IDENTIFICATION OF SMALL ENDOGENOUS VIRAL ELEMENTS WITHIN HOST GENOMES

by

Edward C. Davis, Jr.

A thesis

submitted in partial fulfillment

of the requirements for the degree of

Master of Science in Computer Science

Boise State University

May 2016

BOISE STATE UNIVERSITY GRADUATE COLLEGE

DEFENSE COMMITTEE AND FINAL READING APPROVALS

of the thesis submitted by

Edward C. Davis, Jr.

Thesis Title: Identification of Small Endogenous Viral Elements within Host Genomes

Date of Final Oral Examination: 04 March 2016

The following individuals read and discussed the thesis submitted by student Edward C. Davis, Jr., and they evaluated his presentation and response to questions during the final oral examination. They found that the student passed the final oral examination.

Timothy Andersen, Ph.D.	Chair, Supervisory Committee
Amit Jain, Ph.D.	Member, Supervisory Committee
Gregory Hampikian, Ph.D.	Member, Supervisory Committee

The final reading approval of the thesis was granted by Timothy Andersen, Ph.D., Chair, Supervisory Committee. The thesis was approved for the Graduate College by John R. Pelton, Ph.D., Dean of the Graduate College.

Dedicated to Elaina, Arianna, and Zora.

ACKNOWLEDGMENTS

The author wishes to express gratitude to the members of the supervisory committee for providing guidance and patience.

ABSTRACT

A parallel string matching software architecture has been developed (incorporating several algorithms) to identify small genetic sequences in large genomes. Endogenous viral elements (EVEs) are sequences originating in the genomes of viruses that have become integrated into the chromosomes of sperm or egg cells of infected hosts, and passed to subsequent generations. EVEs have been identified in all seven classes of viruses and in the species of all kingdoms of life. Viruses from groups V and VI are considered in this thesis, including HIV and Ebola, within host genomes ranging from bacteria to humans. This database of small endogenous viral elements (SEVEs) contains homology between the viruses and every chromosome of the ten multicellular organisms in this study, including human, chimpanzee, gorilla, mouse, fruitbat, nematode, and thale cress.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

EVE – Endogenous Viral Element

SEVE – Small Endogenous Viral Element

ERV – Endogenous Retroviral Element

ENRV – Endogenous Non-Retroviral Element

HERV – Human Endogenous Retrovirus

EDI – EVE-derived Immunity

HIV – Human Immunodeficiency Virus

SIV – Simian Immunodeficiency Virus

HTLV – Human T-cell Leukemia Virus

LINE – Long Interspersed Nuclear Element

SINE – Short Interspersed Nuclear Element

UTR – Untranslated Region

ORF – Open Reading Frame

ncDNA – Noncoding DNA

ncRNA – Noncoding RNA

lncRNA – Long noncoding RNA

mRNA – Messenger RNA

siRNA – Small Interfering RNA

ceRNA – Competing Endogenous RNA

RISC – RNA-induced silencing complex

LTR – Long Terminal Repeat

ENCODE – Encyclopedia of DNA Elements

GPCR – G-protein Coupled Receptor

CD – Cluster of Differentiation

LIST OF SYMBOLS

$\leq$	Less than or equal to
$\mathcal{O}$	Big-O
α	Alpha
β	Beta
μ	Mu

CHAPTER 1

INTRODUCTION

1.1 Endogenous Viral Elements

The sequencing of the complete human genome by the Human Genome Project ranks among the most momentous achievements of modern science. One of the surprising results was the relatively small percentage of actual protein encoding genes, a mere 1.5% according to *Nature* [1]. The remainder consists of various types of noncoding DNA (ncDNA), including introns (approximately 6%), regulatory sequences (8-20%), mobile elements such as long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs), and noncoding RNA (ncRNA) such as the 5' and 3' untranslated regions (UTRs) of mRNAs. Other sequences encode for small RNAs, such as small interfering RNAs (siRNA) and micro RNAs (miRNA).

Among the approximately 73% of noncoding DNA are 5-8% of endogenous retroviruses (ERVs) [2]. Until recently, all of these noncoding sequences in vertebrate genomes were relegated to the dustbin of evolution by being labeled as "junk" DNA. Analyses from the Encyclopedia of DNA Elements (ENCODE) project indicate that the junk DNA hypothesis may be incorrect with up to 80% of the human genome being either actively transcribed, serving a regulatory purpose or being at least biochemically active [3]. David Baltimore, known for his famed virus classification scheme, once quipped that in places the human genome resembled a sea of dead

viruses. Much research has been focused on the study of ERVs because retroviruses, Group VI on the Baltimore classification, are the only class of viruses that must insert their genome into the host chromosome as a requisite part of their life cycle. However, endogenous viral elements (EVEs) have been identified from all seven groups of viruses in the Baltimore classification [4]. The mechanisms for the incorporation of nonretroviral sequences into host genomes are not well understood, but it has been suggested that rogue nucleic acids from viral infections are inserted with the help of the reverse transcriptase and integrase enzymes of the retrotransposons that function similarly to those in retroviruses [5].

