SUPPLEMENTAL MATERIALS FOR A META-ANALYSIS OF BAT PHYLOGENETICS BASED ON WHOLE GENOMES AND TRANSCRIPTOMES FROM 18 SPECIES

John A. Hawkins[1,2], Maria E. Kaczmarek[3,4], Marcel A. Müller[5,6,7], Christian Drosten[5,6], William H. Press[1,2,3], Sara L. Sawyer[4,8,*]

[1] Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, TX 78712, USA
[2] Institute for Cellular and Molecular Biology, The University of Texas at Austin, Austin, TX 78712, USA
[3] Department of Integrative Biology, The University of Texas at Austin, Austin, TX 78712, USA
[4] BioFrontiers Institute, University of Colorado Boulder, Boulder, CO 80303, USA
[5] Institute of Virology, Charité-Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Charitéplatz 1, 10117 Berlin, Germany
[6] German Centre for Infection Research (DZIF), Berlin, Germany
[7] Martsinovsky Institute of Medical Parasitology, Tropical and Vector Borne Diseases, Sechenov University, 2-4 Bolshaya Pirogovskaya st., 119991 Moscow, Russia
[8] Department of Molecular, Cellular, and Developmental Biology, University of Colorado Boulder, Boulder, CO 80303, USA
* Correspondence: ssawyer@colorado.edu

Quantification of Genomic-Transcriptomic Data Separation

During manual inspection of uncleaned sequence alignments, we observed non-random separation by data type into genomic and transcriptomic sequences, as described in the text. We here wish to quantify this observation.

The Mismatching Isoform eXon Remover (MIXR) algorithm directly detects and filters exons observed to contain alternate consensus runs. The significant majority of such runs are actually filtered out during the exon-filtering step prior to MIXR filtering (Figure 2E), but those results are mixed with filtering due to incomplete transcript assembly and other artifacts, so we constrain ourselves to this smaller set of observations. Each observed alternate consensus run provides four parameters of interest: N, the total number of species in the alignment; K, the number of genomic species in the alignment; n, the total number of species in the alternate consensus run; and k, the number of genomic species in the alternate consensus run. We take the distribution of N, K from the observed data, omitting alignments with only genomic or transcriptomic species as they contain no relevant information. There are a few options for modeling the distribution of n, from which we take the most conservative: also using the distribution in the data. We then ask whether the observed values of k tend to be different than expected by chance.

We consider two null models. First, the data could be completely random, as colored balls drawn from an urn without replacement. For our problem, this models the hypothesis that isoforms for each species are

www.pnas.org/cgi/doi/10.1073/pnas.1814995116

selected at random, independent of species identity or data type. In this case, k would be distributed according to the hypergeometric distribution.

Second, we model the possibility that alternate consensus runs are in fact real many-amino-acid-long mutations which propagate to all descendant species, thereby creating alternate consensus runs. We call this model tree-random, as the randomness is constrained by the phylogenetic tree. We assume it is safe to consider such mutations unique event polymorphisms, and we assume complete lineage sorting. For this model, we estimate the conditional probability of k as follows. First, we start with the species tree shown in Figure 3. For a given N and K, we randomly sample K genomic and N-K transcriptomic species and prune the tree to contain only those species. We then randomly select a branch of the tree, weighted by the branch length. This branch implies a tree bipartition, the smaller side of which contains a sample draw: n species, k of which are genomic. After generating many such samples, we normalize for each set of N, K, and n to estimate the conditional probability of k.

Finally, we define the genomic-transcriptomic separation statistic, which we label F, to summarize the bias of this distribution for all N, K, and n. We define F to be the normalized distance of k from the expected value under the hypergeometric distribution:

$$F(N,K,n,k) = \frac{1}{Z}\left|k - n\frac{K}{N}\right|$$

where the normalization constant must account for the possibilities of the expected value being closest to zero, n, or K:

$$Z = \max\left(n\frac{K}{N}, \min\left(n - n\frac{K}{N}, K - n\frac{K}{N}\right)\right)$$

We note that while this statistic is inspired by the hypergeometric distribution, it can be used equally well for the tree-random model.

