionary relationships between the 21 considered species by concatenating the alignments of 895 widespread genes with one-to-one orthologs in all species, and reconstructing a maximum likelihood tree using RAxML [26]. The species tree shows the expected topology (Fig. 3).

To make detailed comparisons between three species of particular interest, i.e., cotton rat, mouse, and human, we also reconstructed both the mouse and human phylomes based on the same set of 21 species. A total of 64,415 maximum likelihood gene phylogenies were reconstructed from their phylomes. We then examined gene trees in the phylomes to detect gene duplication events [27] that occurred in one or more of these three species of interest. Our initial estimates of the duplication frequencies were not normalized for time, because at first we did not develop or use a dated tree. Both rodent

species harbor an approximately two-fold increase in the number of species-specific duplications when compared to human, with an average of 0.37 duplications per gene in cotton rat and 0.31 per gene in mouse, compared with just 0.16 in human. This difference may be explained by different divergence times for these three species from their most recent common ancestor, as represented by branches in the species tree (Fig. 3).

To evaluate this possibility further, we obtained divergence times of interest [28] from timetree.org [29]. We then divided the number of duplications per gene by the amount of time estimated to separate each species from the common ancestor (Fig. 3). Even after applying this normalization, we still estimated that the human genome harbors roughly half the number of gene duplications compared with the rodents (i.e., cotton rat, 0.013; mouse, 0.015; and human, 0.0054). We also calculated divergence times using treePL [30] and applied the same correction. The normalized results were similar (i.e., cotton rat, 0.014; mouse, 0.022; and human, 0.0055), as they again revealed higher rates of gene duplication events in rodents [31].

Phylomes provide a view of what happens to each gene over evolutionary time, and yield redundant data about gene duplications. To assess how many species-specific expansions are present, we clustered them using UPGMA [32]. We applied the condition that if at least 50% of proteins that have expanded within a cluster were found to overlap between clusters, then they were combined into larger clusters. We identified 206 such clusters containing five or more members each, with the largest cluster containing 269 proteins, including proteins of unknown function harboring the conserved domain “MTH889-like” (also of unknown function).

Gains and losses of immune response gene homologs

Gains and losses of genetic orthologs and paralogs have been described among members of various gene families when studied across species. Such changes in gene counts particularly are anticipated when comparisons are made between species that diverged more than

(See figure on next page.)

Fig. 2 RNA-seq of pooled cotton rat tissues reveals expression patterns of gene family clusters involved in key immune response pathways. Poly(A)-positive mRNAs were extracted from multiple tissues of a healthy donor cotton rat and of a second individual that was exposed to house dust mite antigen, pooled, and reverse transcribed for preparation of strand-specific RNA-seq libraries. Approximately 50 million sequencing read pairs (2 × 150 bp paired end) were obtained per pool. Sequences were aligned to the refined reference cotton rat genome and compared with gene annotation models. Shown in tracks are as follows: *top*, counts of aligned reads; *middle*, Sashimi plots showing RNA splicing; and *bottom*, gene model schematics depicting exons and introns, at several immune gene family loci including the **A** *Ifit1* gene family (Chr. 2); **B** *Cxcl* gene family (Chr. 1); and **C** *Gbp7* gene (Chr. 16). We detected transcript isoform expression at each annotated gene and multiple instances of alternative splicing. For comparison, recently reported, independent RNA-seq reads from cotton rat individuals exposed to respiratory syncytial virus infection also were aligned to our annotated cotton rat genome assembly (Additional file 1: Fig. S5), revealing comparable splicing isoforms but at lower sequencing coverage

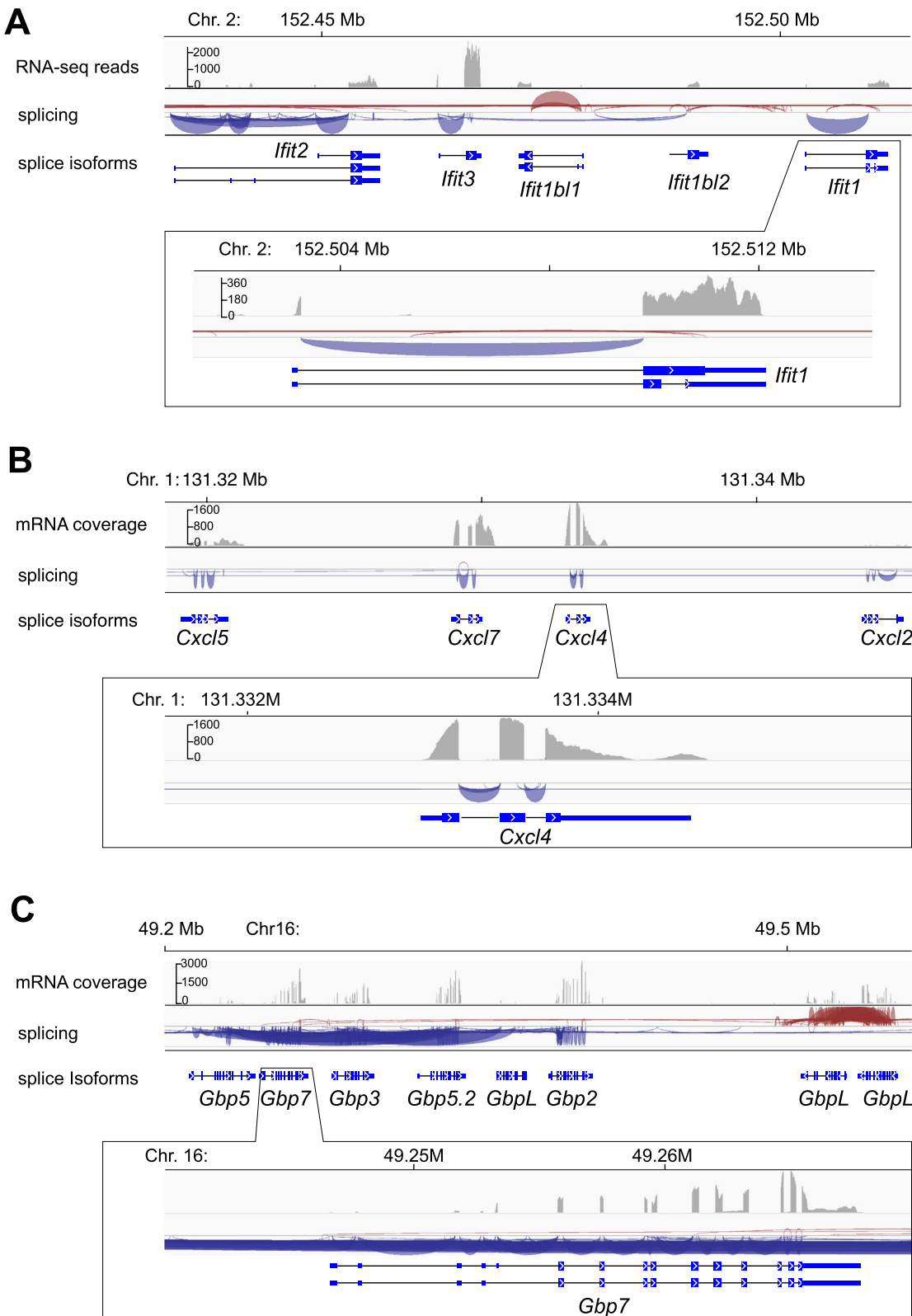


Fig. 2 (See legend on previous page.)

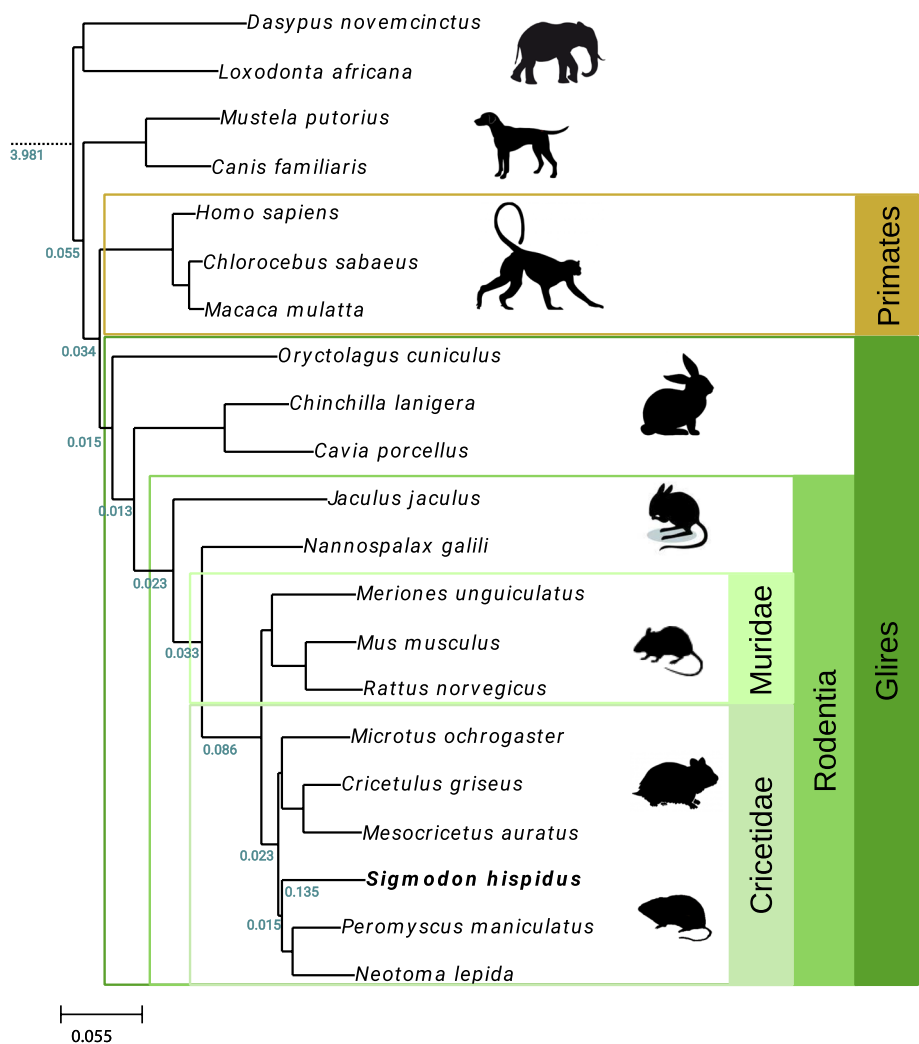


Fig. 3 Species tree obtained from phylome analysis of multiple vertebrate species. A phylogenetic tree displays evolutionary relatedness among multiple vertebrate, mammalian, and rodent species as labeled (right, species phyla). Support for all nodes was at 100% using the rapid bootstrap approach as implemented in RaxML [26]. Duplication values were calculated after removing large species-specific expansions. Left, numbers in green, numbers of duplications per gene per branch; bottom left, key, reference rate of duplication as indicated