1.1.1 Conservation of Endogenous Elements

The conservation of nucleotide sequences over extended evolutionary time scales typically implies some necessary functionality that increases the genetic fitness of the species. This is true of organisms with slow mutation rates, such as vertebrates, as well as those with rapid mutation rates, such as viruses. Once viral genomes are inserted into a host genome, as occurs with endogenous viral elements, the viral sequence assumes the host neutral mutation rate resulting in an approximately million fold rate reduction [6]. Therefore, any highly conserved regions common between relatively slowly evolving eukaryotic organisms and their comparatively fast paced viral antagonists are likely to be significant. It is not sufficient for the sequences of viral invaders to penetrate the nuclei of somatic cells, as the insertion would end with the death of the cell. The establishment of a provirus in the germ cells (e.g., sperm or egg) is necessary to be included in subsequent generations. Following endogenization in the host germ line, the viral elements can exert influence over the evolution of the host. Conservation of EVE sequences common with recent viral isolates could be

particularly significant given rapid viral mutation rates.

Conserved viral sequences could act as agents of infection, aiding and abetting future viruses in the infection of the host by providing homologous targets for viral integration. Alternatively, the sequences might confer some type of antiviral defense or immunity to the host. One potential mechanism could be the production of small RNA molecules derived from the viral sequences that interfere with some stage of viral reproduction when expressed. Such RNA interference has been observed in plants, invertebrates, and mammals [7]. It is also possible that the small RNAs interfere with immune system function of the host. As an example, a recent study by Chuong et al. indicates that ERV transcripts affect the transcriptional regulation of the interferon network (IFN) [8].

There are many constraints imposed on viral sequences, so the location of the sequences within the viral genomes is also important. Viruses must maintain compact genomes to be contained within a tiny icosahedral capsid with a mean diameter of 5 nm [9]. Of particular interest from an immunological perspective are the viral glycoprotein (GP) genes. The glycoproteins form on the viral envelope and in many cases are what allow viruses to evade the immune system and gain entry into the host cell via endocytosis. The glycoproteins enable the immune system to determine the critical distinction between self and non-self. Further increasing the difficulty of integration is that to gain access to the genomic DNA viral sequences must also pass the double nuclear membrane of the eukaryotic cell.

1.1.2 Motivating the Search for EVEs

Human viral diseases play a key role in the human experience, from the annual cycles of influenza viruses to the lethal HIV retroviruses that lead to AIDS, and

Ebola filoviruses that cause human hemorrhagic fever. Other recent epidemics of note include the Rift Valley bunyavirus outbreak in Kenya in 2006 and the H1N1 influenza pandemic of 2009. Measles, caused by the morbillivirus, is another high profile infection enjoying a recent resurgence [10].

Retroviruses are of particular interest due to their innate reverse transcriptase activity and high levels of virulence. A relative of the well known HIV virus is the similarly structured Simian immunodeficiency virus (SIV). Together such immune targeting viruses are members of the lentivirus family. Other retroviruses can even lead to cancer by converting proto-oncogenes into oncogenes. Examples of these include the human T-cell leukemia viruses (HTLV) [11], and the mouse mammary tumor virus (MMTV) [12].

Another class of viruses capable of host genome endogenization are the negative sense single-stranded RNA (-ssRNA) viruses, Group V on the Baltimore classification. They are referred to as negative sense because their genomes are encoded in the 3' to 5' direction, opposite of the 5' to 3' direction of the mRNA transcripts to be translated on ribosomes. Therefore, an RNA replicase enzyme is required to generate the mRNA transcript that will be translated into viral proteins. Examples of -ssRNA viruses include the *Zaire ebolavirus* of the family Filoviridae, the bornavirus of family Bornaviridae, measles of family Paramyxoviridae, and rabies of the family Rhabdoviridae.

Evidence suggests that Baltimore groups V and VI viruses, -ssRNA and retroviruses respectively, are relatively recent evolutionary innovations as they are especially well equipped to attack eukaryotic cells. The discovery of reverse transcriptase activity that reverse transcribes single-stranded RNA into double-stranded DNA was so profound as to require an exception to the central dogma of biology proposed

by Francis Crick (DNA $\rightarrow$ RNA $\rightarrow$ protein). No such mechanism is known to exist in prokaryotes [13] and such viruses target vertebrates specifically. Lentiviruses and filoviruses infect many mammalian species, and bornaviruses are able to infect mammals or birds [14].

Endogenous viral elements can be divided into two broad categories, endogenous retroviruses (ERVs) that rely on retroviral reverse transcriptase encoded by the virus itself for insertion, and endogenous non-retroviruses (ENRVs) that require other means of insertion. The process by which ERVs are endogenized is well understood. The reverse transcriptase enzyme generates the double-stranded DNA copy of the viral RNA genome, complete with identical flanking long terminal repeats (LTRs). The dsDNA copy of the genome is transported to the centrosome along microtubules, enters the nucleus through a nuclear pore, and covalently binds to the genomic DNA. Once integrated, it begins the latent proviral stage of its lifecycle where it is indistinguishable from the host DNA. In this way, the mutation rate slows to the host's neutral rate [15].

The mechanism by which ENRVs are integrated is less well understood, as insertion is not a necessary part of the viral reproductive cycle. The group V -ssRNA viruses, for example, do not possess a reverse transcriptase enzyme. The segmented -ssRNA viruses, such as bunyaviruses, replicate their genomes in the nucleus, but the nonsegmented -ssRNA viruses such as bornaviruses and filoviruses do not. Nevertheless, their genetic sequences have been found in eukaryotic genomes [16]. The -ssRNA viruses instead produce RNA-dependent polymerases to transcribe their genomes into mature mRNA molecules complete with post-transcriptional modifications, including 5' methyl-G caps and polyadenylated 3' tails. The ribosomes within the host cell are then pressed into service to translate the mRNA transcript into viral proteins.