Distributions of the genomic-transcriptomic separation statistic for our data and the two null models are shown in Figure S2, and show the data to be significantly biased toward the separation of genomic and transcriptomic data. Figure S2A shows the distribution for all data and Figure S2B restricts to the distribution of consensus runs with $n \geq 5$, where our summary statistic is more informative. The distribution of the data is significantly shifted to the right, i.e. shifted toward higher separation of genomic and transcriptomic data (KS test, p-value $< 10^{-12}$). For $n \geq 5$, more than 10% of the data is completely separated (F=1), an event with near-zero probability in both null models. By eye, the distribution of the data is very different from hypergeometric, but resembles the tree-random model except for the high-value data. Even excluding the data with $F > 0.75$, however, the distribution is still significantly shifted to the right (KS test, p-value $< 10^{-5}$), consistent with bias throughout the distribution.

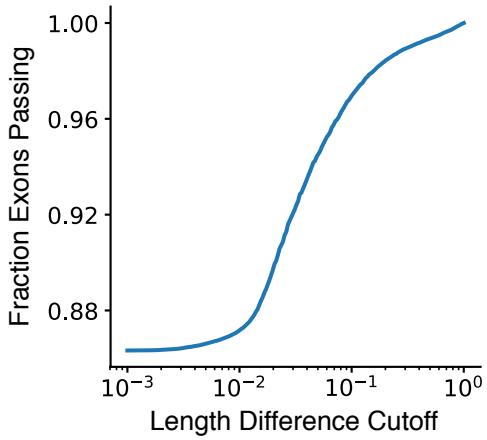

Figure S1. Exons accepted vs. length difference cutoff. The transition to large-length-discrepancy exons is seen to occur at approximately 1% length difference, which we thus used as the cutoff for our exon filtering strategy.

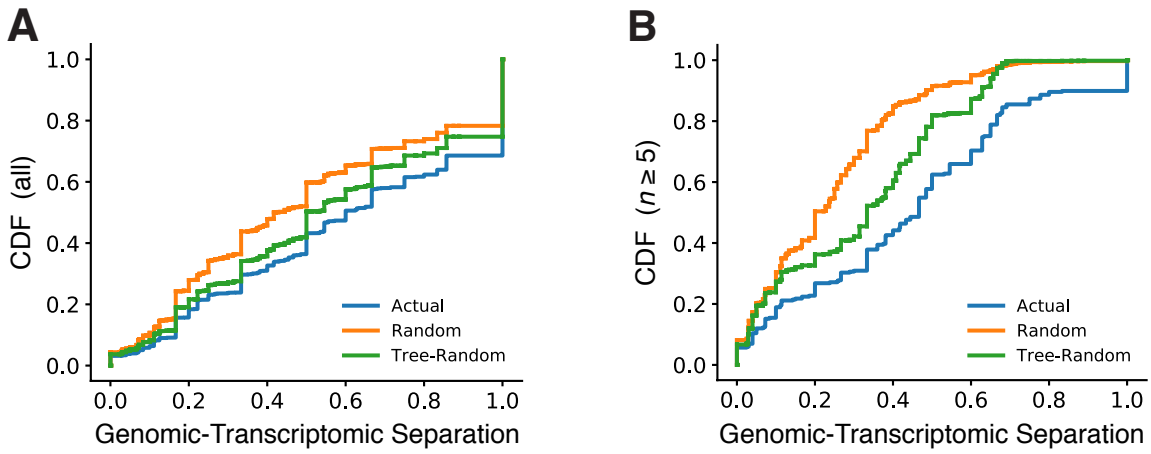

Figure S2. Genomic-Transcriptomic Separation Statistic. The genomic-transcriptomic separation statistic, as defined in the supplemental text, for (A) all alternate consensus runs and (B) alternate consensus runs with at least 5 species.