20 million years ago [33, 34]. We used FatiGO [35] to identify enriched gene ontology (GO) terms categorizing genes that are duplicated in these species of interest. For cotton rat duplicons, a total of 199 GO terms were enriched, including several terms related to the immune system such as antigen binding, immunoglobulin production, immune response, complement activation, and cellular response to interferon-gamma (Additional file 1: TableS5A; adjusted p -value < 0.05). We also identified GO terms for genes that are duplicated in mouse: 92 GO terms were found, of which several were related to the immune system (Additional file 1: Table S5B). Similarly, 50 enriched GO terms for genes duplicated in human were found; again, several

were related to immune functions (Additional file 1: Table S5C). We searched phylome trees for genes duplicated both in cotton rat and one of the other two species examined, i.e., duplicated either in cotton rat and human, or in cotton rat and mouse. This analysis identified 263 duplicated genes in both cotton rat and human, and 497 duplicated genes in both cotton rat and mouse. GO enrichment analysis of these duplicated genes showed several enriched ontology terms (9 in cotton rat and human, and 58 in cotton rat and mouse), which again included several immune system-related genes (Table 4). Focusing further on particular genes whose involvement in human or mouse immune responses was established

Table 4 Comparisons of ontologies and biological processes involved in gene duplication events across three focus species. Yes/No, the listed GO term is or is not significantly enriched in the indicated species, i.e., the number of gene duplications in the ontological category in the species is or is not significant

GO_TERM representative	Cotton rat	Mouse	Human
Antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	No	Yes	No
Complement activation	Yes	No	Yes
Endocrine pancreas development	Yes	No	No
Epoxygenase P450 pathway	No	No	Yes
Establishment or maintenance of epithelial cell apical/basal polarity	No	No	Yes
Flavonoid biosynthesis	No	No	Yes
Glycolytic process	Yes	No	No
Macrophage apoptotic process	No	No	Yes
Mitotic prometaphase	Yes	No	No
Negative regulation of cell differentiation	Yes	Yes	Yes
Negative regulation of retinoic acid receptor signaling pathway	No	No	Yes
Protein localization to adherens junction	No	No	Yes
Receptor-mediated endocytosis	Yes	No	No
Response to stimulus	No	No	Yes
Sensory perception of smell	Yes	Yes	Yes
Sulfation	Yes	No	No
Zonula adherens assembly	No	No	Yes

experimentally, we evaluated their gains or losses in the phylome trees (Fig. 4; Additional file 1: Fig. S6). We tabulated 1378 human immune response genes, of which 1318 are found in phylomeDB, a public repository of gene phylogenies [25]. Of these, 1287 had orthologs in at least one other species, including 949 proteins with a cotton rat ortholog, and 1024 with a mouse ortholog.

We carefully examined 352 immune-related genes belonging to 17 gene families (Additional file 1: Additional Results and Fig. S7) [8, 34, 37–45]. Applying the neighbor-joining method to generate phylogenetic trees for each of these families (Additional file 1: Fig. S6), we identified significant gains and losses of gene orthologs in all three species of interest.

Complicating this analysis, we observed that many genes annotated with the same name in both the mouse and human reference genomes were not direct orthologs. For example, human *MX1* and *MX2* paralogs share the same genetic ancestor, but they are distinct from both the mouse and cotton rat *Mx1* and *Mx2* paralogs that have a distinct molecular ancestor (Additional file 1: Fig. S6). Similarly, the human *IFITM1*, *IFITM2*, and *IFITM3* paralogs originated from a human-specific duplication event, but they are only distantly related to the mouse *Ifitm1*, *Ifitm2*, *Ifitm3*, *Ifitm6*, and *Ifitm7* paralogs (Additional file 1: Figs. S6, S7). This result documents that despite similar name assignments, gene paralogs can diverge markedly, which results in confusing or misleading gene

nomenclature. In the cotton rat genome, we identified 11 *Ifitm* gene family members.

In tabulating losses of immune genes, we counted 63 absent from primates, 24 genes lost from humans, and 103 genes lost from cotton rat. We focused analysis on several functionally important genes absent from the cotton rat genome. For example, while human and mouse have one and two copies of *IFIT3* (human) and *Ifit3* and *Ifit3b* (mouse), respectively, cotton rat lacks an orthologous gene. Similarly, while both human and mouse genomes encode for *Apol6*, *Ccl1*, *Ang*, *Gsdmc*, *Sp100*, and *Sp140*, each of these genes is absent from cotton rat. An important minor histocompatibility antigen (MHA) locus that is present in mouse and human, i.e., the *Raet1/H60* locus [46–48], is absent from the cotton rat genome assembly (Additional file 1: Figure S8).

Retrotransposons and other repetitive elements

Using the RepeatMasker default database of rodent elements, derived from the Dfam database [49], we identified 4.58 million repetitive elements in the HispidOSU genome [16] (Additional file 1: Table S6), including ~1.29 million SINE elements, ~1.35 million LTR elements and ~810,000 LINEs. In accordance with the reported total occupancy of the human and mouse genomes by L1 elements (i.e., 17.5% and 19.9%, respectively), we estimate that 18.1% of cotton rat genomic DNA is made up of L1 elements. We count

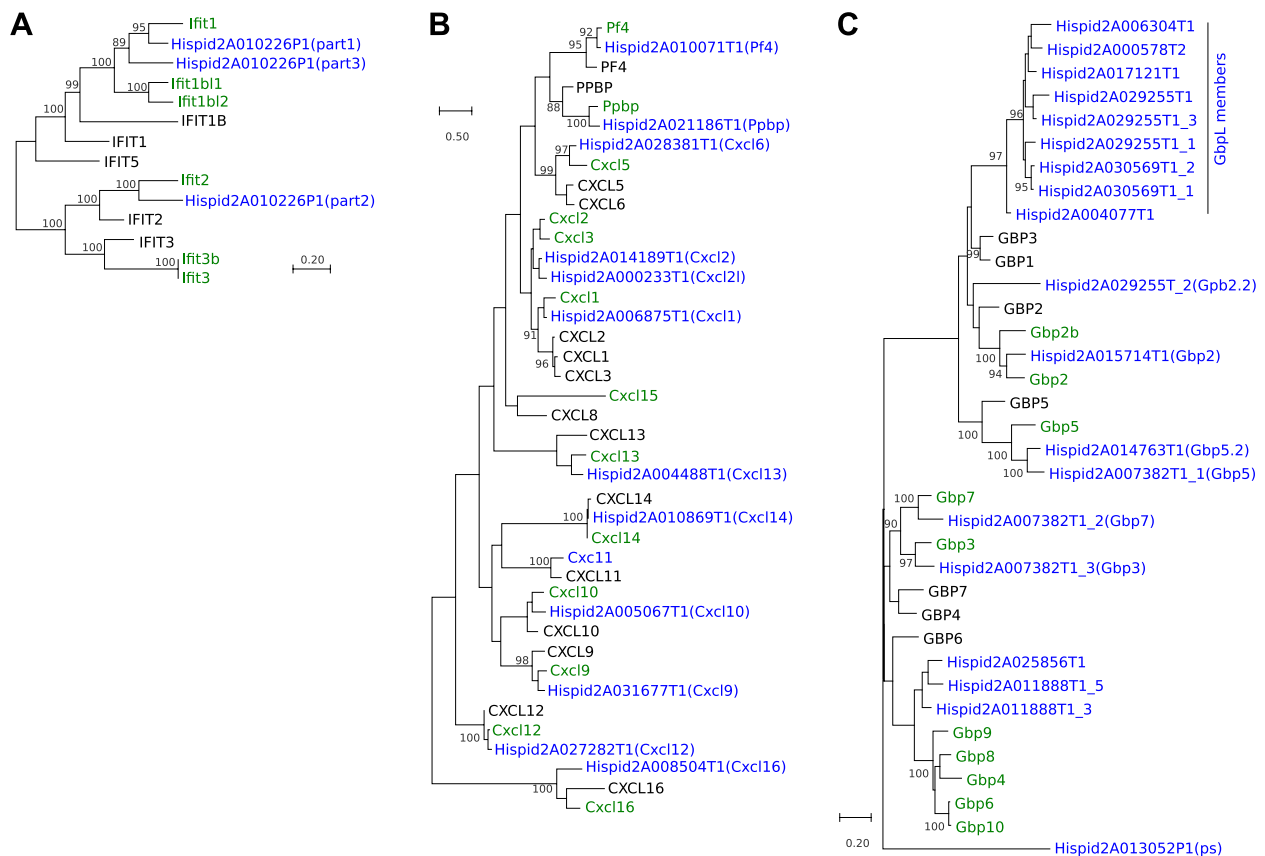


Fig. 4 Phylogenetic analysis of key immune response gene families in cotton rat vs. mouse and human genomes. Analysis of gene relatedness was conducted to construct phylogenetic trees based on amino acid maximum likelihoods. The results revealed the counts of and relatedness between individual gene family members from the **A** *Ifit*, **B** *Cxl*, and **C** *Gbp* families detected in the (blue font) de novo cotton rat (SigHisOSU), (green) mouse (GRCm39) and (black) human (GRCCh38) genomes. Gene annotation IDs and models for cotton rat genes are freely available [36]. Bootstrap values, i.e., the frequency at which the indicated tree structure was supported by a bootstrap method, are shown if > 90. Distance legend is based on the evolutionary distance between amino acids

a total of approximately 950 million nucleotides comprising interspersed repetitive elements in cotton rat. The total fraction of the cotton rat genome comprised of repetitive elements overall is 41.1% (Additional file 1: Table S6). This compares reasonably well with 43.5% occupancy by repetitive elements in the mouse (mm10) assembly.

