Genome analysis

Exon Targeted Retrieval and Classification Toolbox (ExTRaCT): a gene search pipeline to find APOBEC3 Z-domains in novel bat genomes

Brenda Delamonica 1,2,*, Bat1K 21-Families Group 3, Mani Larijani 4, Thomas MacCarthy 1,2†, Liliana M. Dávalos 5

[1]Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11794, USA, [2]Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, NY 11794, USA, [3]DOI 10.5281/zenodo.19001721, [4]Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC V5A 1S6, Canada,[5]Department of Ecology and Evolution, Consortium for Inter-Disciplinary Environmental Research, and Institute for Applied Computational Science, Stony Brook University, Stony Brook, NY 11794, USA

*To whom correspondence should be addressed.

†Deceased.

Abstract

Motivation: Several computation gene search tools exist to identify and annotate an ever-growing body of newly sequenced genomes of different species. Many annotation tools, however, fall short when the target species diverges from well-studied model organisms, and when searching for short genes with multiple copies.

Results: We have developed the Exon Targeted Retrieval and Classification Toolbox, ExTRaCT, an automated pipeline to identify any gene exon with conserved structure in novel species genome assemblies. In the use cases presented here, we applied our search tool to 102 bat genomes to find APOBEC3 gene family members. We show that our homolog search algorithm is efficient (run time average of 5 hours for over 100 genomes), works well with reference sequences distantly related to the target (1 out of 498 misclassifications, 0 false positives and 2 false negatives), and is easy to use. As genomic sequencing becomes faster and more accessible, ExTRaCT has downstream applications in phylogenetic, biochemical and genomic studies. It is a simple computational tool that provides a solution to target gene identification, requiring neither whole-genome-assembly annotations, nor prior knowledge of closely related species.

Availability: https://doi.org/10.5281/zenodo.15769018
Contact: Brenda.delamonica@stonybrook.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

An initiative to sequence all living bat genomes, known as the Bat1K project, aims to discover unique adaptations in bats (Bat1K 21 Family Group, 2026; Morales et al.). However, existing computational tools to identify and annotate gene family members in non-model organism genome assemblies often rely on computationally expensive methods, substantive training, and fail to find small gene copies. Identifying gene family members is difficult in bats, which have evolved through numerous diversification events. Gene homology search tools, such as BLAST, rely on arbitrary hit thresholds which may lead to missing results, and subsequent manual analysis becomes tedious (Nestor et al., 2023). Machine learning annotation and gene prediction tools, such as TOGA (Tool to Infer Orthologs from Genome Alignments) (Kirilenko et al., 2023), can conduct full genome annotations, but may miss sections of complicated multi-domain sequences, especially when the target diverges from the reference, and training on new species requires computational expertise. Furthermore, full annotations may require isoform sequencing with long-read technology to identify complex or missing gene homologues based on RNA, which may not be available for purely computational studies (Jebb et al., 2020). In the case of small antimicrobial peptides in bats, MAKER (Cantarel et al., 2008), an integrated gene prediction pipeline, could only find single gene copies and omitted short

multi-genes (Castellanos et al., 2023). To address gaps in pipelines that optimize for single-copy homologues, we introduce the Exon Targeted Retrieval and Classification Toolbox (ExTRaCT), a streamlined fully computational gene family search pipeline which integrates homologue search tools with a custom algorithm that can accurately identify sequences, relies on a small set of input reference data, requires little experience with complicated techniques and unburdens the user from excessive manual post evaluation (Delamonica, 2025).

The Apolipoprotein B mRNA Editing Catalytic Polypeptide-like (APOBEC) family of cytidine deaminating proteins plays an important role in mammalian immunity (Conticello, 2008; Harris & Dudley, 2015; Pecori et al., 2022). A3s are interferon-stimulated genes (ISG) that act on DNA or RNA viruses by causing C > T point mutations at hotspot motifs to cause changes to viral protein that may be deleterious (Pecori et al., 2022; Stavrou & Ross, 2015). Viruses with A3 mutational signatures, or A3 mutational signatures, or "footprints," have been observed in several human exogenous virus such as herpesviruses, polyomaviruses, hepadnaviruses, and papillomaviruses (Poulain et al., 2020; Shapiro, Krug, & MacCarthy, 2021; Willems & Gillet, 2015) as well as endogenous retroviruses or retroelements (Anwar, Davenport, & Ebrahimi, 2013; Ito, Gifford, & Sato, 2020), zoonotic diseases that have spilled over to humans, including Mpox (Delamonica et al., 2023; Forni et al., 2023; O'Toole et al., 2023) and SARS-CoV-2 (Begum et al., 2024; Kim et al., 2022; Poulain et al., 2020). However, A3s are also known to be inhibited or antagonized by viral proteins (Cheng et al., 2019; Grant & Larijani, 2017; Li et al., 2023; Lochelt et al., 2005; Wood et al., 2009; Zhang et al., 2021). These findings have raised the question of whether A3 activity in the long term is proviral or antiviral, since A3 can play a role in driving viral evolution (Jonathan & Ikeda, 2023). As a result, the endemic pattern of human A3 mutation after viral spillover can *increase* viral diversity (Stavrou & Ross, 2015). A3s are therefore an important immune spillover barrier, albeit one with varying outcomes (Uriu et al., 2021). By understanding A3 characteristics in species known to harbor viruses, such as bats, we may better predict the mutational landscape of disease agents with zoonotic potential.

Bats are known hosts of viral agents that, when spilled over to other mammals, can produce zoonosis and pathogenicity (Guth et al., 2022; Irving et al., 2021). An extremely diverse mammalian order with over 1,500 living species that make up more than 20% of extant mammals spanning the globe, bats have varying physical and ecological phenotypes (Frick, Kingston, & Flanders, 2020; Martínez-Fonseca et al., 2024; Teeling et al., 2005). Efforts to understand bats as viral reservoirs and predict zoonosis come from previous zoonotic spillovers involving viruses that have relatives that circulate or are directly hosted in various bat species, such as Nipah, Marburg, Hendra, SARS and MERS (Becker et al., 2022; Becker et al., 2023; Forero-Munoz et al., 2024; Irving et al., 2021; Pandit et al., 2023; Streicker & Gilbert, 2020; Weber et al., 2023). Bat species vary in the expansion or contraction of immunity genes such as interferons (INFs) (which induce ISGs like A3) (Clayton et al., 2024; Streicker & Gilbert, 2020), antimicrobial peptides (Castellanos et al., 2023), and major histocompatibility complex (MHC) (Moreno Santillan et al., 2021) in ways that are lineage-specific, associated with viral interactions(Vazquez et al., 2025), and sometimes because of differences in diet or diet-related microbiome evolution changing their interactions with potential pathogens. Bat-centered research on viral zoonosis is complicated by bat diversity, varying longevity, data biases and lack of longitudinal data (Carlson et al., 2025; Streicker & Gilbert, 2020; Weber et al., 2023). Nevertheless, two groundbreaking studies (Hayward et al., 2018; Jebb et al., 2020) have endeavored to study A3s in bats, positing that increased copies of A3 in bats may contribute to virus resistance.