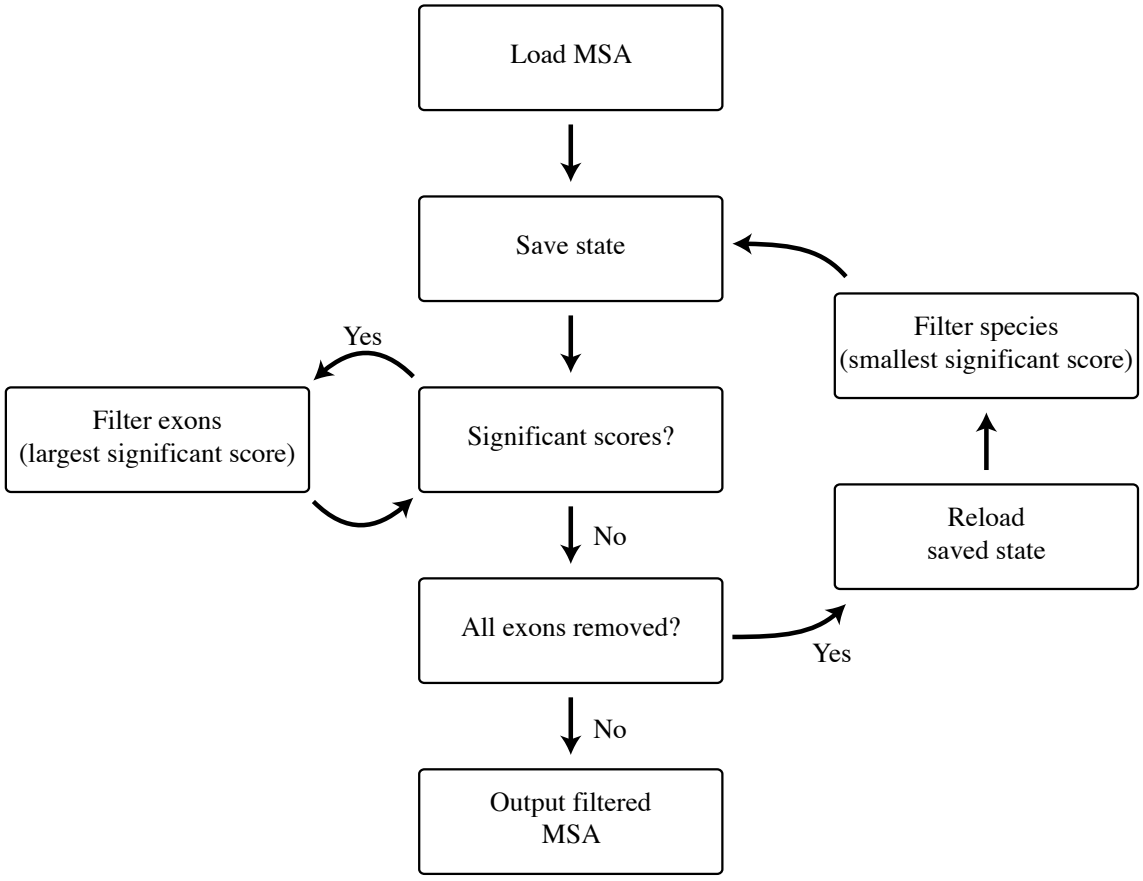

Figure S3. MIXR Flowchart. The flowchart of exon and/or species filtering used by the Mismatching Isoform eXon Remover (MIXR) algorithm. This algorithm attempts to preserve the full set of species except in the event exon removal erases the entire alignment.