We also used RepeatModeler to identify 358 distinct repetitive element families in the cotton rat genome [50, 51]. Resulting consensus sequences were used as inputs for RepeatMasker, to compare counts of elements in the BIUU assembly [19] vs. in our HispidOSU assembly [16]. We also sought to investigate divergence of individual family members from each family's consensus. The large majority of repetitive elements identified in cotton rat have diverged more than 2% from their consensus sequences, regardless of which genome assembly was analyzed (Fig. 5A and B). In general, these more divergent elements are considered to be ancient, inactive elements.

By contrast, the low-divergence elements represent young, active (i.e., recently retrotransposed) repeat family members. We identified 180,869 repetitive elements with less than 2% sequence divergence from the consensus repeat sequences in the BIUU assembly [19], and only 32,339 such elements in the HispidOSU assembly [16]. Of these, 77.2% of the low-divergence repeats in HispidOSU are SINE B2 elements ($n=24,955$), and 20.6% are LINE (L1) elements ($n=6658$; Fig. 5B).

We identified rnd-1_family-3 L1 as a consensus full-length, active L1 element in cotton rat. We refer to this unique consensus L1 sequence as L1sh (*S. hispidus*) (Additional file 1: Fig. S9). This consensus L1sh sequence is 5839 nt in length and is closest to previously reported SigHis-1.18 (Dfam database, acc. no. DF000280615) [49]. When we allowed identified L1s to diverge more than 2% from this consensus sequence, we counted 258,978 L1s in the BIUU assembly [19], compared with only 37,868 in HispidOSU [16]. Focusing on the subset of unique L1

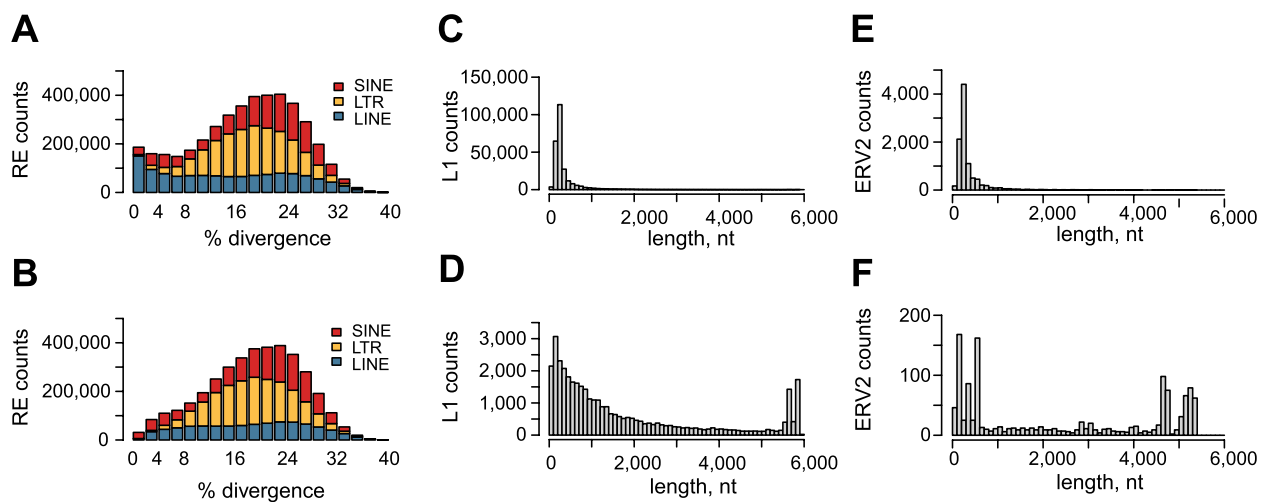


Fig. 5 Repetitive elements in cotton rat genome assemblies. We compared counts of repetitive elements identified in the **A, C, E** BIUU genome assembly [19] vs. **B, D, F** in our HispidOSU assembly [16]. **A, B** Stacked bar graphs depict (y-axis) counts of (key, upper right) repeat family members which (x-axis) diverge to various extents from consensus sequences as indicated. Key, colors: red, SINE retrotransposons; yellow, LTR retroelements; blue, LINE retrotransposons. Consensus sequences were derived from RepeatModeler analysis of our HispidOSU genome assembly. **C, D** Distributions of cotton rat L1 elements' lengths detected in **C** BIUU vs. **D** HispidOSU assemblies. Our consensus full-length cotton rat L1 sequence, termed L1sh, is shown in Additional file 1: Fig. S9. (Y-axis) Counts of unique L1 elements with (x-axis) indicated lengths. **E, F** Element length distributions of cotton rat ERV2 elements detected in **E** BIUU vs. **F** HispidOSU assemblies. Our consensus full-length cotton rat ERV2 sequence is presented in Additional file 1: Fig. S10. (Y-axis) Counts of unique ERV2 elements with (x-axis) indicated lengths

elements which were >90% full-length, we found only 372 in BIUU, compared with 4160 nearly full-length L1s in HispidOSU (Fig. 5C and D, respectively).

Our analysis of LTR elements revealed a unique consensus cotton rat ERV, referred to as ERV2sh (Additional file 1: Fig. S10). This consensus ERV2sh is closest to SigHis-1.326_int (Dfam, acc. no. DF000282675; not full-length, 5312 nt in length). When we allowed identified ERV elements to diverge by >2% from this consensus sequence, we counted 9897 ERV2s in BIUU genome assembly [19], compared with 1326 ERV2s in the HispidOSU assembly [16]. Focusing on nearly full-length ERV2 elements, we found only 14 in the BIUU assembly, compared with 250 in HispidOSU (Fig. 5E and F, respectively).

Members of one cotton rat SINE family represented 92% of all young SINE elements identified. We refer to their consensus (rnd-1_family-2#SINE/B2) as the cotton rat SINEB2sh consensus sequence (Additional file 1: Fig. S11).

We examined the predicted protein-coding sequences in the consensus L1sh element. All mammalian L1 retrotransposons identified to date contain two open reading frames (ORF1 and ORF2), with ORF2 encoding endonuclease and reverse transcriptase activities. As expected, two ORFs were identified in L1sh. Comparison of L1sh ORF2 amino acid sequences revealed 74.6% identity with rat L1 ORF2 and 73.1% identity with mouse L1 ORF2.

L1sh ORF1 has 67% identity to rat L1 ORF1 and 67% identity to mouse L1 ORF1. In general, mammalian L1 ORF2 proteins typically are more highly conserved across species than are L1 ORF1 proteins (Wagstaff et al. 2011). This trend also holds in comparisons of cotton rat L1sh to other identified L1 sequences, as expected.

To examine evolution of L1 retrotransposons across rodent and other mammalian species, we compared L1 ORF2 protein sequences across 9 and independently across 23 species, by creating phylogenetic trees (Fig. 6A and Additional file 1: Fig. S12, respectively) [38]. Because L1 elements are vertically transmitted through the germline, and because their evolution frequently mirrors the evolution of their host species, we compared the L1 ORF2 phylogenetic trees against the species phylogenies determined by analysis of their phylomes (i.e., compare trees in Fig. 6A vs. B, and compare trees in Additional file 1: Fig. S12 vs. Figure 3, respectively). The results confirm that the evolutionary relationships between L1 ORF2 sequences consistently match those between the species' phylogenies and divergence times.

Discussion

The cotton rat (*S. hispidus*), a rodent native to the Americas, has emerged as an animal model useful in the study of human respiratory virus pathogenesis and in the development of vaccines and antiviral therapeutic agents. Cotton rats have innate susceptibility to a variety

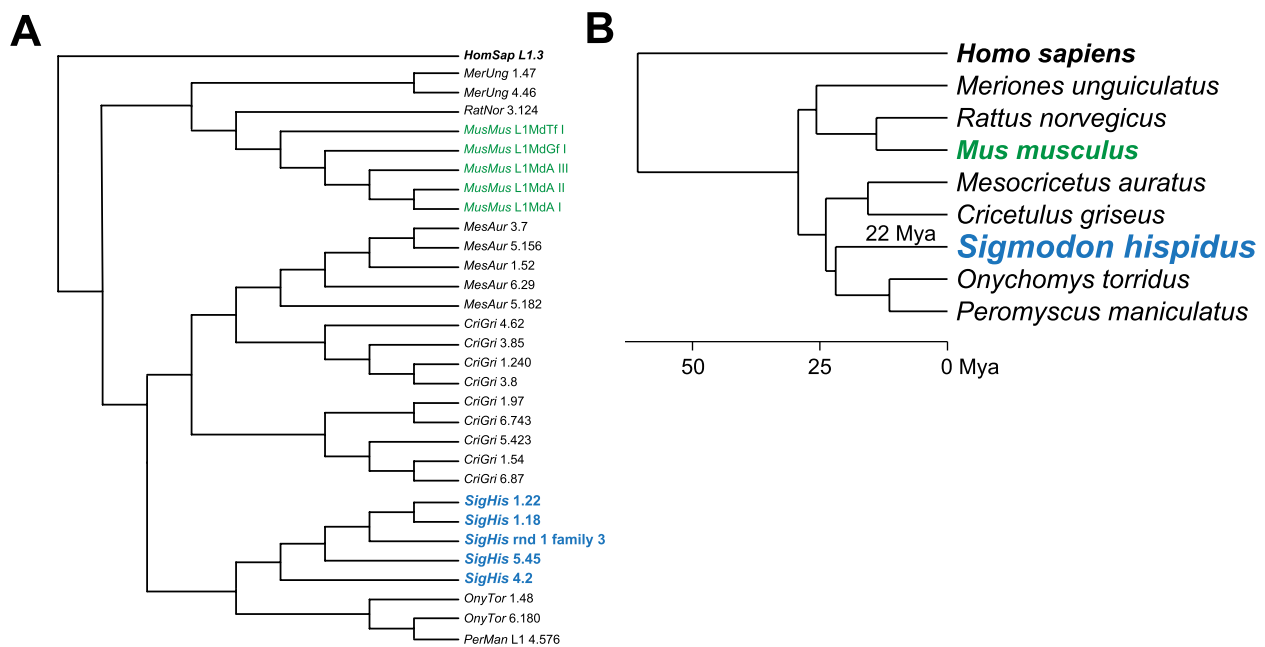


Fig. 6 Comparison between phylogenetic trees based on L1 ORF2 sequences and genomes from the same mammalian species. **Bold black font**, human; **green**, mouse; **blue**, cotton rat. **A** Phylogenetic tree was constructed from L1 ORF2 amino acid sequences from the nine indicated species, as annotated in the Dfam v. 3.8 database [49]. **B** Species phylogenetic tree was constructed, and *bottom*, divergence times were calculated as described in “Methods”

of human pathogens, although the genetic basis for this susceptibility has remained unknown. Here we report a new, high-quality genome assembly, based first upon our linked-read and PCR-free WGS data. By incorporating data from chromosome paints derived from *M. musculus* and comparative fluorescence microscopy analysis of metaphase spreads of several rodent species, i.e., *S. hispidus*, *M. musculus*, and *P. leucopus*, we were able to make chromosome-level assignments in the assembly. While long stretches of chromosomal sequences display synteny with other rodent species, we also identified numerous examples of genetic innovation in gene families, supported by phylogeny analysis. A large majority of RNA-seq reads from pooled cotton rat tissues and independently from infected animals aligned well to annotated genes and even revealed alternative splicing isoforms, suggesting broad utility of this new genome assembly as a valuable reference in future studies.