The first study (Hayward et al., 2018) analyzed the A3s of two bat species in the Pteropodidae family. They biochemically tested A3 gene expression, mutational activity and viral restriction from transcriptomes identified using BLAST. They found up to 13 possible genes in *Pteropus alecto*. For context, humans have 7 different A3s (A3A/B/C/D/F/G/H) and mice have a single copy (Salter, Bennett, & Smith, 2016). More recently, Jebb et al. (2020) examined six bat species, hereafter referred to as the "six bats" (*Rousettus aegyptiacus* in Pteropodidae and *Rhinolophus ferrumequinum* in Rhinolophidae, both in Yinpterochiroptera, *Phyllostomus discolor* in Phyllostomidae, *Molossus molossus* in Molossidae, and *Pipistrellus kuhlii* and *Myotis myotis* in Vespertilionidae, all in Yangochiroptera). The original six-bat study (Jebb et al., 2020) used isoform sequencing with long-read technology to supplement computational findings by TOGA to identify A3s. The species with the highest A3 count was *Pipistrellus kuhlii*, with a total of 8 distinct genes, although it is unclear if all are active or functional. In contrast, *Rousettus aegyptiacus* had 2 A3s, a marked difference in A3 count from *Pteropus*. It is therefore still unclear whether all bats have more A3s than other mammals or even more than primates, but as more bat genome assemblies become available, we need easy-to-use tools to extract A3 sequences to study their evolution and better understand their role in bat immunity.

Here we expanded these findings by searching for the catalytic Zinc coordinating domains (Z-domains) of A3s in 102 different bat species, including the six bats from Jebb et al. (2020), using ExTRaCT. Salter was the first to classify human cytidine deaminating Z-domains into four types, zinc-dependent deaminase sequence motif (ZDD), attributed to Activation induced deaminase (AID), and the APOBECs A1, A2 and A4, and the three A3 Z-domains, Z1, Z2 and Z3 (Salter, Bennett, & Smith, 2016). A3s can have either one or two catalytic Z-domain regions, while the rest have a single ZDD. Along with theories of expansion events in A3 Z-domains among bats, both Hayward and Jebb propose new bat-specific Z-domain structures, diverging from the canonical motifs described in Salter, Bennett and Smith (2016). Not only is ours the largest study of bat A3 Z-domains to date, but the bat assemblies analyzed include species from all 21 bat families which could further support evolutionary associations with immunity and introduce novel definitions of bat-specific A3 Z-domain motif structures.

Our results support the idea that A3 genes are expanded in some bat lineages. We used freely available computational tools and built ExTRaCT to streamline the process of finding genes in gene families in an easy-to-use suite of Python functions so that it may be used to find other structurally conserved genes for any species. Our algorithm is fast, does not require high computational power or expertise and has lenient user-defined search criteria to facilitate modifications and ensure accuracy of results. We show that our pipeline correctly found the desired sequences and identified copies previously missed by TOGA from the six-bat study. Considering the drawbacks of existing gene family annotation tools, our novel suite, ExTRaCT, has applications in evolutionary, phylogenetic or functional study of genes families beyond bat A3.

2 Methods

Assemblies from 103 genomes were downloaded from the Bat1K database (Bat1K 21 Family Group, 2026; Morales et al.). One species, *Myotis thysanodes*, was excluded from the analysis because the scaffolds were incomplete and none of the results passed quality control. Reference homologue A3 gene nucleotide FASTA files can be found in (NCBI Resource Coordinators, 2016) and the Hiller Lab databases (Jebb

et al., 2020; Perez et al., 2025; UCSC, 2025). All data sources and accession numbers are in Supplementary Materials.

The ExTRaCT pipeline was written in Python version 3.7.10 (Python Software Foundation, 2021) and requires both Biopython (v1.81) modules SeqIO and Align IO (Cock et al., 2009), PyBedTools (v 0.10.0) (Dale, Pedersen, & Quinlan, 2011), HMMER version 3.1b2 (Eddy, 2011). EMBOSS (specifically the program getorf) (v6.6) (Rice, Longden, & Bleasby, 2000), Scipio (v1.4.1) (Keller et al., 2008)MAFFT (v 7.490) (Katoh & Standley, 2013), and RAxML (v8.2.18) (Stamatakis, 2014). The custom functions in ExTRaCT are prep_hmmer.py, which converts a given FASTA to the appropriate input HMMER profile structure for the main algorithm, Gene_search.py, and three post processing functions, Scipio_run.py, table_to_fasta.py and gene_Tree.py, which help analyze the results. ExTRaCT is efficient as it relies on limited information about the gene family and replaces manual activities with programmable automation.

Our approach begins by generating a reference homologue gene profile required for ExTRaCT. We used HMMER (v3.1b2) (Eddy, 2011) instead of BLAST (v2.9) (Camacho et al., 2009) to find homologous sequences because BLAST yielded fewer viable hits and required more manual steps to filter out target non-overlapping sequences. The core search algorithm (Gene_search.py) has four crucial steps which provide nucleotide and amino acid sequences that closely match previously known gene structures in a novel species where gene annotations may not yet be known or provided. The steps are: (1) find homologous sequence start and stop locations, (2) extract the sequence from the genome assembly, (3) identify open reading frames in the sequence and (4) filter protein sequences that match with a given structure. In the first step we use the HMMER function nhmmer to search for nucleotide homologues based on a profile created using hmmbuild (Eddy, 2011). Of these, the HMMER homologue search is the most time-consuming step in the process as it searches through the entire genome. The homologue search may also take longer if there are more known sequences in the FASTA homologue input file. HMMER returns start and end locations of homologues as a bed file, which need to then be extracted by PyBedTools (v0.8.1) (Dale, Pedersen, & Quinlan, 2011), the second step of our pipeline. Given a series of nucleotide sequences, ExTRaCT identifies which hits match a gene protein structure by identifying open reading frames (ORF) (EMBOSS v6.6 getorf) (Rice, Longden, & Bleasby, 2000) within a nucleotide sequence. However, true start and stop locations may change at step (3) and be different from the initial HMMER homologue search results. Lastly, sequences containing the canonical Z-domain motif are saved based on categorizations defined by the user. To ensure all gene copies are extracted, the result list includes all sequences, even duplicated non overlapping sequences within the same species. Overall, the main algorithm can take approximately 5 hours for over 100 species, depending on the size of the input files. However, once the ORF files are generated, there is an option to only test filtering of various motif structures, as ExTRaCT will automatically skip steps 1-3 if the HMMER, bed and ORF files already exist in the expected location. These first three steps do not depend on the motif category file that finally filters the gene hits. That way, a single test to try different categorization methods is easy, requires less memory and can take as little as 5 minutes for approximately 100 species. To optimize processing a large scale of species targets, it is best to split the input FASTA file with target genome paths and run smaller batches in parallel. Sample code is provided in the GitHub repository.