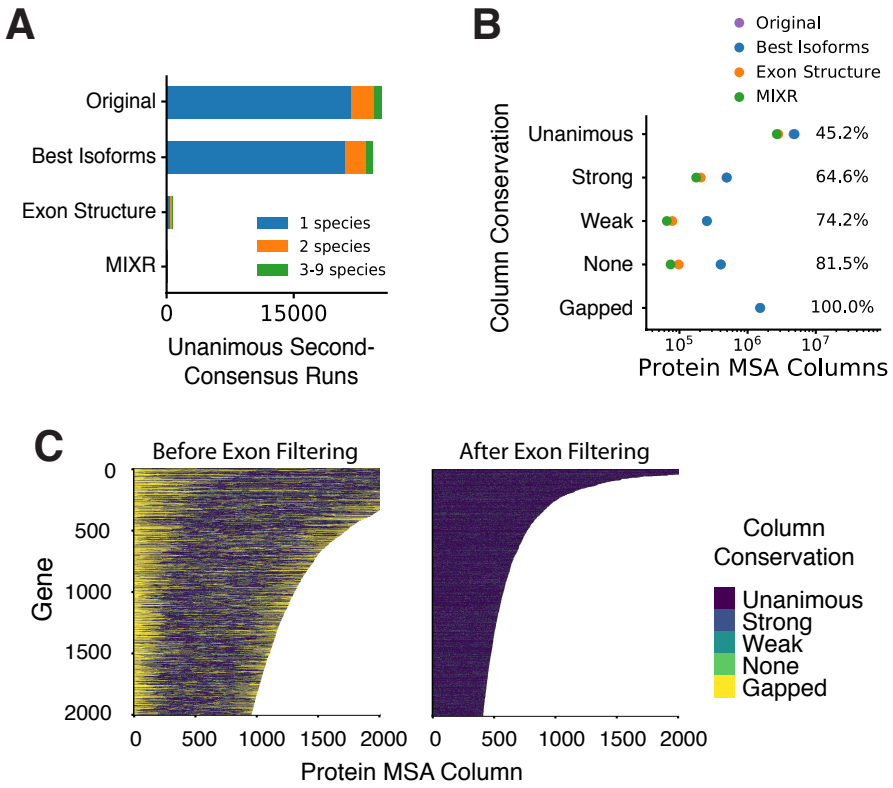

Figure S4. Quality measures after filtering requiring equal-length exons. Figures equivalent to Figures 2E-F, except that the exon length filter required exact matching of exon lengths rather than up to 1% length difference. The results are qualitatively the same as Figures 2E-F, except that fewer unanimous second consensus runs survived the exon structure filtering step (but were filtered by MIXR in both cases) and gaps were removed completely. See caption for Figures 2E-F.

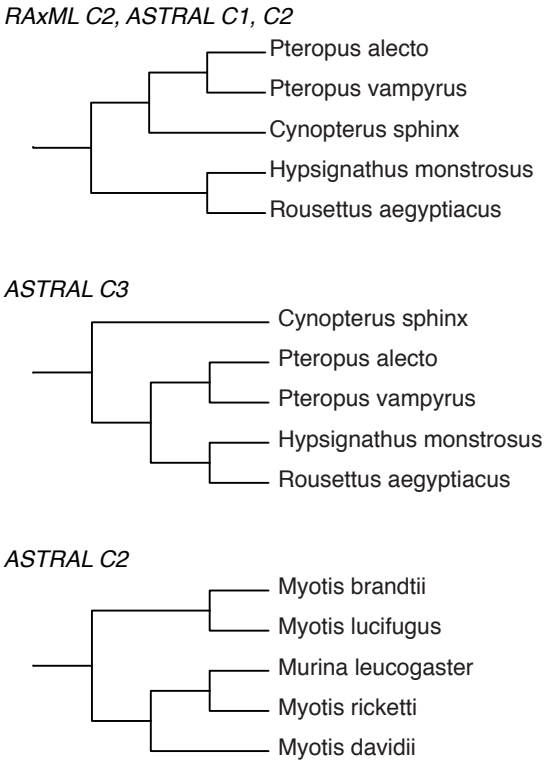

Figure S5: Alternative topologies in our constructed phylogenies. Four of our twelve phylogenetic construction methods disagreed with the consensus tree on one or two nodes, as shown above. Corresponding construction methods shown next to each alternative topology.