The highly contiguous HispidOSU genome assembly [16] was generated by combining 10×linked-read and PCR-free WGS data, i.e., from short sequencing reads and linked reads only (Table 1). Two features of our experimental design provided advantages that improved the resulting genome assembly: high physical coverage of the genome was attained without PCR amplification-induced artifacts; and long-distance genomic connectivity between barcoded, linked reads was reached because

they were derived from the same physical molecules (frequently > 50–100 kb) [13]. Despite having a high scaffold N50 length, our intermediate assembly still was characterized by a relatively high number of small gaps. We patched them by generating complementary assemblies made from the PCR-free paired end libraries, followed by a second step where the remaining gaps were closed using Compass, a gap-filling software tool [52].

Comparisons between our HispidOSU cotton rat genome assembly [16] and another independent, recently released genome assembly, i.e., BIUU [19], confirm the overall size and general composition of the cotton rat genome (Table 1). However, a limitation of the BIUU assembly is highlighted by its much higher counts of scaffolds, most of which are single-contig scaffolds reflecting its reduced contiguity and increased fragmentation overall. We observed examples of such fragmentation disrupting gene cluster structures in this recent BIUU assembly, which by contrast were resolved in our new genome assembly (Fig. 2; data not shown). An additional example of fragmentation plaguing the earlier assembly was provided by comparing counts of full-length, active, non-divergent L1 and LTR elements identified in the two assemblies (Figs. 5C–F). The counts of nearly full-length L1s and ERV2s were about tenfold higher in the HispidOSU assembly than in the BIUU assembly, likely because the linked-read sequences underlying

HispidOSU improved connectivity across long repetitive sequences.

We took advantage of a relatively high degree of orthology between the *S. hispidus* genome and other rodents including the well-characterized *M. musculus* genome, by analyzing metaphase spreads from the former using chromosome-specific paints derived from the latter. This analysis facilitated assignment of *S. hispidus* genomic sequences into chromosomes on the basis of their lengths, banding patterns and synteny, resulting for the first time in a chromosome-level cotton rat genome assembly.

Genes in the new cotton rat reference genome were identified and annotated by making ab initio gene predictions, adding alignments of RNA sequencing data, and conducting cross-species comparisons. We estimated 24,878 protein-coding genes based on ab initio predictions (Table 2). This count is on par with other mammalian and rodent species. Of these, we detected expression of transcripts for 21,417 (96%) in RNA-seq libraries pooled from multiple cotton rat tissues (Table 3). Detailed examination of transcripts expressed from particular genes in several conserved immune gene families, upon alignment of RNA-seq reads against the annotated genome assembly, corroborated their predicted exonic structures.

Phylome analysis facilitated investigation of the evolutionary relatedness among members of gene families conserved among the cotton rat, mouse, human, and other species. This revealed numerous genomic innovations manifested as gene family gains and losses distinguishing cotton rat from other rodents. The finding that cotton rat has undergone a more pervasive loss of immune-related genes compared to mouse and humans may have implications about its increased susceptibility to infection by viruses and other pathogens. However, we acknowledge that some of these apparent gene losses may be attributed in part to technical difficulties in assembling short-read sequencing data de novo, even when the short reads are locally associated as linked reads. In particular, numerous immune genes (e.g., in the major histocompatibility complex and immunoglobulin gene loci) are encoded in large, repetitive gene clusters that are particularly difficult to resolve from short reads [21], even when connected as linked reads as those underlying HispidOSU [16].

Functional studies have highlighted both some differences between mouse and cotton rat immune responses and concomitant similarities between the latter and humans. In contrast to the mouse, but similar to humans, cotton rats have columnar epithelial cells in their respiratory tract, which are the target of infection by viruses such as RSV and influenza virus [53, 54]. Toll-like receptor 9 molecules are expressed at much lower levels in

cotton rat or human lymphoid tissues than in the mouse [55]. Human *TLR9* agonists also stimulate cotton rat cells better than agonists developed for mouse *Tlr9* [55, 56]. As in human macrophages, cotton rat macrophages produce little nitric oxide, implying that nitric oxide is mainly used as a signal transduction molecule in both species, and not as an anti-bacterial effector as it is in mice [57]. Another similarity between humans and cotton rats lies in the constitutive expression of heat shock protein 70 (*HSP70*) in tissues, which is not the case in mice (Niewiesk, unpublished). We anticipate that the availability of the HispidOSU reference genome [16] will facilitate further genetic studies on the basis for these and other functional properties in cotton rat.

More than 800 cotton rat genes' coding sequences previously were cataloged in the National Center for Biotechnology Information (NCBI) database [58]. Most of these are related to the immune response. We compared individual orthologs in cotton rat, mouse, human, and other species, in an effort to decipher differences between their susceptibility and immune responses to various pathogens. Overall, we found that the cotton rat genes are more closely related to hamster genes than to mouse and rat genes, as expected (Fig. 3). This evolutionary relationship was illustrated by orthologs of *CD1*, a restriction element of natural killer (NK) cells [59]. The observed differences in homology are consistent with the split between the *Muridae* (mouse, rat) and *Cricetidae* (hamster, cotton rat) families 23.3 to 24.9 million years ago, whereas the split between the *Arvicolinae-Cricetinae* (hamster) and *Sigmodontinae-Netominae* (cotton rat) occurred 18.7–19.6 million years ago [60]. However, all of these rodent genes are more closely related to each other than to human genes, again as expected (Fig. 3, Fig. 6).

CD150 (a lymphocyte activation molecule) is another annotated gene with important functions, as it encodes a receptor for measles virus (MV) with functional orthologs in human and cotton rat but a nonfunctional ortholog in mouse. Key differences in amino acids at positions 60, 61, and 63 may explain these interspecific differences in MV receptor function [61]. In contrast, no such differences could be identified in the orthologous RSV receptor protein CX3CR1 encoded in cotton rats, mice, and humans [54]. In this case, unidentified downstream factors may influence or mediate differential susceptibility to RSV infection. Like humans, cotton rats express a functional set of Mx proteins encoding the antiviral proteins Mx1 and Mx2 [62]. These proteins severely reduce infection of the cotton rat lung tissue by influenza virus. In contrast, most mouse strains lack functional Mx proteins, and therefore rely on an incomplete innate type I interferon response.

The draft reference genome of cotton rat defines the structures of numerous gene candidates involved in virus susceptibility or immunity, and therefore will facilitate future research on many aspects of viral replication and immunity in cotton rats. We have documented a significant number of losses of particular immune-related genes, absent from the cotton rat genome. Some genes that are conserved between mouse and human but absent from cotton rat include *Ccl1*, *Ang* gene family members, *Ifit3*, *Rarres2*, *Nlrp2*, *Gsdmc*, *Reat1e*, *Sp100*, and *Sp140*. Other genes are lost from both human and cotton rat but present in mouse, including AIM2-like receptor members (*Pydc3* and *Pydc4*), *H60b*, *Ifitm6*, *Irgb10*, *Irga6*, and *Apol6*. The loss of these genes from the cotton rat genome may help explain the host's exquisite susceptibility to human viruses and other pathogens. Further characterization of the potentially protective roles of these genes in immunity is warranted.

Mobile genetic elements comprise approximately half of mammalian genomes, including those of all rodent species sequenced to date. They also play important roles in shaping genomes and contributing to speciation, as they both contribute to random variation and are objects of natural selection through evolution. Transposons also have been implicated in the etiology of diseases in human and other species [63–65].

The cotton rat is an outgroup to other rodent species that lost active L1 retrotransposition [66]. Moreover, the cotton rat itself does not harbor active SINE B1 retrotransposition. For the first time, we have defined unique, nearly full-length, consensus cotton rat L1, ERV2 and SINEB2 element sequences, which we termed L1sh, ERV2sh and SINEB2sh (Additional File 1: Figs. S9 – S11).

Our ERV2sh consensus sequence (Additional File 1: Fig. S10) aligns well with reported pro/pol region sequences of active mysTR elements in cotton rat (e.g., GenBank DQ139757.1 to DQ139766.1) [67], with nucleotide sequence identities ranging from 86 to 95% (and mostly from 92.5 to 94.6%) within the pro/pol region. We identified approximately 1326 ERV2 elements in our HispidOSU assembly [16] (Fig. 5F), including 250 nearly full-length elements. These counts compare favorably with the ~1000 mysTR elements estimated previously with PCR-based methods in the cotton rat genome [67].