To test performance, we compared hit counts base on inputs by 1) reference species taxa (bat, primate and laurasiatherian), 2) domain specification (single, double or both) and 3) user-designed categorization based on Salter, Hayward and Jebb (motif sets A, B and C, in order). We report three different input strategies for a total of 27 tests in Table 1. We conducted a pairwise comparison for all 27 tests whose summaries can be found in Supplementary Materials. To test accuracy, we then compared the results to known bat A3s found in Jebb et al. (2020).

There are three post processing functions that finetune nucleotide sequence results using Scipio (v1.4.1) (Keller et al., 2008) (custom function Scipio_run.py), convert tables with amino acid and nucleotide sequences into FASTA files (custom function table_to_fasta.py), and generate multiple sequence alignments using MAFFT (v7.490) (Katoh & Standley, 2013) and infer phylogenetic using trees RAxML (v.8.2.18) (Stamatakis, 2014). We chose to keep the Scipio nucleotide track separate from the main algorithm since many pertinent analyses may be conducted with amino acid structure alone. However, we provide the Scipio code to extract accurate nucleotide sequence information, as the start and stop bounds are sometimes lost in step (3) of Gene_search.py and not updated when running getorf. The corrected nucleotide sequences could later be used to express and synthesize proteins to conduct experiments of gene activity and function. The optional post-processing step to examine phylogenetic trees of the output helps to validate which genes may be misclassified. These trees could be used for comparative phylogenetic studies that model the duplication and loss processes by which a given gene has evolved. Both Scipio and tree steps can be skipped if primary protein structure is all that is required.

Because of the known divergence between A3s among bats, we provide a final gene tree of amino acid sequences, rooted after phylogenetic inference using known AID genes as outgroups in FigTree (v1.4.4) (Rambaut, 2018). False positive, or non-homologues, results can be identified if sequences form distinct clades that render a known gene subfamily paraphyletic or polyphyletic (e.g., by being nested within the outgroup clade). In contrast, a false-negative results would be a gene omitted from the results of a known species, which we evaluated manually with the six bats (Jebb et al., 2020). However, misclassification errors can be found using a species tree if a clade comprises different gene categories, i.e., results yield non monophyletic groups. While this analysis can identify homologues, we are unable, without proper phylogenetic comparison to the species tree, to distinguish paralogs (which derive from gene duplication) or orthologs (which derive from speciation). In Figure 1 we removed the AID outgroup and kept all the APOBECs, (A1, A2 and A4 in grey). To test if any of the sequences have been mislabeled and belong to different domain groups, we verify that each of the domains form clades as well.

Various packages in R (v4.3.2) (R Core Team, 2023) were used to plot the phylogenetic tree (gtree (v3.10.0) (Yu, 2020; Yu et al., 2020), ggplot2(v3.5.1) (Wickham et al., 2024), phytools(v2.1-1) (Revell, 2024), ape (v5.7-1) (Paradis & Schliep, 2019), ggmsa (v 1.8.0) (Zhou et al., 2022), Biostrings (2.70.1) (Pagès et al., 2023), seqinr (v4.2-36) (Charif & Lobry, 2007) and msa (v1.34.0) (Bodenhofer et al., 2015)). All tables were reformatted in R using the packages dplyr (v1.1.4) (Wickham et al., 2023), tidyr (v1.3.0) (Wickham, Vaughan, & Girlich, 2023), openxlsx (v4.2.7.1) (Schauberger & Walker, 2024), readxl (v1.4.3) (Wickham & Bryan, 2023), writeexl (v1.5.1) (Ooms, 2024) and purrr (v1.0.2) (Wickham & Henry, 2023).

3 Results

We tested three input strategies, compared the best results to known bat annotations and inferred a gene tree for all 102 bat A3Z genes identified using ExTRaCT.

Table 1. Count of A3 Z-domains found in 102 bat genomes for 27 input tests

Input Clade	Domain Type	A (Salter)	B (Hayward)	C (Jebb)
Bat	Single	121	123	496
	Double	120	122	489
	Both	121	123	496
Primate	Single	120	122	495
	Double	121	120	471
	Both	121	123	495
Laurasiatheria	Single	120	122	487
	Double	121	123	496
	Both	121	123	496

ExTRaCT sequence counts of A3 Z-domain search conducted on 102 novel bat genome assemblies grouped by different input options: species homologue input, domain type and labeling protocol (A, B and C). "Both" means the input includes single and double domain A3s.

3.1 ExTRaCT is robust across different mammalian clades

When considering different clade inputs, for the same domain type and motif set, closely related species should result in similar hits. For domain type both the counts across the different clades are equal under motif set A (121) and B (123) but for motif set C the primate test is missing one hit (primate = 495, chiroptera and laurasiatherian = 496). ExTract has a very low error rate (1/496) when using a different input clade (primate) than the target species, it is otherwise robust. The input sequences need not be of the same order as the target species to garner the same results. Input sequences of the same superorder are only slightly preferred. This is beneficial if no known relatives to a target species have gene information or annotated genomes.

While the counts may be the same, the actual sequences found might differ, so we refer to a pairwise match inspection for all test scenarios (Supplementary Materials). We verified that Laurasiatheria and bat domain type both and motif set C input tests contained all the sequence hits, which we elaborate on in the next section. Furthermore, testing multiple domain types is A3-specific. Catalytic domain configurations vary greatly across species and to capture the correct sequences we found that including all genes (both) is best. For the rest of the analysis and for brevity, we focus on domain type 'both.'