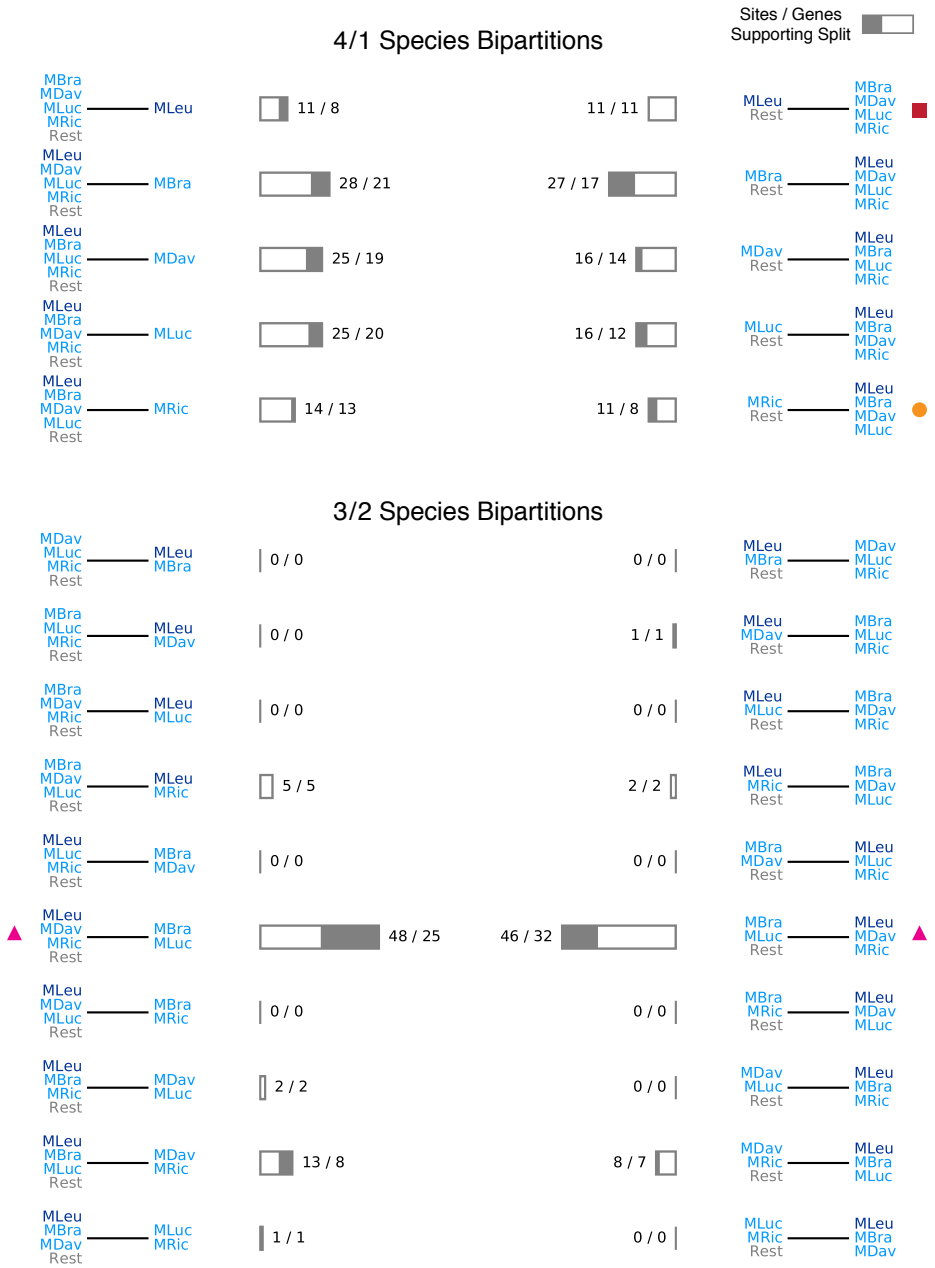

Figure S6: Myotis and *M. leucogaster* bipartition support in the 100 manually inspected genes. Each possible bipartition is shown with a bar graph that shows the number of amino acid sites supporting the bipartition (grey) and the number of genes those sites are found in (white), also shown as text to the side. "Rest" indicates which side of the bipartition agrees with the consensus sequence from the full alignment. The bipartition of *M. leucogaster* as outgroup to Myotis (red square) has less than or equal the number of supporting sites of any other Myotis bat being outgroup to Myotis + *M. leucogaster*, including the transcriptomic *M. ricketti* as outgroup (orange circle). Meanwhile, the bipartitions of *M. leucogaster, M. ricketti,* and *M. davidii* split from the other two are the most supported bipartitions (pink triangles).

Species	Assembly Accession Number	Total bp	Scaffold Count	Scaffold N50
Eptesicus fuscus	EptFus1.0	2,026,629,342	6,789	13,454,942
Myotis brandtii	ASM41265v1	2,107,242,811	169,750	3,225,832
Myotis davidii	ASM32734v1	2,059,799,708	101,769	3,454,484
Myotis lucifugus	Myoluc2.0	2,034,575,300	11,654	4,293,315
Pteropus alecto	ASM32557v1	1,985,975,446	65,598	15,954,802
Pteropus vampyrus	Pvam_2.0	2,198,284,804	36,094	5,954,017

Table S1. Genome assembly accession numbers and statistics.

Species	SRA Accession Numbers
Artibeus jamaicensis	SRR539297
Carollia brevicauda	SRR327705, SRR327706, SRR327707
Cynopterus sphinx	SRR837385
Desmodus rotundus	SRR327702, SRR327703, SRR327704, SRR606899, SRR606902, SRR606908, SRR606911
Hypsignathus monstrosus*	SRR7734571, SRR7734572
Macrotus californicus	SRR1023040
Miniopterus schreibersii	SRR974728, SRR974729, SRR974730, SRR974731, SRR974732, SRR974733, SRR974734, SRR974735, SRR974736, SRR974737, SRR974738, SRR974739, SRR974740, SRR974741
Murina leucogaster	SRR636860, SRR636861, SRR636910, SRR636954, SRR636955
Myotis ricketti	SRR837386
Rhinolophus ferrumequinum	SRR1048140, SRR1048142
Rousettus aegyptiacus*	SRR7735101, SRR7735102
Tadarida brasiliensis	SRR636883, SRR636884, SRR636885

Table S2: Sequence Read Archive (SRA) accession numbers for raw transcriptomic data used. Asterisks indicate species whose data was generated in this study.