We acknowledge several ways by which the HispidOSU assembly [16] could be improved further. First, although linked-read sequences are derived from long physical DNA molecules and therefore facilitate analysis of long-range connectivity spanning repetitive genomic elements, they nevertheless are discontinuous. Despite improvements in linked-read library preparation methods, identical barcodes still frequently label

multiple independent genomic DNA segments, thereby introducing potential artifacts. Other methods such as continuous long-read sequencing, optical mapping, and Hi-C methods have been optimized recently. Use of such methods is likely to improve connectivity spanning across widespread repetitive elements, required for high-quality genome assemblies [68]. Second, the individuals sequenced here were highly inbred, so extensive genomic homozygosity would be expected. Genomic studies in diverse, wild-caught, outbred cotton rats would help define population-level allelic variation frequencies and elucidate the population size. Third, while some chromosome-level scaffolds may have represented second haplotypes in genomic regions where heterozygosity persisted, we did not explicitly address allelic variation or diploidy here, nor did we conduct phasing. Identification of additional haplotypes will be facilitated by analysis of diverse cotton rat individuals. For example, inclusion of additional, independently collected sequencing data supporting the BIUU assembly [19] may add such haplotype information. In addition, the resulting, combined genome assembly would be improved from increased depth of sequencing coverage. Another limitation is that a mean of only 1.17 transcripts was detected per annotated gene. Deeper RNA-seq across a full complement of tissues and diverse experimental and developmental conditions and infections would improve quantification of tissue-specific expression of most annotated genes and identification of novel transcripts. In addition, many more examples of alternative splicing and diverse transcript isoforms would be identified.

Conclusions

We assembled and annotated a high-quality, reference assembly of the cotton rat genome, by combining multiple lines of evidence including PCR-free whole genome sequencing, linked-read sequencing, chromosome paint data, and RNA-seq data. This genome reference is extended further by chromosome assignments (karyotype $2n=52$), demonstrating significant similarity and synteny with other new world rodents, and extensive rearrangements when compared to the mouse. Several unique, active transposable element families were identified in *S. hispidus*, including L1sh, ERV2sh, and SINEB2sh, which actively contributed to its genomic structure. We anticipate that this annotation and characterization of HispidOSU, a chromosome-level cotton rat genome assembly, will serve as a valuable resource, and will facilitate and accelerate ongoing investigations into its host defenses against viral and other pathogens, genome biology, and mammalian evolution.

Methods

Animals

Inbred cotton rats (*S. hispidus*) that were between 4 and 8 weeks of age and free of specified pathogens (as specified by the breeder) were purchased from Envigo, Inc. (Indianapolis, IN). They were maintained in a barrier system in accordance with a protocol approved by the Ohio State University Institutional Animal Care and Use Committee. Environmental conditions were maintained at $20 \pm 2^\circ \text{C}$ and 30–70% relative humidity with a 12-h light cycle. Euthanasia was performed via CO_2 inhalation.

Genomic DNA extraction, library preparation, and sequencing

To obtain high molecular weight DNA for linked-read sequencing and WGS library preparation, DNA was extracted from an individual male cotton rat's ear pinna tissue, using MagAttract HMW DNA kit (Qiagen). The isolation protocol including RNase treatment followed the manufacturer's recommendations. DNA quality and concentration and the DNA integrity number were measured using a Nanodrop spectrophotometer, Qubit fluorimeter, and an Agilent TapeStation, respectively.

To generate PCR-free libraries, we fragmented genomic DNA using a Covaris S2 sonicator, resulting in DNA fragments of various size distributions. We optimized shearing to yield median fragment lengths of 350 and of 550 nt in two independent aliquots. An Illumina TruSeq DNA Sample Prep kit was used to add adapter and sample barcode sequences to the resulting genomic DNA fragments, following the manufacturer's protocol. Library quality was assessed using an Agilent TapeStation, and concentrations were determined using a Qubit fluorimeter. Sequencing was performed on an Illumina HiSeq2500, resulting in 2×150 bp paired-end reads with indexes incorporated for each library.

A linked-read genomic DNA library was prepared from approximately 1.25 ng high molecular weight genomic DNA as template, using the 10× Genomics Chromium genome library and gel bead kit to create Gel Bead-In-Emlusions (GEMs), following the manufacturer's protocol (10× Genomics). Isothermal incubation of the GEMs produced DNA fragments sampling from the high molecular weight DNA inputs, barcoded with 10× Genomics linked-read indexes. Additional sequencing primers and sample indexes were added by end repair, A-tailing, and adaptor ligation. After amplification, the barcoded library was size selected. The resulting library structure and concentration were assayed by quantitative PCR (KAPA Biosystems). Sequencing was conducted on an Illumina HiSeq2500, yielding 2×150 bp paired-end reads along with the sample and 10× linked-read molecular indexes.

Bioinformatic analysis of sequencing data

In an initial pre-processing step, reads from the two PCR-free libraries were screened for adapter sequences and low-quality bases ($Q < 10$) and trimmed using Cutadapt 1.8.1 [69]. After this step, read pairs containing a single read shorter than 50 bp were discarded. To filter out spiked-in sequences, remaining reads were mapped against the PhiX reference sequence using GEM mapper (edit distance $\leq 10\%$) [70]. Finally, processed PCR-free reads were error-corrected using Lighter v1.1.1 ($k = 21$) [71].

$10 \times$ Chromium linked-read data also were error-corrected, using bloom-filters generated from the PCR-free data which were characterized by lower error rates. Barcodes and an additional seven nucleotides were clipped off from each Read 1 prior to error-correction. They were re-included subsequently, to ensure valid inputs for the assembler.

Genome assembly

Error-corrected $10 \times$ linked reads were used as inputs into Supernova v2.0.1 [14]. A pseudo-haploid representation of the assembly was generated using the subcommand *mkoutput*. The assembly was further scaffolded using ARKS v1.0.2 [15].

Processed PCR-free data was used to produce nine contig assemblies using ABySS 2.0.2 [72], exploring different K -mer sizes (i.e., 37, 47, 57, 67, 77, 87, 97, 107, and 117). Flanks of decreasing lengths (starting at 1 kb and ranging down to 100 bp, in decrements of 100 bp) around each gap in the Supernova assembly were searched for in these assemblies. When both flanks mapped unambiguously to the same contig and in the correct order and orientation, using GEM mapper [70], the sequence between the outermost mapping coordinates was extracted and used to patch the gap, giving priority to sequences originating from assemblies of larger K -mer size. Remaining gaps were filled using Compass (<https://github.com/nygenome/compass>), exploring the same K -mer sizes listed above. Finally, all scaffolds shorter than 200 kb were searched for in the assembly, using MegaBLAST [73]. Scaffolds that fully aligned to a larger scaffold (coverage = 100%, identity $\geq 99\%$) were considered redundant and therefore removed.

Gene completeness was evaluated using CEGMA 2.5 [17] with a default 248 core eukaryotic gene set, and using BUSCO 5.4.0 [18] with the *mammalia_odb10* gene set. In another check of completeness, PCR-free data were mapped in paired-end mode against the final assembly with BWA-MEM v0.7.17 [74]. Corresponding mapping statistics were computed using CollectMultipleMetrics

from the Picard toolkit v2.16.0 (<https://broadinstitute.github.io/picard>).

Assigning pseudo-chromosomes

Preparations of cotton rat chromosomes were made from embryo fibroblast cells that were isolated from six 12–14-day-old cotton rat embryos using a mouse embryonic fibroblast isolation kit (Pierce). Metaphase chromosome preparations and cross-species chromosome painting were performed using laboratory mouse chromosome probes as described previously [75] (Additional File 1: Fig. S1). Paint probes specific for *S. hispidus* Chrs. 1, 2, 3, 8, 9, 11, 22, 23, 24, and 25, generated from flow-sorted mouse chromosomes, also were hybridized onto *O. torridus* metaphase chromosomes to evaluate concordance between chromosomes of *S. hispidus* and *O. torridus*.

Pseudo-chromosome models were constructed using Syn2Chr (<https://github.com/igcbioinformatics/Syn2Chr/>), based on synteny between cotton rat and the mouse reference genome GRCm38. Ambiguous genome structures were corrected based on an updated *Peromyscus leucopus* genome assembly (i.e., UCI_PerLeu_2.1, https://www.ncbi.nlm.nih.gov/assembly/GCF_004664715.2) [76] and on the *O. torridus* genome (i.e., mOncTor1.1, GCA_90399543 GCF_903995425.1, https://www.ncbi.nlm.nih.gov/assembly/GCF_903995425.1) [77].

RNA extraction and sequencing

Tissues included for analysis by RNA sequencing (RNA-Seq) were harvested freshly from euthanized male cotton rats. RNA was extracted from brain, heart, intestine, kidney, liver, lung, mediastinal lymph node, muscle, Peyer's patch, spleen, and thymus of a healthy, naive adult cotton rat male. Independently, a second cotton rat male was injected intraperitoneally with 100 µg of HDM adsorbed to AdjuPhos (Brenntag) in a 1:1 v/v ratio [78]. Eight days later, it was challenged intranasally with 100 µg of HDM in a 100 µl volume. The cotton rat was euthanized 4 days post-challenge, and RNA was extracted from lung, mediastinal lymph nodes, and spleen. RNA was isolated using the RNeasy Microarray Tissue kit (Qiagen).

RNAs from various tissues were pooled for each individual. RNA pool concentrations and quality were checked using a NanoDrop (Thermo Scientific) and Bioanalyzer (Agilent), respectively, and were determined sufficient for RNA sequencing. Messenger RNAs were enriched based on their poly(A) tails, reverse transcribed to cDNA and barcoded using indexed adapters to permit multiplexing of individual sample pools. The cDNA libraries were prepared and finished, and quality and concentration were measured, in the Ohio State University Comprehensive Cancer Center Genomics Shared

Resource. Sequencing was carried out in a single lane in an Illumina HiSeq4000 instrument, generating 2 × 150 bp paired-end reads output in fastq file format.

RNA-seq analysis

RNA-seq reads were aligned to the cotton rat genome assembly mSigHis_REL_1907.fa, harboring 26 chromosomes and more than 16,000 unplaced scaffolds, with annotation file mSigHis_REL_1907.gff3. We used STAR v2.5.3a [79], using command lines including genomeDir genome --readFilesIn../evidence/NaiveCR_1-11_RNAseq/1_11_S28_L007_P0001_R1.fastq.gz;../evidence/NaiveCR_1-11_RNAseq/1_11_S28_L007_P0001_R2.fastq.gz; --readFilesCommand zcat; --runThreadN 4 --outFileNamePrefix star/Naive; --outSAMstrandField intronMotif --outSAMtype BAM SortedByCoordinate; --outSAMattrIHstart 0; --outFilterIntronMotifs RemoveNoncanonical; --outTmpDir \$TMPDIR/Naive.