3.2 Motif definition is the most critical and flexible input

What impacts the final count the most is the last step of Gene_search.py which filters and categorizes hits that match a user-provided labeling criterion, motif sets A, B and C. In our test, motif set A has the fewest hits. This may be because the canonical motifs were based on human A3s, have fewer domain categories than those that have since been seen in bats, and are more restrictive protein structures than in Hayward and Jebb. Despite this, motif set B only captured two more sequences than set A. It is possible that despite the inclusion of a new Z-domain in motif B, other changes are more restrictive, limiting the final output. Also, since motif B is only based on two bat species in the same family, and given the known diversity among bats, these categories may not capture sequence diversity of the whole order. The difference in results between motif set C and either motif set A or B is remarkable. Across all test scenarios anywhere from 373-375 more sequences were found in the motif C set test result list.

We found in our pairwise match inspection that two sequences in *Furipterus horrens* were found in the Laurasiatheria reference motif set B test but neither in Laurasiatheria nor in bat motif set C. We verified that this is due to the differences in structure of the Z2B domain motif between Hayward and Jebb. Out of all 27 pairwise inspections, these two sequences were the only sequences that are missing completely from the combined domain Laurasiatheria or bat motif set C test. We consider these false negative results for this test and add them manually to produce our final list of 498 genes across the 103 bat species (Supplementary Materials).

3.3 Validating results with known bat A3s

To ensure that our algorithm finds true A3 Z-domains, and identify false negatives, we count the number of known Z-domains from the six bat species in Jebb et al. (2020) that were found using ExTRaCT in **Error! Reference source not found.**. Motif set C tests returned all the 25 known Z-domains across all clade and domain type tests. However, motif set A and B tests had the same results, with a high rate of false negatives (22/25) and only 3 matching Z-domains to known bat A3s. This further emphasizes the importance of motif categorization. Motif C returns no false negatives because it is the most up to date and relevant structure for bats in different families.

In the final gene list, we note that among the six bat species, an additional eight sequences were found, different from the 25 known bat A3 Z-domains found using TOGA. These may be A3 paralogs or orthologs, depending on when and how the genes diverged. The sequences may be copies from duplication events or speciation events, and they may be non-coding sequences or nonfunctional genes. This is a primary example of the benefit of ExTRaCT over other annotation techniques to better understand the evolution of gene families, as it can find gene family members other conventional computational annotation methods miss.

3.4 Producing a reliable gene list for phylogenetic evolutionary analysis and novel Z-domain discovery in bats

A typical step when conducting a manual gene homology search is to align the sequences, infer a phylogeny and validate hits are in the correct gene family or subfamily. We built table_to_fasta.py and gene_Tree.py to automate this post-processing. This results in Figure 1, a phylogeny of the 498 ExTRaCT sequence hits from the Laurasiatheria combined domain motif set C test with two added sequences (false negative results). Known bat cytidine deaminases are the outgroup (A1, A2, A4=grey). The different Z-domain categories are labeled by color (Z1=purple, Z2=green, Z3=orange). None of the sequences form a clade with the outgroup so we can be confident that all results are monophyletic. This is probably because our algorithm focuses on filtering out by structure rather than a heuristic similarity score threshold, enforcing finer structure similarity.

All the sequences, except one tip labeled Z3 (orange dot) in the Z2 clade, form a clade with the correctly labeled domain categories. This sequence is in the *Nycteris thebaica* (Nt1) assembly and is the second Z3 domain found in that species. Having two Z3 domains is rare (Munk, Willemsen, & Bravo, 2012; OhAinle et al., 2008) as they are thought to be genotoxic (Ito, Gifford, & Sato, 2020; Suspene et al., 2011) so it is

likely that this sequence is mislabeled and a possible misclassification error. In an MSA of the Z3 domains (Supplementary Materials), Nt1 has a unique structure, unlike that of other Z3s, and is unlike any of the Z-domain sequences identified in bats so far. It has both a TWS triplet, a defining distinction of Z3 domains, and a WF couplet in position 9-10, which
is also an indicator of Z2. It is possible this is either a result of recombination or some other evolutionary event. Since most of the sequence is clearly not well aligned to other Z3s and aligns more closely to other Z2s, we relabel it to Z2 in our final gene list. However, this raises the possibility that bats have even more distinct Z-domain types than defined most recently by Jebb et al. (2020).

The 25 known six bat species Z-domain sequences are labeled in brown triangles and the 8 possible paralog or orthologs are labeled by empty pink triangles. We note that many of the known genes and ExTRaCT-identified homologs are closely related, but without a gene tree-species tree reconciliation it is not clear which are paralogs, and which are orthologs. A gene tree-species tree reconciliation could identify duplication, speciation and loss events across lineages over evolutionary time. For example, from our gene list we see that count of Z-domains varies drastically by species, where total counts range from 1-23 total Z-domain hits per species. A phylogenetic reconciliation analysis would shed some light into the evolutionary relationships between the species and A3 genes. Further work could relate these evolutionary dynamics to immunity and host-virus associations.

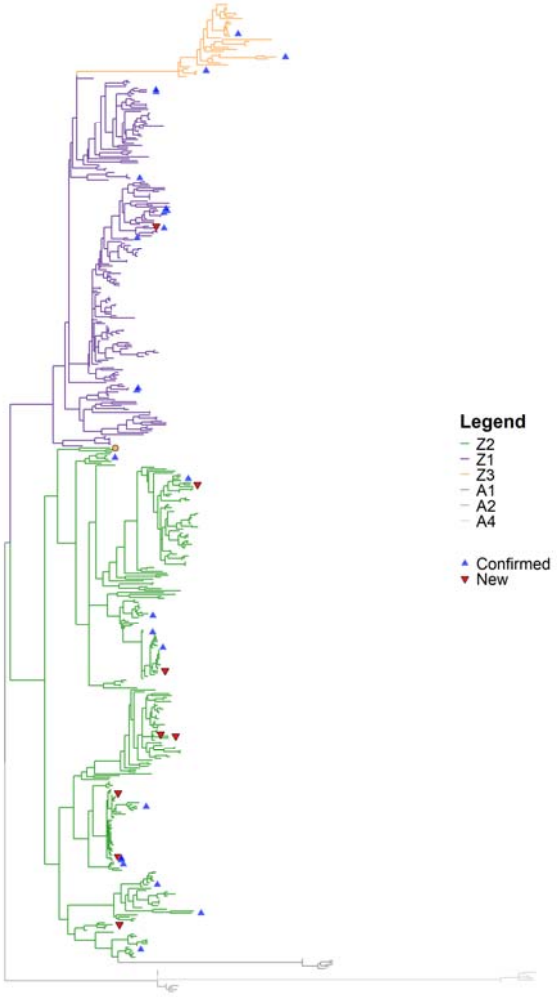

Fig. 1. Gene tree of bat A3 Z-domains. Phylogenetic tree with the final 498 sequences from ExTRaCT pulled from 102 otherwise unannotated bat genomes. Outgroup of known six bat ZDDs of the APOBECs A1, A2 and A3 in grey. Different Z-domain subfamily members are labeled by color (Z1 is purple, Z2 is green and Z3 is orange). Red upside-down triangles indicate hits which are not previously annotated Z domains of known A3s. Blue triangles indicate confirmed A3 Z domains of known six bats.