Most likely, it is the mRNA molecules that become endogenized in the germ line rather than the viral genomes themselves. The reverse transcriptase activity of the retrotransposons within the host genome, such as LINEs, are the most likely candidates for endogenization.

Whether EVEs confer immunity or encourage susceptibility to viral infection remains uncertain. One hypothesis suggests they become incorporated into EVE-derived immunity genes (EDI). The Fv1 and Fv4 genes in mice have been shown to act as inhibitors of murine leukemia [17]. Another hypothesis is that the sequences encode for small RNA elements that either work to block viral RNA translation or as miRNAs acting as competing endogenous RNA (ceRNA). Such elements encoded by expressed pseudogenes have been studied with respect to human cancers [18]. Given the established link between viruses and cancer, such a mechanism could also exist in viruses.

The hypotheses discussed above rely on the assertion that endogenous viral elements provide resistance or immunity to the host. The alternative is that they proffer aid to the attacking virion. As mentioned previously, glycoproteins are cell membrane integral proteins that expose a carbohydrate chain into the extracellular fluid, helping to mediate cell-to-cell communication and allowing the cell to be identified by the host immune system. Enveloped viruses exploit this technique by exposing surface glycoproteins, allowing them to masquerade safely within the host organism and initiate fusion with the cell membrane. The HIV virus, for example, exposes the gp120 glycoprotein, allowing it to target the CD4 receptor on the surfaces of helper T-cells [19]. An analogous receptor in the filovirus family is the GP glycoprotein. Sequences similar to the GP have been identified in vertebrate genomes [20]. Homology with the viral nucleoprotein (NP) gene sequences have also been observed. The NP is

responsible for encapsulating the viral genome and so is an obvious frontline target for host immune systems.

CHAPTER 2

LITERATURE REVIEW

2.1 Biological Computation

The relationship between biology and computer science dates back at least fifty years to the 1960s. From Sanger's successful sequencing of the insulin protein, to Watson and Crick's discovery that the DNA molecule is the coding language for life, it quickly became apparent that biomolecules are information carriers, much like a silicon transistor. This revelation created a conceptual link between molecular biology and Shannon's information theory [21]. Zuckerkandl and Pauling compared nucleic and amino acid sequences to semantemes, the fundamental unit of information in linguistics. They coined the term semant*ides*, a fundamental chemical unit [22]. This gave rise to the field of *paleo*genetics, now better known as molecular evolution. The once controversial idea that phylogenetic relationships could be inferred simply from sequence analysis, combined with the advent of the molecular clock, helped form the foundation of the field of bioinformatics [23].

2.1.1 Substitution Matrices

Computational biology originated with Margaret Dayhoff. Working in the FORTRAN language on IBM computers, she developed the first molecular biology database. Another of her innovations was the Percent Accumulated Mutation or PAM sub-

stitution matrix for sequence alignment [24]. PAM matrices remain in use today, with the PAM250 being the most common. The BLOcks SUbstitution Matrix or BLOSUM is another important substitution matrix that applies observed rather than extrapolated local alignment scores as PAM does [25]. The BLOSUM62 matrix is a notable example.

Whereas substitution matrices employ heuristic techniques for pairwise comparisons, the Smith and Waterman algorithm for local sequence alignment always produces the same results. It is a dynamic programming algorithm that makes use of matrices to reward matches and penalize gaps [26].

The FASTA set of programs for pairwise sequence similarity scoring was created by Lipman and Pearson [27]. FASTA programs allow direct comparison of nucleotide and amino acids sequences by performing translation on the fly. The RDF2 program evaluates similarity scores with a shuffling method that permits the preservation of the original sequence. The LFASTA program generates dot matrix plots of similarity greater than a given threshold and supports a number of different scoring matrices. The best sequences are evaluated by collecting the top ten sequences and rescoring them. One of the most enduring contributions of this work is the FASTA file format, which is now ubiquitous within the field of bioinformatics.

2.1.2 Bioinformatics Toolkits

The Basic Local Alignment Search Tool (BLAST), originally developed by Altschul, Gish, and Miller, has become the *de facto* standard software package for performing nucleotide or amino acid sequence comparisons [28]. Following in the footsteps of the Smith-Waterman algorithm for local similarity and the faster heuristic approach of FASTA, BLAST directly approximates optimal local alignments with a maximal

segment pair (MSP) score. BLAST sacrifices the accuracy of Smith-Waterman for speed, but with greater sensitivity than FASTA. The BLAST algorithm filters out low complexity regions (meaning highly repetitive), then converts the query sequence into a k-word list ($k = 11$ for nucleotides, 3 for amino acids). Each word is compared with a substitution matrix such as BLOSUM62 and those obtaining scores greater than a threshold T are kept. The remainder are added to a tree used to search the database for exact matches that are then extended with gaps considered to yield high score segment pairs (HSPs). These high scoring pairs are evaluated for significance and a Smith-Waterman alignment is performed on the highest scoring of all. An e-value is calculated from these alignments based on gap penalties and those achieving a value greater than the threshold E (expected value) are reported to the user.