For comparisons with RNA-seq data generated independently [9], Illumina raw reads were downloaded from the short read archive, SRA database [24]. Alignments against our cotton rat genome assembly were performed using STAR 2.79a with command lines: STAR --genomeDir../Star2.79a --readFilesIn<(zcat SRR23104982_1.fastq.gz)<(zcat SRR23104982_2.fastq.gz) --alignIntronMax 500,000 --outSAMtype BAM SortedByCoordinate --outFileNamePrefix Inf1 --sjdbOverhang 99 --sjdbGTFfile../mSigHis_REL_1907.gff3 --runThreadN 16.

Gene annotation

Gene annotation of the cotton rat genome assembly was generated by combining transcript alignments, protein alignments, and ab initio gene predictions. Transcript models were prepared using Cufflinks v2.2.1 [80]. With the addition of 107 *Sigmodon* genes downloaded from NCBI, PASA assemblies were produced with PASA v2.3.3 [81]. TransDecoder, part of the PASA package, was run on the PASA assemblies to detect coding regions in the transcripts. The complete human, rat, and mouse proteomes were downloaded from Uniprot, and aligned to the genome using spaln [82] (v2.2.2). Ab initio gene predictions were performed on the repeat masked cotton rat assembly using three different programs: GeneID v1.4 [83], Augustus v3.2.3 [84] and Genemark-ES v2.3e [85], with and without incorporating evidence from RNAseq data. The gene predictors were run with the human trained parameters, except Genemark, which runs in a self-trained manner. All results were combined into consensus gene coding sequence (CDS) models using EvidenceModeler v1.1.1 [81]. Additionally, untranslated regions (UTRs) and alternative splicing isoforms were updated through two rounds of PASA annotations. Functional annotations

were assigned to proteins with Blast2go [86]. A Blastp [87] search was run using the nr database, and then Interproscan [88] was run to detect protein domains on annotated peptides. All resulting data were combined using Blast2go which produced the final functional annotation results.

Annotations of noncoding RNAs (ncRNAs) were produced by running the following steps. First, the program cmsearch v1.1 [89], part of Infernal [90], was run against the RFAM [91] database of RNA families (v12.0). In addition, tRNAscan-SE v1.23 [92] was run to detect transfer RNA genes present in the genome assembly. To detect long-noncoding RNAs (lncRNAs), we selected those PASA-assemblies that had not been included into the annotation of protein-coding genes in order to evaluate expressed but untranslated transcripts. Particular PASA assemblies lacking protein-coding annotations that exceeded 200 bp and whose length was not covered at >80% by a small ncRNA were incorporated into the ncRNA annotation as lncRNAs. Resulting transcripts were clustered into genes using shared splice sites or significant sequence overlap as criteria for designation as the same gene.

Phylome reconstruction

We constructed a phylome, i.e., a complete collection of phylogenetic trees depicting the relatedness of each gene across a set of evolutionarily distinct genomes. Cotton rat, mouse, and human genomes were included in this analysis to identify gene gains and losses among these key species. Twenty additional species included 11 Rodentia species were investigated in each phylome. Phylomes were constructed using an automated pipeline [25, 27]. Briefly, for each predicted protein encoded by a particular genome, a Smith-Waterman search was performed against a proteome database [93]. Results were filtered using an e-value cut-off $<1E-5$ and a continuous overlapping region of 0.5. Up to 150 homologous sequences for each protein were identified, which were then aligned using MUSCLE v3.8 [94], MAFFT v6.712b [95], and kalign [96]. Alignments were performed in forward and reverse orientations and in each possible reading frame using the Head or Tail approach [97]. The resulting six alignments were combined with M-COFFEE [98] and then trimmed with trimAl v1.3 [99], with a consistency score cutoff of 0.1667 and gap score cutoff of 0.9. Trees were reconstructed using the best-fitting evolutionary model. The selection of the model best fitting each alignment was performed as follows: a Neighbor Joining (NJ) tree was reconstructed as implemented in BioNJ [100]. The likelihood of each topology was computed, allowing branch-length optimization, using 7 different models (JTT, LG, WAG, Blosum62, MtREV, VT, and Dayhoff), as

implemented in PhyML v3.0; and then the model best fitting the data, as determined by the AIC criterion [101], was used to derive ML trees. Four rate categories were used, and invariant positions were inferred from the data. Branch support was computed using an approximate likelihood ratio test (aLRT), based on a chi-square distribution. Resulting trees and alignments are stored in phylomeDB (<http://phylomedb.org>), labeled with phylomeIDs 19 (cotton rat), 20 (mouse) and 21 (human). Trees were scanned using ETE v3.0 [102].

We reconstructed additional phylomes to include *Onychomys torridus*. We then used the single-copy orthologs to reconstruct the species trees and calculate divergence times as described above (Fig. 6B). Topologies remained unchanged and divergence times varied little across repetitions and trees. Resulting phylomes are archived under phylomeIDs 276 (large phylome) and 350 (small phylome) in phylomeDB.

Species tree reconstruction

A total of 895 proteins were found encoded as single-copy genes in all 21 species. These proteins were used to reconstruct a species tree by concatenating the clean alignments produced during phylome reconstruction. The concatenated alignment contained 635,820 amino acid positions. RAXML HPC-PTHREADS-SSE3 version 8.2.4 [26] was then used to reconstruct a phylogenetic tree based on the PROTGAMMALG model. Branch support was calculated using the rapid bootstrap approach implemented in RAXML. Species trees followed the taxonomic classification as expected.

Divergence times were calculated using treePL [30], based on the species tree calculated from phylome analysis and using three calibration points obtained from [103]: divergence between primates and rodents (estimated between 61 and 164 million years ago, MyA), base of Rodents (56–66 MyA) and Divergence between Mouse and Rat (10–14 MyA). Parameters were established after running cross-validation. Different smoothing parameters were tested with very little difference in the results.

Orthology and paralogy determination

Trees were scanned for orthologs and paralogs using a species overlap algorithm as implemented in ETE v3.0 [102]. For each tree, nodes were annotated as speciation or duplication nodes, depending on whether there were common species at both sides of the node or not. When common species were present, a duplication node was annotated. In this case, sequences on either side of the node were considered as paralogs. If no common species were found, a speciation node was annotated. In such a case, the sequences were considered as orthologs.

GO term assignment and enrichment analysis

GO terms for proteomes included in the phylomes were downloaded from phylomeDB. The GO terms were transferred between one-to-one and many-to-one orthologs to cotton rat genes. Enrichment of GO terms was calculated using a python adaptation of FatiGO [35], which uses Fisher's exact test with multiple testing correction to identify enriched GO terms, adjusted p -value < 0.05 .

Annotation of transposable elements

RepeatModeler version 2.0.5 [51] was used to detect repeat families in the HispidOSU genome assembly. Resulting consensus sequences were used as an input library for RepeatMasker version 4.1.6 [104] to count repetitive elements and to map their coordinates in the cotton rat assembly. We focused on those elements with $< 2\%$ divergence from the consensus sequences as active in cotton rat.

We used NCBI ORFfinder [105] with the -s 2 (any sense codon) option to detect open reading frames in cotton rat L1 elements and determine amino acid sequences of L1 ORF2 reverse transcriptases. We compared cotton rat L1 ORF2 amino acid sequences against other species' L1 sequences as reported in Dfam database (version 3.8) [49]. Multiple sequence alignments were analyzed using ClustalX version 2.0 [106]. Phylogenetic trees were built using the Bootstrap NJ tree method.

Abbreviations

Bp	Basepairs
Chr.	Chromosome
GO	Gene ontology
HDM	House dust mite
kb	Kilobasepairs
lncRNA	Long noncoding RNA
MyA	Million years ago
N/A	Not applicable
NCBI	National Center for Biotechnology Information
RNA-seq	RNA sequencing
RSV	Respiratory syncytial virus
v/v	Volume-to-volume
SRA	Short-read archive
WGS	Whole genome sequencing

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12915-025-02316-6>.

Additional file 1: Additional Results; Additional Figures S1–S12; Additional Tables S1–S6; Additional References. Additional Figures—Fig. S1, Chromosomal homology in cotton rat vs. mouse. Fig. S2, Orthology and synteny in cotton rat vs. mouse. Fig. S3, Comparisons between cotton rat and white-footed mouse chromosomes. Fig. S4, Comparisons between *S. hispidus*, *O. torridus* and *P. maniculatus*. Fig. S5, Expressed gene counts in cotton rat tissues. Fig. S6, Immune gene families' phylogenetic trees. Fig. S7, Gains and losses of immune genes in human, mouse and cotton rat. Fig. S8, Loss of *Raet/MHA* locus from cotton rat. Fig. S9, Sequence of cotton rat L1. Fig. S10, Sequence of cotton rat ERV2. Fig. S11, Sequence of cotton rat SINE/B2. Fig. S12, Phylogeny of mammalian L1 ORF2s. Additional

Tables—Table S1, Cotton rat pool RNA-seq features. Table S2, Immune response genes. Table S3, Undetected cotton rat immune response transcripts. Table S4, RSV-mediated differential gene expression in cotton rat. Table S5, GO terms of duplicated immune genes. Table S6, Comparison of repetitive element calls in BIUU vs HispidOSU gene assemblies.

Acknowledgements

Next-gen sequencing was performed in the Ohio State University Comprehensive Cancer Center (OSUCCC) Genomics Shared Resource, supported by NCI Cancer Center Support Grant P30CA016058.

Authors' contributions

Conception of the work: DES, SN; design of the work: JL, AC, JG-G, KA, TA, TG, BH, DES, SN; acquisition, analysis or interpretation of data: JL, AC, JG-G, FY, KA, GG, BLN, BF, UPC, SW, MM-H, TK, JCM, TA, TG, BH, DES, SN; creation of new software used in the work: none; drafted or substantially revised manuscript: JL, AC, JG-G, KA, TK, TA, TG, BH, DES, SN. All authors read and approved the final manuscript.

Funding

SN was supported by research funds from the College of Veterinary Medicine, Ohio State University.