3.5 ExTRaCT is fast and scalable

Our core algorithm in the pipeline is fast and does not require much memory. Depending on the size of the inputs, a single genome assembly can be analyzed in 2-3 minutes. The largest genome assembly in our test case was *Saccopteryx bilineata*, approximately 2.5 Gb. For a single run, HMMER is the longest step, and it takes around 2 minutes. The subfunctions, pybedtools, getorf and our filtering code, all took less than a minute. The total time to run our pipeline on a genome of 2.5Gb in size was 3 minutes. Depending on computing capabilities, at scale, the run of 103 genomes took an average of 5 hours. Genome sequencing is ubiquitous and there is a growing stream of novel genomes becoming publicly available every day. Out of 498 sequences, our algorithm only misclassified one sequence, demonstrating a low misclassification error rate, zero false positives (each gene subfamily is monophyletic) and two false negatives (found in other tests). Having fast, accurate and easy-to-use methods to conduct targeted annotations of gene families is very beneficial as the landscape of data availability increases.

This order-of-magnitude scaling in annotation of bat A3s has not been done before, and we demonstrate that as genome assemblies become more available, ExTRaCT may be used at a massive scale. There are more genomic data than ever before, but to garner insights into the evolution of gene families requires methods to identify genes. We chose to investigate A3s, a gene family at the forefront of immunity and virus evolution, in bats, a highly diverse clade that includes several species known to be reservoirs of viruses with zoonotic potential and in which A3s appear to have undergone highly dynamic evolutionary changes. With ExTRaCT, others can start to identify A3s or any other salient genes, study the evolution of a gene family or characterize gene expression and classify gene activity. In our case molecular biologists can use our bat A3 results to study A3 and virus coevolution to continue to answer the question of the role A3 mutagenesis plays, whether anti- or pro-viral, in bats. These types of studies could improve the prediction of mutagenesis and recognize viruses with potential risk to humans. By identifying the catalytic regions of bat A3 genes, our case study is the first step in understanding how A3s are involved in the exceptional immunity of bats.

Acknowledgements

Funding

This work has been supported by the grants NIH R01AI132507 to T.M., a Hurst-Della Pietra donation, NSF-OISE 2020577, NSF-IOS 2032063 and NSF-IOS 2217296 to L.M.D., and NFRFE-2021-00933 (Canada) to M.L., L.M.D., and T.M.

Research reported in this publication was supported by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number K12GM102778. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of Interest: none declared.

References

Anwar, F., Davenport, M. P., & Ebrahimi, D. (2013). Footprint of APOBEC3 on the genome of human retroelements. *J Virol*, *87*(14), 8195-8204. https://doi.org/10.1128/JVI.00298-13

Bat1K 21 Family Group. (2026). *Bat1K 21 Family Group* Zenodo. https://doi.org/https://doi.org/10.5281/zenodo.19001722

Becker, D. J., Albery, G. F., Sjodin, A. R., Poisot, T., Bergner, L. M., Chen, B., Cohen, L. E., Dallas, T. A., Eskew, E. A., Fagre, A. C., Farrell, M. J., Guth, S., Han, B. A., Simmons, N. B., Stock, M., Teeling, E. C., & Carlson, C. J. (2022). Optimising predictive models to prioritise viral discovery in zoonotic reservoirs. *Lancet Microbe*, *3*(8), e625-e637. https://doi.org/10.1016/S2666-5247(21)00245-7

Becker, D. J., Eby, P., Madden, W., Peel, A. J., & Plowright, R. K. (2023). Ecological conditions predict the intensity of Hendra virus excretion over space and time from bat reservoir hosts. *Ecol Lett*, *26*(1), 23-36. https://doi.org/10.1111/ele.14007

Begum, M. M., Bokani, A., Rajib, S. A., Soleimanpour, M., Maeda, Y., Yoshimura, K., Satou, Y., Ebrahimi, D., & Ikeda, T. (2024). Potential Role of APOBEC3 Family Proteins in SARS-CoV-2 Replication. *Viruses*, *16*(7). https://doi.org/10.3390/v16071141

Bodenhofer, U., Bonatesta, E., Horejs-Kainrath, C., & Hochreiter, S. (2015). msa: an R package for multiple sequence alignment. *Bioinformatics*, *31*(24), 3997-3999. https://doi.org/10.1093/bioinformatics/btv494

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, *10*, 421. https://doi.org/10.1186/1471-2105-10-421

Cantarel, B. L., Korf, I., Robb, S. M., Parra, G., Ross, E., Moore, B., Holt, C., Sanchez Alvarado, A., & Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res*, *18*(1), 188-196. https://doi.org/https://doi.org/10.1101/gr.6743907

Carlson, C. J., Brookson, C. B., Becker, D. J., Cummings, C. A., Gibb, R., Halliday, F. W., Heckley, A. M., Huang, Z. Y. X., Lavelle, T., Robertson, H., Vicente-Santos, A., Weets, C. M., & Poisot, T. (2025). Pathogens and planetary change. *Nature Reviews Biodiversity*, *1*(1), 32-49. https://doi.org/10.1038/s44358-024-00005-w

Castellanos, F. X., Moreno-Santillan, D., Hughes, G. M., Paulat, N. S., Sipperly, N., Brown, A. M., Martin, K. R., Poterewicz, G. M., Lim, M. C. W., Russell, A. L., Moore, M. S., Johnson, M. G., Corthals, A. P., Ray, D. A., & Davalos, L. M. (2023). The evolution of antimicrobial peptides in Chiroptera. *Front Immunol*, *14*, 1250229. https://doi.org/10.3389/fimmu.2023.1250229

Charif, D., & Lobry, J. R. (2007). *Structural approaches to sequence evolution: Molecules, networks, populations* (U. B. a. M. P. a. H. E. R. a. M. Vendruscolo, Ed.). Springer Verlag.