Inspired by BLAST is the BLAST-like Alignment Tool (BLAT) developed by W. J. Kent, which claims improved accuracy and efficiency over BLAST when performing cross-species comparisons [29]. Speed improvements are claimed of up to 500 times for nucleic acids and 50 times for protein sequences when compared to BLAST. The improvement is attributed to the BLAT technique of indexing non-overlapping k-mers in the genome. The index can be cached in memory in most cases and computed only once per genome. The algorithm uses the index to locate regions of likely homology within the query sequence. It then performs local alignments between homologous regions as in the Smith-Waterman algorithm. The aligned regions are spliced together (much like exons) into larger regions (much like genes). The last step is to revisit the smaller aligned regions to adjust gap boundaries for increased sensitivity. The algorithm is benchmarked against TBLASTX using 1000 mouse genome reads against human chromosome 22 for an average speed increase of 45% and sensitivity increase from 84.5 to 86.7%.

Other noteworthy bioinformatics tools include Thompson's ClustalW method for multiple sequence alignments (MSA) [30], Hidden Markov Models using Bayesian networks [31], and genetic algorithms [32]. For the automated construction of phylogenetic trees, there is Hall's Molecular Evolutionary Genetics Analysis using maximum likelihood (MEGA) [33].

2.2 String Matching

The online exact string matching problem has broad applications in computer science, not merely in computational biology or chemistry but also in text and speech analysis, digital signal processing, databases, and compression. Generally stated, it is the task of finding all occurrences of a pattern string p of length m within a given text t of length n over an alphabet Σ of size σ. The worst case lower bound of the string matching problem is $\mathcal{O}(n)$.

2.2.1 Match Table Algorithms

The first algorithm to achieve that lower bound was devised by Morris and Pratt and hence bears their names. Knuth provided some improvements to the original algorithm and so the eponymous algorithm has three initials [34]. The Knuth-Morris-Pratt (KMP) algorithm maintains a partial match table to prevent the reprocessing of already matched characters. The partial match table is updated whenever a mismatch occurs, allowing the algorithm to skip ahead to the next possible position where a match can possibly occur, thus eliminating backtracking. The construction of the match table for pattern p occurs in $\mathcal{O}(m)$ time and the scanning of the text t requires

$\mathcal{O}(n)$ time for an overall complexity of $\mathcal{O}(m+n)$. Given the obvious assumption that $m \leq n$, the overall time is simplified to $\mathcal{O}(n)$.

The distinction of creating the first string matching algorithm to achieve sublinear average time complexity belongs to Boyer and Moore [35]. The success of the Boyer-Moore algorithm is the innovative revelation that the end of a string (the suffix) should be used to scan for matches rather than the first because it allows more of the text to be skipped. Matches continue back to front until the first character of the pattern is matched. Similarly to the KMP algorithm, a preprocessing table based on the pattern is constructed in linear time $\mathcal{O}(m)$, and is accessible in constant time $\mathcal{O}(1)$. The overall worst case performance of the algorithm is $\mathcal{O}(m+n)$ when the pattern does not occur in the text ($p \not\subset t$), and $\mathcal{O}(mn)$ when the pattern does occur in the text ($p \subset t$).

2.2.2 Hashing Algorithms

The Rabin-Karp algorithm is a solution to the exact string matching problem that employs hashing to find instances of pattern p in text t [36]. The hash function converts a given string to an integer, taking advantage of the fact that the same string will be hashed to the same number as the pattern p. The challenge is dealing with collisions where a non-matching string hashes to the same index as p. Collisions must be resolved by comparing the entire strings. However, the selection of a reasonably good hash function ensures collisions will be infrequent. For practical purposes, this requires the generation of large prime numbers for use in the hashing. The R-K algorithm is best suited to multiple pattern matching and hence is commonly used in plagiarism detection. The worst case time complexity is $\mathcal{O}(mn)$ like many string matching algorithms.

2.2.3 Automata Algorithms

Another class of algorithms capable of achieving sublinear average time complexity are those that make use of factor automata. The automata are data structures that can identify all factors of a given pattern p. The Backward-Oracle-Matching (BOM) algorithm from Allauzen, Rochemore, and Affinot is one of the more efficient examples, particularly for long patterns (large m) [37]. In an attempt to combine the best of both worlds, Faro and Lecroq introduced the Extended-Backword-Oracle-Matching (EBOM) fast string matching algorithm [38]. It is a variant of the Boyer-Moore algorithm with the suffix lookup table replaced with an automata based oracle like BOM. The oracle is a deterministic finite automaton that accepts all of the suffixes of a word. The automaton is built with the reverse of the pattern p in $\mathcal{O}(m)$ time and searches with a sliding window moving from right to left, hence it is a backward oracle match. The worst case time complexity is quadratic $\mathcal{O}(mn)$ like Boyer-Moore, but the average time complexity is $\mathcal{O}(n \log m/m)$.

2.2.4 Suffix Trees

A suffix tree is a data structure that represents all of the suffixes of a string. It is similar to a trie and has applications in many string algorithms, including the exact string matching problem. Ukkonen provides a linear time tree construction algorithm [39]. Each trie in the tree is an automaton as in the Aho-Corasick algorithm. Suffix links take the place of the failure transitions in the automaton. Each node in the tree corresponds to a state in the automaton. The construction proceeds left to right over the text t. States with at least two transitions are branching, states with one transition other than root are implicit, and nodes with no transitions are the leaves.