Data availability

RNA-seq, WGS data, and the genome assembly and annotations from this study have been deposited in the NCBI repository under BioProject PRJNA720389 with RNA-seq data (SUB157345), WGS data (SUB9369558), and genome assembly [16]. Annotated chromosome-level genome assembly and aligned RNA-seq data will be submitted as BED files for species-specific tracks at the UCSC browser. Genome annotations and metadata are available for visualization and searching via a genome browser and BLAST server [36].

Declarations

Ethics approval

Housing, experimental treatment, and euthanasia of cotton rats as described in this manuscript was reviewed and approved by the Institutional Animal Care and Use Committee at Ohio State University, under protocol 2009A0183.

Consent for publication

Not applicable.

Competing interests

DES was employed at 10x Genomics after linked-read data were generated and analyzed.

Author details

¹European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. ²The Instituto Gulbenkian de Ciência, Oeiras, Portugal. ³Oujiang Laboratory, Zhejiang, China. ⁴New York Genome Center, New York, NY, USA. ⁵Centro Nacional de Análisis Genómico (CNAG), Barcelona, Spain. ⁶Universitat de Barcelona (UB), Barcelona, Spain. ⁷Department of Cancer Biology and Genetics, Ohio State University, Columbus, OH, USA. ⁸Department of Thoracic/Head and Neck Medical Oncology, University of Texas MD Anderson Cancer Center, Houston, TX, USA. ⁹Department of Veterinary Biosciences, College of Veterinary Medicine, Ohio State University, Columbus, OH, USA. ¹⁰Wellcome Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. ¹¹Mechanisms of Disease Program, Institute for Research in Biomedicine, Barcelona, Spain. ¹²Life Sciences Department, Barcelona Supercomputing Center, Barcelona, Spain. ¹³Genomics Shared Resource, Ohio State University Comprehensive Cancer Center, Columbus, OH, USA. ¹⁴National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. ¹⁵Department of Experimental and Health Sciences, Universitat Pompeu Fabra (UPF), Barcelona, Spain. ¹⁶Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain. ¹⁷CIBER de Enfermedades Infecciosas, Instituto de Salud Carlos III, Madrid, Spain. ¹⁸Department of Lymphoma & Myeloma, University of Texas MD Anderson Cancer Center, Houston, TX, USA. ¹⁹Present address: School of Life Sciences and Medicine, Shandong University

of Technology, Zibo, China. ²⁰Present address: Ansa Bio, Emeryville, CA, USA.

²¹Present address: Natera, New York, USA, NY. ²²Present address: Department of Medicine, VA Boston Healthcare System, U.S. Department of Veterans Affairs, Boston, MA, USA. ²³Exponential Molecular, Inc., Palo Alto, CA, USA.

Received: 25 March 2024 Accepted: 27 June 2025

Published online: 18 July 2025

References

- Prince GA. The cotton rat in biomedical research. *AWIC Newsletter*. 1994;5:3–5.
- Niewiesk S, Prince G. Diversifying animal models: the use of hispid cotton rats (*Sigmodon hispidus*) in infectious diseases. *Lab Anim*. 2002;36(4):357–72.
- Green MG, Huey D, Niewiesk S. The cotton rat (*Sigmodon hispidus*) as an animal model for respiratory tract infections with human pathogens. *Lab Anim (NY)*. 2013;42(5):170–6.
- Armstrong C. The Experimental Transmission of Poliomyelitis to the Eastern Cotton Rat. *Sigmodon hispidus hispidus Public Healths Reports*. 1939;54:1719–21.
- Prince GA, Jenson AB, Horswood RL, Camargo E, Chanock RM. The pathogenesis of respiratory syncytial virus infection in cotton rats. *Am J Pathol*. 1978;93(3):771–91.
- Boukhalova MS, Prince GA, Blanco JC. The cotton rat model of respiratory viral infections. *Biologicals*. 2009;37(3):152–9.
- Cullen LM, Blanco JC, Morrison TG. Cotton rat immune responses to virus-like particles containing the pre-fusion form of respiratory syncytial virus fusion protein. *J Transl Med*. 2015;13:350.
- Rajagopala SV, Singh H, Patel MC, Wang W, Tan Y, Shilts MH, et al. Cotton rat lung transcriptome reveals host immune response to Respiratory Syncytial Virus infection. *Sci Rep*. 2018;8(1):11318.
- Strickland BA, Rajagopala SV, Kamali A, Shilts MH, Pakala SB, Boukhalova MS, et al. Species-specific transcriptomic changes upon respiratory syncytial virus infection in cotton rats. *Sci Rep*. 2022;12(1):16579.
- Mejias A, Dima B, Suarez NM, Garcia C, Suarez-Arrabal MC, Jartti T, et al. Whole blood gene expression profiles to assess pathogenesis and disease severity in infants with respiratory syncytial virus infection. *PLoS Med*. 2013;10(11): e1001549.
- Heinonen S, Velazquez VM, Ye F, Mertz S, Acero-Bedoya S, Smith B, et al. Immune profiles provide insights into respiratory syncytial virus disease severity in young children. *Sci Transl Med*. 2020;12(540):eaaw0268.
- Colella JP, Tigano A, MacManes MD. A linked-read approach to mus-eomics: Higher quality de novo genome assemblies from degraded tissues. *Mol Ecol Resour*. 2020;20(4):856–70.
- Ott A, Schnable JC, Yeh CT, Wu L, Liu C, Hu HC, et al. Linked read technology for assembling large complex and polyploid genomes. *BMC Genomics*. 2018;19(1):651.
- Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. Direct determination of diploid genome sequences. *Genome Res*. 2017;27(5):757–67.
- Coombe L, Zhang J, Vandervalk BP, Chu J, Jackman SD, Birol I, et al. ARKS: chromosome-scale scaffolding of human genome drafts with linked read kmers. *BMC Bioinformatics*. 2018;19(1):234.
- Ohio State University. Cotton rat genome consortium, genome sequencing. NCBI BioProject PRJNA720389. 2021. <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA720389/>.
- Parra G, Bradnam K, Korfi I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*. 2007;23(9):1061–7.
- Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015;31(19):3210–2.
- Broad Institute. SigHis_v1_BIUU Genome assembly 1ed. NCBI BioProject PRJNA399433. 2019. <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA399433>.
- Elder FF. Tandem fusion, centric fusion, and chromosomal evolution in the cotton rats, genus *Sigmodon*. *Cytogenet Cell Genet*. 1980;26(2–4):199–210.
- Lilue J, Doran AG, Fiddes IT, Abrudan M, Armstrong J, Bennett R, et al. Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci. *Nat Genet*. 2018;50(11):1574–83.
- Swier VJ, Bradley RD, Rens W, Elder FF, Baker RJ. Patterns of chromosomal evolution in *Sigmodon*, evidence from whole chromosome paints. *Cytogenet Genome Res*. 2009;125(1):54–66.
- Long AD, Baldwin-Brown J, Tao Y, Cook VJ, Balderrama-Gutierrez G, Corbett-Detig R, et al. The genome of *Peromyscus leucopus*, natural host for Lyme disease and other emerging infections. *Sci Adv*. 2019;5(7):eaaw6441.
- Vanderbilt University Medical Center. Reference transcriptome and RSV infection in two species of cotton rats: *Sigmodon fulviventer* and *Sigmodon hispidus*. NCBI BioProject PRJNA816878. 2022. <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA816878>.
- Fuentes D, Molina M, Chorostecki U, Capella-Gutierrez S, Marcet-Houben M, Gabaldon T. PhylomeDB V5: an expanding repository for genome-wide catalogues of annotated gene phylogenies. *Nucleic Acids Res*. 2022;50(D1):D1062–8.
- Stamatakis A. RAXML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014;30(9):1312–3.
- Huerta-Cepas J, Gabaldon T. Assigning duplication events to relative temporal scales in genome-wide studies. *Bioinformatics*. 2011;27(1):38–45.
- Pacifici M, Santini L, DiMarco M, Baisero D, Francucci L, Marasini GG, Visconti P, Rondini C. Generation length for mammals *Nature Conservation*. 2013;5:88–94.
- Hedges SB, Marin J, Suleski M, Paymer M, Kumar S. Tree of life reveals clock-like speciation and diversification. *Mol Biol Evol*. 2015;32(4):835–45.
- Smith SA, O'Meara BC. treePL: divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics*. 2012;28(20):2689–90.
- Thybert D, Roller M, Navarro FCP, Fiddes I, Streeter I, Feig C, et al. Repeat associated mechanisms of genome evolution and function revealed by the *Mus caroli* and *Mus pahari* genomes. *Genome Res*. 2018;28(4):448–59.
- Ponce de Leon Senti E, Diaz, Elva, Muro, Hector, Cuellar Garrido, Luis, Guerra, Juan, Torres, Aurora, Torres, Dolores and Hernandez-Aguirre, Arturo. A Distance Measure for Building Phylogenetic Trees: A First Approach. *Research in Computing Science*. 2017;139:149–62.
- Lilue J, Muller UB, Steinfeldt T, Howard JC. Reciprocal virulence and resistance polymorphism in the relationship between *Toxoplasma gondii* and the house mouse. *Elife*. 2013;2:e01298.
- Nakaya Y, Lilue J, Stavrou S, Moran EA, Ross SR. AIM2-Like Receptors Positively and Negatively Regulate the Interferon Response Induced by Cytosolic DNA. *mBio*. 2017;8(4):e00944–17.
- Al-Shahrour F, Diaz-Uriarte R, Dopazo J. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*. 2004;20(4):578–80.
- HispidOSU Genome Browser and BLAST Server. HispidOSU assembly: Genome browser and BLAST server CNAG2021 [Available from: <https://denovo.cnag.cat/cottonrat>].
- Green R, Wilkins C, Thomas S, Sekine A, Hendrick DM, Voss K, et al. Oas1b-dependent Immune Transcriptional Profiles of West Nile Virus Infection in the Collaborative Cross. *G3 (Bethesda)*. 2017;7(6):1665–82.
- Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol*. 2011;7:539.
- Saito M, Sejima H, Naito T, Ushirogawa H, Matsuzaki T, Matsuura E, et al. The CC chemokine ligand (CCL) 1, upregulated by the viral transactivator Tax, can be downregulated by minocycline: possible implications for long-term treatment of HTLV-1-associated myelopathy/tropical spastic paraparesis. *Virology*. 2017;14(1):234.
- Qi H, Chu V, Wu NC, Chen Z, Truong S, Brar G, et al. Systematic identification of anti-interferon function on hepatitis C virus genome reveals p7 as an immune evasion protein. *Proc Natl Acad Sci U S A*. 2017;114(8):2018–23.
- Li M, Kao E, Gao X, Sandig H, Limmer K, Pavon-Eternod M, et al. Codon-usage-based inhibition of HIV protein synthesis by human schlafen 11. *Nature*. 2012;491(7422):125–8.