Cheng, A. Z., Yockteng-Melgar, J., Jarvis, M. C., Malik-Soni, N., Borozan, I., Carpenter, M. A., McCann, J. L., Ebrahimi, D., Shaban, N. M., Marcon, E., Greenblatt, J., Brown, W. L., Frappier, L., & Harris, R. S. (2019). Epstein-Barr virus BORF2 inhibits cellular APOBEC3B to preserve viral genome integrity. *Nat Microbiol*, *4*(1), 78-88. https://doi.org/10.1038/s41564-018-0284-6

Clayton, E., Atasoy, M. O., Naggar, R. F. E., Franco, A. C., Rohaim, M. A., & Munir, M. (2024). Interferon-induced protein with tetratricopeptide repeats 5 of black fruit bat (Pteropus alecto) displays a broad inhibition of RNA viruses. *Front Immunol*, *15*, 1284056. https://doi.org/10.3389/fimmu.2024.1284056

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422-1423. https://doi.org/10.1093/bioinformatics/btp163

Conticello, S. G. (2008). The AID/APOBEC family of nucleic acid mutators. *Genome Biol*, *9*(6), 229. https://doi.org/10.1186/gb-2008-9-6-229

Dale, R. K., Pedersen, B. S., & Quinlan, A. R. (2011). Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics*, *27*(24), 3423-3424. https://doi.org/10.1093/bioinformatics/btr539

Delamonica, B. (2025). *EXon Targeted Retrieval and Classification Toolbox (ExTRaCT)* Zenodo. https://doi.org/10.5281/zenodo.15769019

Delamonica, B., Davalos, L., Larijani, M., Anthony, S. J., Liu, J., & MacCarthy, T. (2023). Evolutionary potential of the monkeypox genome arising from interactions with human APOBEC3 enzymes. *Virus Evol*, *9*(2), vead047. https://doi.org/10.1093/ve/vead047

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Comput Biol*, *7*(10), e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Forero-Munoz, N. R., Muylaert, R. L., Seifert, S. N., Albery, G. F., Becker, D. J., Carlson, C. J., & Poisot, T. (2024). The coevolutionary mosaic of bat betacoronavirus emergence risk. *Virus Evol*, *10*(1), vead079. https://doi.org/10.1093/ve/vead079

Forni, D., Cagliani, R., Pozzoli, U., & Sironi, M. (2023). An APOBEC3 Mutational Signature in the Genomes of Human-Infecting Orthopoxviruses. *mSphere*, *8*(2), e0006223. https://doi.org/10.1128/msphere.00062-23

Frick, W. F., Kingston, T., & Flanders, J. (2020). A review of the major threats and challenges to global bat conservation. *Ann N Y Acad Sci*, *1469*(1), 5-25. https://doi.org/10.1111/nyas.14045

Grant, M., & Larijani, M. (2017). Evasion of adaptive immunity by HIV through the action of host APOBEC3G/F enzymes. *AIDS Res Ther*, *14*(1), 44. https://doi.org/10.1186/s12981-017-0173-8

Guth, S., Mollentze, N., Renault, K., Streicker, D. G., Visher, E., Boots, M., & Brook, C. E. (2022). Bats host the most virulent-but not the most dangerous-zoonotic viruses. *Proc Natl Acad Sci U S A*, *119*(14), e2113628119. https://doi.org/10.1073/pnas.2113628119

Harris, R. S., & Dudley, J. P. (2015). APOBECs and virus restriction. *Virology*, *479-480*, 131-145. https://doi.org/10.1016/j.virol.2015.03.012

Hayward, J. A., Tachedjian, M., Cui, J., Cheng, A. Z., Johnson, A., Baker, M. L., Harris, R. S., Wang, L. F., & Tachedjian, G. (2018). Differential Evolution of Antiretroviral Restriction Factors in Pteropid Bats as Revealed by APOBEC3 Gene Complexity. *Mol Biol Evol*, *35*(7), 1626-1637. https://doi.org/10.1093/molbev/msy048

Irving, A. T., Ahn, M., Goh, G., Anderson, D. E., & Wang, L. F. (2021). Lessons from the host defences of bats, a unique viral reservoir. *Nature*, *589*(7842), 363-370. https://doi.org/10.1038/s41586-020-03128-0

Ito, J., Gifford, R. J., & Sato, K. (2020). Retroviruses drive the rapid evolution of mammalian APOBEC3 genes. *Proc Natl Acad Sci U S A*, *117*(1), 610-618. https://doi.org/10.1073/pnas.1914183116

Jebb, D., Huang, Z., Pippel, M., Hughes, G. M., Lavrichenko, K., Devanna, P., Winkler, S., Jermiin, L. S., Skirmuntt, E. C., Katzourakis, A., Burkitt-Gray, L., Ray, D. A., Sullivan, K. A. M., Roscito, J. G., Kirilenko, B. M., Davalos, L. M., Corthals, A. P., Power, M. L., Jones, G., . . . Teeling, E. C. (2020). Six reference-quality genomes reveal evolution of bat adaptations. *Nature*, *583*(7817), 578-584. https://doi.org/10.1038/s41586-020-2486-3

Jonathan, M., & Ikeda, T. (2023). APOBEC3 family proteins as drivers of virus evolution. *Frontiers in Virology*, *3*. https://doi.org/10.3389/fviro.2023.1332010

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*, *30*(4), 772-780. https://doi.org/10.1093/molbev/mst010

Keller, O., Odronitz, F., Stanke, M., Kollmar, M., & Waack, S. (2008). Scipio: using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. *BMC Bioinformatics*, *9*, 278. https://doi.org/10.1186/1471-2105-9-278

Kim, K., Calabrese, P., Wang, S., Qin, C., Rao, Y., Feng, P., & Chen, X. S. (2022). The roles of APOBEC-mediated RNA editing in SARS-CoV-2 mutations, replication and fitness. *Sci Rep*, *12*(1), 14972. https://doi.org/10.1038/s41598-022-19067-x

Kirilenko, B. M., Munegowda, C., Osipova, E., Jebb, D., Sharma, V., Blumer, M., Morales, A. E., Ahmed, A. W., Kontopoulos, D. G., Hilgers, L., Lindblad-Toh, K., Karlsson, E. K., Zoonomia Consortiumdouble, d., & Hiller, M. (2023). Integrating gene annotation with orthology inference at scale. *Science*, *380*(6643), eabn3107. https://doi.org/10.1126/science.abn3107