2.3 Endogenous Viruses

Viral genomes are among the most rapidly evolving in nature. This allows them the flexibility to keep one step ahead of host immune systems, quickly adapting and crossing interspecies boundaries. Such rapid mutation gives researches the opportunity to observe evolution in nearly real-time by sequencing viral isolates. However, it becomes much more difficult to track viral evolution across great expanses of time. Fortunately, viruses tend to leave behind markers of their passage in the genomes of the hosts they infect. These molecular "fossils" can be analyzed by viral archaeologists to gain a greater understanding of both viral and host evolution.

2.3.1 Virology

In one of the seminal papers in the field of virology, Baltimore proved to be incredibly prescient considering the limited amount of data available at the time [40]. He provided a group-based classification system of viruses derived from the nature of their genetic material. Class I consists of all viruses with double-stranded DNA (dsDNA) and Class II encompasses those with single-stranded DNA (ssDNA). The genomes of these DNA viruses can be directly transcribed by the host cell machinery. Class III and Class IV consist of double-stranded and single-stranded RNA (dsRNA and ssRNA, respectively). Class IV viruses require a template strand to be synthesized before transcription can occur. Class V contains the negative sense ssRNA viruses, with single strand genomes that are the inverse of mRNAs, and therefore must carry an intermediate RNA polymerase to enable transcription of mature mRNA molecules. Class VI includes RNA viruses that encode their genomes via a DNA intermediate, now known as retroviruses, but previously known as tumor viruses due to their

association with cancer. A Class VII for pararetroviruses (e.g., hepatitis B) was later added, but otherwise Baltimore's system of classification has required very little modification.

2.3.2 Endogenous Retroviruses and Nonretroviruses

In one of the early treatments of this topic, Katzourakis and Gifford provide a rather exhaustive analysis of endogenous viral element integration in animal genomes, both retroviral and nonretroviral [14]. They performed *in silico* analysis (i.e., BLAST searches) in a wide array of animal and viral genomes. Homologous sequences were observed between DNA, RNA, and RT viruses, within animal hosts ranging from insects to vertebrates, including mammals and birds. Both nuclear and cytoplasmic replicating viruses were covered. Phylogenetic analyses were also performed with wide ranging results. The function of EVEs and whether they are advantageous or deleterious to the host remained unanswered.

Noting that 8% of the human genome is composed of endogenous retroviral elements, Horie et al. set out to determine the extent to which nonretroviral elements are also endogenized [41]. They found that nonsegmented negative-sense RNA viruses such as bornavirus and ebolavirus also have this potential. Sequences homologous to the bornavirus nucleoprotein (NP) gene were identified within several species of mammals, including humans and other primates, rodents and even elephants. The phylogenetic analysis indicates that these elements can be traced back to insertions that occurred more than 40 million years ago (Mya). These results indicate that not only are nonretroviral endogenizations possible, but they have taken place numerous times throughout evolutionary history.

In a related work by Horie, et al., the authors perform a comprehensive search for endogenous bornavirus-like elements (EBLs) [16]. Despite being nonsegmented -ssRNA viruses that replicate their genomes in the cytoplasm, bornaviruses can cause persistent infections in the nuclei of host cells. This means their mRNA transcripts also find their way into host genomes via endogneous germline integrations just like their retroviral counterparts. They provide a review of the presence of EBLs in eukaryotic genomes, including invertebrates. In terms of host function, they note the existence of endogenous nucleoprotein sequences in mice impacting the murine leukemia virus, the remnants of open reading frames in primates, and the fragments of *env* genes in endogenous retroviruses that resulted in the development of placental mammals. Experiments were even conducted to insert modern bornavirus DNA into cultured mouse cells, albeit with limited success.

Lee et al. also considered the ERV-L mutation that gave rise to the mammalian placenta [15]. The authors conducted a study tracking ERV lineage back to 104-110 Mya. Other sequences, selfish genetic elements (SGEs), are found inserted within the ERV sequences. For example, the ERV-L endogenous retroviral gene has homologs in the chromosomes of four mammalian species, including boar, horse, chimpanzee, and human (on chromosome 17). The study included multiple bioinformatics methods, such as BLAST, MUSCLE, Needle, and RepeatMasker.

Belyi, Levine, and Skalka focused on the endogenous viral elements derived from -ssRNA viruses (group V), such as bornavirus and ebolavirus, in a similar study [5]. Previously, only retroviruses (group VI) were known to exist in animal genomes (ERVs). The authors identified at least 80 nonretroviral elements (ENRVs) within the genomes of 19 vertebrate species. Most of the elements originated from viruses that cause neurological disease (bornavirus) or hemorrhagic fevers (ebolavirus). Based

on the tell-tale signs of poly-G caps and poly-A tails surrounding the elements, they were identified as former viral mRNA transcripts that had been endogenzed, likely with the help of the reverse transcriptase enzyme from retrotransposons such as long interspersed elements (LINEs). The estimates of the number of integrations are admittedly low due to the limitations of the bioinformatics techniques applied in the analysis.

In a recent review paper, Aiewsakun and Katzourakis explain that endogenous viral elememts from all seven viral groups from the Baltimore classification have been identified within the genomes of eukaryotic organisms [4]. They provide several different dating techniques to trace viral-host interaction routes throughout evolution. This can be accomplished by comparing orthologs or paralogs, assuming the host neutral mutation rate, and augmentation with geographic data from known host migration patterns.