42. Sarvestani ST, McAuley JL. The role of the NLRP3 inflammasome in regulation of antiviral responses to influenza A virus infection. *Antiviral Res.* 2017;148:32–42.
43. Feng B, Zhang Q, Wang J, Dong H, Mu X, Hu G, et al. IFITM Expression Patterns Induced by H9N2 Virus and Inactivated Viral Particle in Human Umbilical Vein Endothelial Cells and Bronchus Epithelial Cells. *Mol Cells.* 2018;41(4):271–81.
44. Verhelst J, Parthoens E, Schepens B, Fiers W, Saelens X. Interferon-inducible protein Mx1 inhibits influenza virus by interfering with functional viral ribonucleoprotein complex assembly. *J Virol.* 2012;86(24):13445–55.
45. Bailey CC, Zhong G, Huang IC, Farzan M. IFITM-Family Proteins: The Cell's First Line of Antiviral Defense. *Annu Rev Virol.* 2014;1:261–83.
46. Jung H, Hsiung B, Pestal K, Procyk E, Raulet DH. RAE-1 ligands for the NKG2D receptor are regulated by E2F transcription factors, which control cell cycle entry. *J Exp Med.* 2012;209(13):2409–22.
47. Chalupny NJ, Sutherland CL, Lawrence WA, Rein-Weston A, Cosman D. ULBP4 is a novel ligand for human NKG2D. *Biochem Biophys Res Commun.* 2003;305(1):129–35.
48. Gilfillan S, Ho EL, Cella M, Yokoyama WM, Colonna M. NKG2D recruits two distinct adaptors to trigger NK cell activation and costimulation. *Nat Immunol.* 2002;3(12):1150–5.
49. Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob DNA.* 2021;12(1):2.
50. Smit AH. RepeatModeler Open-1.0. 2008–2015. 2008.
51. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A.* 2020;117(17):9451–7.
52. Corvelo A, Shelton J, New York Genome Center. Compass: Improving assembly contiguity by using locally mapped reads to bridge gaps via De Bruijn graphs. 2023 [Available from: <https://github.com/nygenome/compass>].
53. Grieves JL, Yin Z, Durbin RK, Durbin JE. Acute and Chronic Airway Disease After Human Respiratory Syncytial Virus Infection in Cotton Rats (*Sigmodon hispidus*). *Comp Med.* 2015;65(4):315–26.
54. Green G, Johnson SM, Costello H, Brakel K, Harder O, Oomens AG, et al. CX3CR1 Is a Receptor for Human Respiratory Syncytial Virus in Cotton Rats. *J Virol.* 2021;95(16):e0001021.
55. Kim D, Niewiesk S. Synergistic induction of interferon alpha through TLR-3 and TLR-9 agonists stimulates immune responses against measles virus in neonatal cotton rats. *Vaccine.* 2014;32(2):265–70.
56. Kim D, Niewiesk S. Synergistic induction of interferon alpha through TLR-3 and TLR-9 agonists identifies CD21 as interferon alpha receptor for the B cell response. *PLoS Pathog.* 2013;9(3):e1003233.
57. Carsillo M, Kutala VK, Puschel K, Blanco J, Kuppusamy P, Niewiesk S. Nitric oxide production and nitric oxide synthase type 2 expression by cotton rat (*Sigmodon hispidus*) macrophages reflect the same pattern as human macrophages. *Dev Comp Immunol.* 2009;33(5):718–24.
58. National Center for Biotechnology Information. *Sigmodon hispidus*: Hispid cotton rat (*Sigmodon hispidus*) is a species of rodent in the family Cricetidae. Taxonomy ID 42415. 2025. <https://www.ncbi.nlm.nih.gov/datasets/taxonomy/42415/>.
59. Fichtner AS, Paletta D, Starick L, Schumann RF, Niewiesk S, Herrmann T. Function and expression of CD1d and invariant natural killer T-cell receptor in the cotton rat (*Sigmodon hispidus*). *Immunology.* 2015;146(4):618–29.
60. Stepan S, Adkins R, Anderson J. Phylogeny and divergence-date estimates of rapid radiations in murid rodents based on multiple nuclear genes. *Syst Biol.* 2004;53(4):533–53.
61. Carsillo T, Huey D, Levinsky A, Obojes K, Schneider-Schaulies J, Niewiesk S. Cotton rat (*Sigmodon hispidus*) signaling lymphocyte activation molecule (CD150) is an entry receptor for measles virus. *PLoS ONE.* 2014;9(10):e110120.
62. Pletneva LM, Haller O, Porter DD, Prince GA, Blanco JC. Interferon-inducible Mx gene expression in cotton rats: cloning, characterization, and expression during influenza viral infection. *J Interferon Cytokine Res.* 2006;26(12):914–21.
63. Cordaux R, Batzer MA. The impact of retrotransposons on human genome evolution. *Nat Rev Genet.* 2009;10(10):691–703.
64. Hancks DC, Kazazian HH Jr. Roles for retrotransposon insertions in human disease. *Mob DNA.* 2016;7:9.
65. Chuang NT, Gardner EJ, Terry DM, Crabtree J, Mahurkar AA, Rivell GL, et al. Mutagenesis of human genomes by endogenous mobile elements on a population scale. *Genome Res.* 2021;31(12):2225–35.
66. Erickson IK, Cantrell MA, Scott L, Wichman HA. Retrofitting the genome: L1 extinction follows endogenous retroviral expansion in a group of murid rodents. *J Virol.* 2011;85(23):12315–23.
67. Cantrell MA, Ederer MM, Erickson IK, Swier VJ, Baker RJ, Wichman HA. MysTR: an endogenous retrovirus family in mammals that is undergoing recent amplifications to unprecedented copy numbers. *J Virol.* 2005;79(23):14698–707.
68. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature.* 2021;592(7856):737–46.
69. Martin M. Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads. *EMBnet Journal.* 2011;17:10–2.
70. Marco-Sola S, Sammeth M, Guigo R, Ribeca P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat Methods.* 2012;9(12):1185–8.
71. Song L, Florea L, Langmead B. Lighter: fast and memory-efficient sequencing error correction without counting. *Genome Biol.* 2014;15(11):509.
72. Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, et al. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. *Genome Res.* 2017;27(5):768–77.
73. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. *J Comput Biol.* 2000;7(1–2):203–14.
74. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM 2013 [Available from: <https://arxiv.org/abs/1303.3997>].
75. Yang F, O'Brien PC, Ferguson-Smith MA. Comparative chromosome map of the laboratory mouse and Chinese hamster defined by reciprocal chromosome painting. *Chromosome Res.* 2000;8(3):219–27.
76. University of California Irvine. UCI_PerLeu_2.1 Genome assembly. 2 ed. NCBI BioProject PRJNA281425. 2020. <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA281425/>.
77. Wellcome Sanger Institute. mOncTor1.1 Genome assembly 100 ed. NCBI 2020. 2 ed. NCBI BioProject PRJNA664182. 2020. <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA664182>.
78. Green MG, Petroff N, La Perle KMD, Niewiesk S. Characterization of Cotton Rat (*Sigmodon hispidus*) Eosinophils, Including Their Response to Respiratory Syncytial Virus Infection. *Comp Med.* 2018;68(1):31–40.
79. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013;29(1):15–21.
80. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol.* 2010;28(5):511–5.
81. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVIDENCEModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* 2008;9(1):R7.
82. Gotoh O. A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence. *Nucleic Acids Res.* 2008;36(8):2630–8.
83. Parra G, Blanco E, Guigo R. GenElD in *Drosophila*. *Genome Res.* 2000;10(4):511–5.
84. Stanke M, Schöffmann O, Morgenstern B, Waack S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics.* 2006;7:62.
85. Lomsadze A, Burns PD, Borodovsky M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res.* 2014;42(15):e119.
86. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics.* 2005;21(18):3674–6.
87. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403–10.
88. Zdobnov EM, Apweiler R. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics.* 2001;17(9):847–8.

89. Cui X, Lu Z, Wang S, Jing-Yan Wang J, Gao X. CMsearch: simultaneous exploration of protein sequence space and structure space improves not only protein homology detection but also protein structure prediction. *Bioinformatics*. 2016;32(12):i332–40.
90. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*. 2013;29(22):2933–5.
91. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, et al. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res*. 2015;43(Database issue):D130–7.
92. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*. 1997;25(5):955–64.
93. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol*. 1981;147(1):195–7.
94. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*. 2004;5:113.
95. Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*. 2005;33(2):511–8.
96. Lassmann T, Sonnhammer EL. Kalign—an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*. 2005;6:298.
97. Landan G, Graur D. Heads or tails: a simple reliability check for multiple sequence alignments. *Mol Biol Evol*. 2007;24(6):1380–3.
98. Wallace IM, O'Sullivan O, Higgins DG, Notredame C. M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res*. 2006;34(6):1692–9.
99. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 2009;25(15):1972–3.
100. Gascuel O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol*. 1997;14(7):685–95.
101. Akaike H. Information theory and extension of the maximum likelihood principle. *Proceedings of the 2nd international symposium on information theory*. 1973:267–81.
102. Huerta-Cepas J, Serra F, Bork P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol Biol Evol*. 2016;33(6):1635–8.
103. Benton MJ, Donoghue PCJ, Asher RJ, Friedman M, Near TJ, Vinther J. Constraints on the timescale of animal evolutionary history. *Palaeontologia Electronica*. 2015;18.
104. Smit AH. RepeatMasker Open-4.0. 2013–2015 2013.
105. Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, et al. Database resources of the National Center for Biotechnology. *Nucleic Acids Res*. 2003;31(1):28–33.
106. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. *Bioinformatics*. 2007;23(21):2947–8.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.