Li, Y. L., Langley, C. A., Azumaya, C. M., Echeverria, I., Chesarino, N. M., Emerman, M., Cheng, Y., & Gross, J. D. (2023). The structural basis for HIV-1 Vif antagonism of human APOBEC3G. *Nature*, *615*(7953), 728-733. https://doi.org/10.1038/s41586-023-05779-1

Lochelt, M., Romen, F., Bastone, P., Muckenfuss, H., Kirchner, N., Kim, Y. B., Truyen, U., Rosler, U., Battenberg, M., Saib, A., Flory, E., Cichutek, K., & Munk, C. (2005). The antiretroviral activity of APOBEC3 is inhibited by the foamy virus accessory Bet protein. *Proc Natl Acad Sci U S A*, *102*(22), 7982-7987. https://doi.org/10.1073/pnas.0501445102

Martínez-Fonseca, J. G., Westeen, E. P., Wan, H. Y., Cushman, S. A., & Chambers, C. L. (2024). A global review of landscape-scale analyses in bats reveals geographic and taxonomic biases and opportunities for novel research. *Biological Conservation*, *299*. https://doi.org/10.1016/j.biocon.2024.110829

Morales, A. E., Liang, Y., Thomas, W. R., Leushkin, E. V., Castellanos, F. X., Larkin, D. M., Brown, T., Fromm, B., Hand, S. J., Huang, Z., Hughes, G. M., Jones, M. F., Lim, B. K., Mai, M., Myers, E. W., Pippel, M., Puechmaille, S. J., Simmons, N. B., Abueg, L. A. L., . . . Teeling, E. C. Reference genomes and fossils rewrite bat family-level phylogeny and global biogeography. *In Review. Nature*.

Moreno Santillan, D. D., Lama, T. M., Gutierrez Guerrero, Y. T., Brown, A. M., Donat, P., Zhao, H., Rossiter, S. J., Yohe, L. R., Potter, J. H., Teeling, E. C., Vernes, S. C., Davies, K. T. J., Myers, E., Hughes, G. M., Huang, Z., Hoffmann, F., Corthals, A. P., Ray, D. A., & Davalos, L. M. (2021). Large-scale genome sampling reveals unique immunity and metabolic adaptations in bats. *Mol Ecol*, *30*(23), 6449-6467. https://doi.org/10.1111/mec.16027

Munk, C., Willemsen, A., & Bravo, I. G. (2012). An ancient history of gene duplications, fusions and losses in the evolution of APOBEC3 mutators in mammals. *BMC Evol Biol*, *12*, 71. https://doi.org/10.1186/1471-2148-12-71

NCBI Resource Coordinators. (2016). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, *44*(D1), D7-19. https://doi.org/10.1093/nar/gkv1290

Nestor, B. J., Bayer, P. E., Fernandez, C. G. T., Edwards, D., & Finnegan, P. M. (2023). Approaches to increase the validity of gene family identification using manual homology search tools. *Genetica*, *151*(6), 325-338. https://doi.org/10.1007/s10709-023-00196-8

O'Toole, A., Neher, R. A., Ndodo, N., Borges, V., Gannon, B., Gomes, J. P., Groves, N., King, D. J., Maloney, D., Lemey, P., Lewandowski, K., Loman, N., Myers, R., Omah, I. F., Suchard, M. A., Worobey, M., Chand, M., Ihekweazu, C., Ulaeto, D., . . . Rambaut, A. (2023). APOBEC3 deaminase editing in mpox virus as evidence for sustained human transmission since at least 2016. *Science*, *382*(6670), 595-600. https://doi.org/10.1126/science.adg8116

OhAinle, M., Kerns, J. A., Li, M. M., Malik, H. S., & Emerman, M. (2008). Antiretroelement activity of APOBEC3H was lost twice in recent human evolution. *Cell Host Microbe*, *4*(3), 249-259. https://doi.org/10.1016/j.chom.2008.07.005

Ooms, J. (2024). *writexl: Export Data Frames to Excel 'xlsx' Format*. In (Version 1.5.1) https://CRAN.R-project.org/package=writexl

Pagès, H., Aboyoun, P., Gentleman, R., & DebRoy, S. (2023). *Biostrings: Efficient manipulation of biological strings* https://doi.org/10.18129/B9.bioc.Biostrings

Pandit, P. S., Anthony, S. J., Goldstein, T., Olival, K. J., Doyle, M. M., Gardner, N. R., Bird, B., Smith, W., Wolking, D., Gilardi, K., Monagin, C., Kelly, T., Uhart, M. M., Epstein, J. H., Machalaba, C., Rostal, M. K., Dawson, P., Hagan, E., Sullivan, A., . . . Johnson, C. K. (2023). Author Correction: Predicting the potential for zoonotic transmission and host associations for novel viruses. *Commun Biol*, *6*(1), 25. https://doi.org/10.1038/s42003-022-04364-y

Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3), 526-528. https://doi.org/10.1093/bioinformatics/bty633

Pecori, R., Di Giorgio, S., Paulo Lorenzo, J., & Nina Papavasiliou, F. (2022). Functions and consequences of AID/APOBEC-mediated DNA and RNA deamination. *Nat Rev Genet*, *23*(8), 505-518. https://doi.org/10.1038/s41576-022-00459-8

Perez, G., Barber, G. P., Benet-Pages, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J. N., Hinrichs, A. S., Lee, C. M., Nassar, L. R., Raney, B. J., Speir, M. L., van Baren, M. J., Vaske, C. J., Haussler, D., Kent, W. J., & Haeussler, M. (2025). The UCSC Genome Browser database: 2025 update. *Nucleic Acids Res*, *53*(D1), D1243-D1249. https://doi.org/10.1093/nar/gkae974

Poulain, F., Lejeune, N., Willemart, K., & Gillet, N. A. (2020). Footprint of the host restriction factors APOBEC3 on the genome of human viruses. *PLoS Pathog*, *16*(8), e1008718. https://doi.org/10.1371/journal.ppat.1008718

Python Software Foundation. (2021). *Python Programming Language*. In (Version 3.7.10) Python Software Foundation. https://www.python.org/downloads/release/python-3710/

R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rambaut, A. (2018). *FigTree*. In (Version v1.4.4) University of Edinburgh. http://tree.bio.ed.ac.uk/software/figtree/

Revell, L. J. (2024). phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things). *PeerJ*, *12*, e16505. https://doi.org/10.7717/peerj.16505