2.3.3 Potential Functions of Endogenous Viruses

Aswad and Katzourakis later turned their attention to one of the potential functions of EVEs, virally derived immunity, asserting that EVEs incorporated into the germline and then passed to progeny via horizontal gene transfer are chronicles of the ongoing battle between viruses and hosts [17]. Recent advances in genomic sequencing and bioinformatics technology make it possible to properly study this evolution and led to increased opportunities to study EVEs *in silico*. The paper focuses on functional viral derived genes with intact open reading frames in multiple species. Those that may act as inhibitors of viral infection are dubbed EVE-derived immunity genes or EDIs. The genes are identified in several animal species, including fruitfly, mouse, cat, sheep, and bat. The EDI gene functions are categorized as either blocking viral entry

into the cell (glycoprotein) or disrupting viral replication (*gag* genes) and immune system anticipation (super antigen or *sag* genes). The functions of many other EVEs remain unknown.

In addition to EDIs, another potential function of EVEs is the encoding of small interfering RNAs (siRNAs) or micro RNA (miRNAs), one of the emerging areas of study in genetics. They appear to be derived from pseudogenes (genes that have lost their regulatory sequences) or other noncoding regions. According to Kalyana-Sundaram et al., the traditional model of post-transcriptional modification may be incomplete [18]. The model holds that introns are excised from RNA transcripts, leaving only the exons to be spliced together in various combinations by the spliceosome to form mature mRNA transcripts. However, endogenous siRNA or miRNA binding sites may provide another level of control. Analysis of miRNA recognizing elements (MREs) in pseudogones has been limited by their similarity to analogous sequences encoding genes. The authors provide an analysis of pseudogene transcription from 280 normal tissue samples and thirteen cancerous ones. They found pseudogene expression to be prevalent, even ubiquitous, and in some cases possibly cancer-specific. They propose a connection to the recently discovered competitive endogenous RNA (ceRNA) networks in the transcriptome. Although this work did not cover viruses specifically, EVEs could be potential ceRNAs given the relationship between retroviruses and cancer (e.g., HTLV).

The `miRBase` created by Kozomara and Griffith-Jones is intended to be the primary online repository for micro RNA sequences and annotation [42]. The latest version from 2014 contains 17,000 sequences from more than 140 species. The database is searchable by sequence, experiment, tissue, and stage. The project objectives are to be human readable and computer parsable.

Sagan and Sarnow determined siRNAs to be involved with antiviral mechanisms, establishing their role in silencing the expression of viral genes and therefore conferring immunity to a host cell [43]. Such RNAs are evolutionarily conserved and triggered by the presence of double-stranded RNA (dsRNA), which is often viral. Hence, they are cleaved by the DICER complex and bind complementary mRNA transcripts to prevent translation, effectively silencing the corresponding gene. This process had previously been shown to provide immunity in plants and invertebrates, but this was the first confirmation of the same function in mammals.

CHAPTER 3

METHODS

3.1 Species Selections

Exhaustive identification of all potential endogenous viral elements in a particular genome is challenging due to the considerable size of vertebrate genomes and the rapid mutation rate of viruses. Here a sample size is generated by subdividing several viral genomes into small fragments of only 18 base pairs, or about the size of a siRNA sequence, with a step size of three base pairs. All of the chromosomes of a given host genome are then scanned for all occurrences of each viral fragment sequence. The primary objective of this research has been to assemble an initial database containing a representative sample of all small endogenous viral elements across multiple viruses and multiple host genomes.

The focus has been on viruses from groups V and VI of the Baltimore classification. The lentiviruses HIV-1 and SIV were selected to represent the retroviruses. These viruses were chosen because their genomes have been well studied and endogenous retroviruses are already known to encompass 8% of the human genome. In terms of potential clinical importance, there are also primates known to possess immunity to SIV, such as the sooty mangabey [44].

Family Filoviridae is represented in this study by the infamous *Zaire ebolavirus*, both because of the attention drawn by the recent 2014 outbreak in West Africa,

and its presence as a blood-borne pathogen. For a nonretrovirus to be endogenized, its genome or the mRNA produced from it must be present within the nucleus of a sperm or egg cell while a retrotransposon is active (assuming that the aforementioned ENRV insertion hypothesis is correct). Such an event seems much more likely to occur in the presence of a virus that can be sexually transmitted and is therefore already in the vicinity of the gonads where germ cells reside. The morbillivirus responsible for measles infections was selected as another representative from group V (-ssRNA viruses) as something of a control against the ebolavirus, as it is known to be highly infectious to humans but not known to be sexually transmitted.

The Influenza A (H7N9) virus from the Orthomyxoviridae family was the final selection from group V due to the long, complex history between influenza and humans. The influenza genome is segmented, whereas the ebolavirus and morbillivirus genomes are not. The H7N9 genome is from a 2013 outbreak of H7N9 in China. The virus is known to infect birds as well as mammals (the A is for avian). Particularly virulent strains emerge when genetic recombination occurs between avian and mammalian versions of influenza [45].

The set of host genomes selected for inclusion in this study reveals a strong primate bias. The first on the list was the GRCh38 version of the human genome, consisting of 22 autosomal chromosomes along with the sex chromosomes X and Y. The genomes of the nearest living genetic relatives of *Homo sapiens* have also been analyzed, including the chimpanzee (*Pan troglodytes*), the gorilla (*Gorilla gorilla*), the orangutan (*Pongo pygmaeus*), and the gibbon (*Nomascus leucogenys*). The most recent common ancestor of humans and chimpanzees dates to at least 13 million years ago or as early as 4 Mya.

In order to provide more inclusive coverage of living systems, several of the model

organisms from more distant branches of the current phylogeny are also included. Arguably the best studied member of class Mammalia is the house mouse, also known as *Mus musculus*. The GRCm38.3 version of the mouse genome was the most recent at the time of this writing. Venturing away from class Mammalia and even phylum Chordata, within the Ecdysozoa are the phyla Nematoda and Arthropoda. Drawn from them are the genomes of the nematode worm *C. elegans* and the pioneering fruitfly of T.H. Morgan's lab, *Drosophila melanogaster*.

Representing the other two eukaryotic kingdoms Viridiplantae and Fungi are the model genomes of the thale cress plant, *Arabidopsis thaliana*, and the haploid yeast *Saccharomyces cerevisiae*. The *E. coli* bacteria is the representative model organism for all prokaryotes.

The work of Pourrut et al. shows that fruit bats may act as reservoirs for the *Zaire ebolavirus* [46]. The black flying fox or *Pteropus alecto* is known to be a host for the Ebola and Nipah viruses, both from group V of the Baltimore classification, and the SARS virus from group IV (+ssRNA) [47]. The black flying fox genome has been sequenced but not annotated, meaning that the raw sequencing data are available as scaffold files, but have not been compiled into chromosomes [48]. Nevertheless, the number of viral matches for the flying fox is of interest from an emerging infection point of view, and so have been included in the study.

3.2 Computational Approach

Several software components were designed and implemented to conduct this study. The first component is an application for scanning the NCBI FTP server and downloading genome files, called the `FTPScanner`. The second is a massively parallel

string search tool called the `GenomeScanner`, and the last a set of tools for storying and analyzing the output data called the `MatchDatabase`.

3.2.1 FTPScanner

One of the common places where genomes are stored for bioinformatics research is on the NCBI FTP site.[1] The files are freely available for any industrious coder to download. The two primary file formats are FASTA (.mfa or fa extension) and Genbank (.gbk extension). The FASTA format is the simplest, consisting of a single comment line beginning with a greater than character followed by a description. All other lines are sequences of nucleotides or amino acids. Genbank files contain more detailed information, including annotations, but are more complex and therefore more difficult to parse.

The contents of the FTP server are arrayed in a sprawling filesystem with many directories and subdirectories. In order to simplify the navigation of this hierarchy, the first software component constructed for this project was the aptly named FTPScanner and is implemented in the Java language. Its purpose is to scan the contents of the FTP server for genome files in FASTA and Genbank format and download or update the files on the local filesystem if desired. Upon encountering a new species, the code performs an automated Wikipedia search to fetch pertinent information about the organism such as kingdom, phylum, class, etc., and then stores the information in a JSON file database, along with the paths to the genomic files.

Once the required genomic data have been acquired from the NCBI FTP server, they can be passed to the next component, the genome scanner.

[1] `ftp://ftp.ncbi.nlm.nih.gov/genomes`

3.2.2 GenomeScanner

The primary component of the software framework implemented for this project is the `GenomeScanner`, a massively parallel string parser and searching engine. Built to be as fast as possible, it is implemented in the C++ programming language. Linux is the target operating environment but the source code could certainly be compiled for another platform.

The user interface is command line driven with an input argument pointing to the data file path where the genome files are stored on the filesystem (`-fp`). The second input argument is a text file containing a line delimited list of viral genome files in FASTA format (`-vf`). The third argument is a similar list containing all of the host genome files to be searched (`-hf`).

Table 3.1: Viral Genome Sizes

Viral Genome	Size (bp)
Human immunodeficiency virus 1	9181
Simian immunodeficiency virus	9519
Zaire ebolavirus	18922
Measles morbillivirus	15895
Influenzavirus A (H7N9)	13441

The scanner iterates over each viral genome file in the list once for each host genome file and creates a new instance of the `GenomeScanner` class for each file pair. The viral genomes are read directly into memory as they tend to be comparatively small. Table 3.1 indicates the sizes of each of the viral genomes included in this study.

After reading the viral genome into memory, the `Scan` method iterates over the viral sequence and generates subsequences of size k where k is the length of the small endogenous viral elements to be identified. The default size of k is 18 but any desired integer value can be specified from the command line using the `-sub` argument. There

```
1:   GGTCTCTCTGGTTAGACC
2:      CTCTCTGGTTAGACCAGA
3:         TCTGGTTAGACCAGATCT
...
n-2: AATAAAGCTTGCCTTGAG
n-1:    AAAGCTTGCCTTGAGTGC
n:         GCTTGCCTTGAGTGCTTC
```

Figure 3.1: Example of the first three and last three viral subsequences of the HIV-1 virus given parameters of k=18 and s=3.

is also a step size parameter s that configures the scanner to step ahead by s characters (nucleotides) when generating the next subsequence. The default value of s is three, but other values can be specified from the command line with the `-step` argument provided that s is less than k. Figure 3.1 illustrates an example of this technique by indicating the first three and the last three subsequences generated from the HIV-1 genome. The default step size of three was selected due to the large number of host genomes and viruses selected for the study.

The host genomes are vast compared to the tiny viral ones. The primate genomes (human, chimpanzee, gorilla, and orangutan) all contain approximately 3×10^9 base pairs. This is not an unreasonable amount of data to be read into memory on any machine with sufficient memory, but not without sacrificing some parallel processing capabilities. For that reason, the `GenomeScanner` reads the host genome files in discrete blocks. Host files are generally stored as one file per chromosome because a chromosome is simply one continuous DNA molecule. The blocks are measured in number of lines of a FASTA file. The block size defaults to one hundred thousand (1×10^5) but can be specified from the command line via the `-bs` argument. Since a typical FASTA line has a length of 70 characters and each character consumes one

byte of space, the average size of a host genome block is about seven megabytes.