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*, *16*(6), 276-277. https://doi.org/10.1016/s0168-9525(00)02024-2

Salter, J. D., Bennett, R. P., & Smith, H. C. (2016). The APOBEC Protein Family: United by Structure, Divergent in Function. *Trends Biochem Sci*, *41*(7), 578-594. https://doi.org/10.1016/j.tibs.2016.05.001

Schauberger, P., & Walker, A. (2024). *openxlsx: Read, Write and Edit xlsx Files*. In (Version 4.2.7.1) https://CRAN.R-project.org/package=openxlsx

Shapiro, M., Krug, L. T., & MacCarthy, T. (2021). Mutational pressure by host APOBEC3s more strongly affects genes expressed early in the lytic phase of herpes simplex virus-1 (HSV-1) and human polyomavirus (HPyV) infection. *PLoS Pathog*, *17*(4), e1009560. https://doi.org/10.1371/journal.ppat.1009560

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312-1313. https://doi.org/10.1093/bioinformatics/btu033

Stavrou, S., & Ross, S. R. (2015). APOBEC3 Proteins in Viral Immunity. *J Immunol*, *195*(10), 4565-4570. https://doi.org/10.4049/jimmunol.1501504

Streicker, D. G., & Gilbert, A. T. (2020). Contextualizing bats as viral reservoirs. *Science*, *370*(6513), 172-173. https://doi.org/10.1126/science.abd4559

Suspene, R., Aynaud, M. M., Guetard, D., Henry, M., Eckhoff, G., Marchio, A., Pineau, P., Dejean, A., Vartanian, J. P., & Wain-Hobson, S. (2011). Somatic hypermutation of human mitochondrial and nuclear DNA by APOBEC3 cytidine deaminases, a pathway for DNA catabolism. *Proc Natl Acad Sci U S A*, *108*(12), 4858-4863. https://doi.org/10.1073/pnas.1009687108

Teeling, E. C., Springer, M. S., Madsen, O., Bates, P., O'Brien S, J., & Murphy, W. J. (2005). A molecular phylogeny for bats illuminates biogeography and the fossil record. *Science*, *307*(5709), 580-584. https://doi.org/10.1126/science.1105113

UCSC. (2025). *Senckenberg Genome Browser*. https://genome.senckenberg.de/

Uriu, K., Kosugi, Y., Ito, J., & Sato, K. (2021). The Battle between Retroviruses and APOBEC3 Genes: Its Past and Present. *Viruses*, *13*(1). https://doi.org/10.3390/v13010124

Vazquez, J. M., Lauterbur, M. E., Mottaghinia, S., Gaucherand, L., Maesen, S., Singer, M., Villa, S., Bucci, M., Fraser, D., Gray-Sandoval, G., Haidar, Z. R., Han, M., Kohler, W., Lama, T. M., Le Corf, A., Loyer, C., McMillan, D., Li, S., Lo, J., . . . Sudmant, P. H. (2025). Extensive longevity and DNA virus-driven adaptation in nearctic Myotis bats. *bioRxiv*. https://doi.org/10.1101/2024.10.10.617725

Weber, N., Nagy, M., Markotter, W., Schaer, J., Puechmaille, S. J., Sutton, J., Davalos, L. M., Dusabe, M. C., Ejotre, I., Fenton, M. B., Knornschild, M., Lopez-Baucells, A., Medellin, R. A., Metz, M., Mubareka, S., Nsengimana, O., O'Mara, M. T., Racey, P. A., Tuttle, M., . . . Reeder, D. M. (2023). Robust evidence for bats as reservoir hosts is lacking in most African virus studies: a review and call to optimize sampling and conserve bats. *Biol Lett*, *19*(11), 20230358. https://doi.org/10.1098/rsbl.2023.0358

Wickham, H., & Bryan, J. (2023). *readxl: Read Excel Files*. In (Version 1.4.3) https://CRAN.R-project.org/package=readxl

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D., Brand, T. v. d., & Posit, P. (2024). *ggplot2*. In (Version 3.5.1) CRAN. https://ggplot2.tidyverse.org/

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A Grammar of Data Manipulation*. In (Version 1.1.4) https://CRAN.R-project.org/package=dplyr

Wickham, H., & Henry, L. (2023). *purrr: Functional Programming Tools*. In (Version 1.0.2) https://CRAN.R-project.org/package=purrr

Wickham, H., Vaughan, D., & Girlich, M. (2023). *tidyr: Tidy Messy Data*. In (Version 1.3.0) https://CRAN.R-project.org/package=tidyr

Willems, L., & Gillet, N. A. (2015). APOBEC3 Interference during Replication of Viral Genomes. *Viruses*, *7*(6), 2999-3018. https://doi.org/10.3390/v7062757

Wood, N., Bhattacharya, T., Keele, B. F., Giorgi, E., Liu, M., Gaschen, B., Daniels, M., Ferrari, G., Haynes, B. F., McMichael, A., Shaw, G. M., Hahn, B. H., Korber, B., & Seoighe, C. (2009). HIV evolution in early infection: selection pressures, patterns of insertion and deletion, and the impact of APOBEC. *PLoS Pathog*, *5*(5), e1000414. https://doi.org/10.1371/journal.ppat.1000414

Yu, G. (2020). Using ggtree to Visualize Data on Tree-Like Structures. *Curr Protoc Bioinformatics*, *69*(1), e96. https://doi.org/10.1002/cpbi.96

Yu, G., Lam, G., Xu, S., Li, L., Jones, B., Silverman, J., Iwasaki, W. M., Xia, Y., & Huang, R. (2020). *ggtree*. In (Version v3.10.0) Bioconductor. https://bioconductor.riken.jp/packages/3.10/bioc/html/ggtree.html

Zhang, Z., Perkovic, M., Gu, Q., Balakrishnan, K., Sangwiman, A., Haussinger, D., Lindemann, D., & Munk, C. (2021). HIV-2 Vif and foamy virus Bet antagonize APOBEC3B by different mechanisms. *Virology*, *554*, 17-27. https://doi.org/10.1016/j.virol.2020.11.013

Zhou, L., Feng, T., Xu, S., Gao, F., Lam, T. T., Wang, Q., Wu, T., Huang, H., Zhan, L., Li, L., Guan, Y., Dai, Z., & Yu, G. (2022). ggmsa: a visual exploration tool for multiple sequence alignment and associated data. *Brief Bioinform*, *23*(4). https://doi.org/10.1093/bib/bbac222