With the input data properly subdivided into viral subsequences (the proverbial needles) and host genome blocks (the corresponding haystacks), the actual substring matching can be performed. This problem is essentially an instance of the online exact string matching problem. An excellent review of the problem space has been provided by Faro and Lecroq [49].

The `GenomeScanner` contains implementations of several algorithms designed to solve this problem. These include preprocessing algorithms such as the Knuth-Morris-Pratt (KMP) algorithm [34] and the Boyer-Moore (BM) algorithm [35]. Also included is the Rabin-Karp (RK) randomized algorithm [36] and an implementation of Ukkonen's online suffix tree construction algorithm [39]. Preliminary benchmarks indicated that the implementation of Faro and Lecroq's extended backward oracle match algorithm (EBOM) [38] yielded the fastest search results within the current architecture. The string matching algorithm to be applied can also be specified from the command line using the `-sa` argument. The EBOM algorithm is the default option due to the performance.

Even the most efficient string matching algorithm would be limited by the input size. The time complexity of this particular problem has four determining factors: (1) the size of the host genome h, (2) the size of the viral genome v, (3) the subsequence length k, and (4) the subsequence step size s. Solving this problem would be quite time intensive in a strictly sequential context, but fortunately it can be considered embarrassingly parallel as little or no communication is required between concurrently running tasks. An overview of the architecture is given in Figure 3.2.

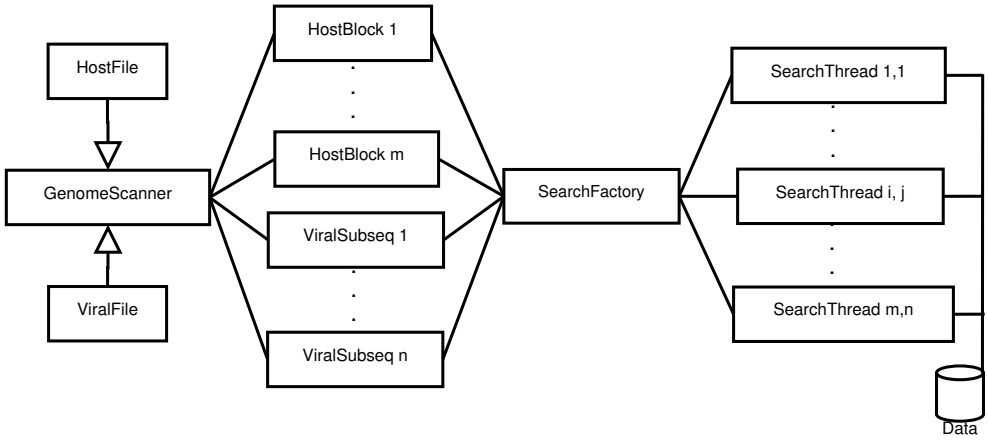

Figure 3.2: `GenomeScanner` Parallel Architecture.

The parallel architecture of the `GenomeScanner` engine is greatly inspired by the Java programming language. First, there is a `Runnable` interface with a virtual `Run` method to be implemented by all implementing classes.

An abstract class `StringSearch` encapsulates the data and methods for each string search, including the block to be searched, the subsequence to be matched, the block size, beginning and end, and a string identifier for each of the source files, e.g., the host genome and viral genome file names. It also implements the `Runnable` interface, which simply invokes the `Search` method. The infrastructure is intentionally generic so the code can easily be extended to any alphabet or search space (e.g., amino acid sequences or written text). Each subclass of the `StringSearch` class implements a different string matching algorithm when it overrides the `Search` method. Each implementation is responsible for updating the vector containing the indices of all matches.

The `StringSearchFactory` class is responsible for generating instances of the appropriate `StringSearch` subclass based on the selected algorithm, `BOMStringSearch`

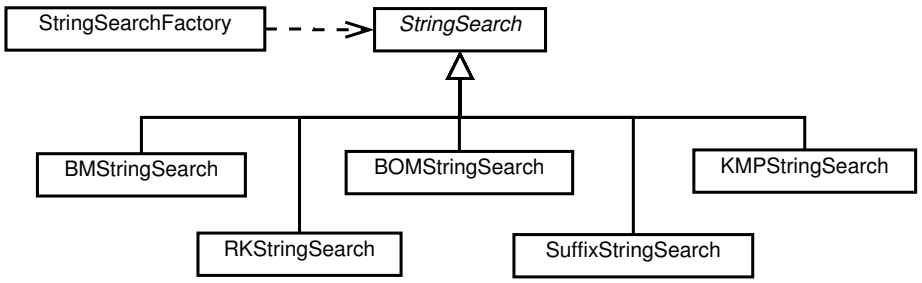

Figure 3.3: UML diagram of `StringSearch` class hierarchy.

for example (Figure 3.3).

Parallelism in the code is implemented (or not) via the thread interface class `ThreadInt`. Each thread consists of a named identifier, a `Runnable` target object and a boolean running status. There are three implementing subclasses of the thread interface, a `SimpleThread`, which simply executes the code sequentially and returns (sequential), a `PThread` implemented with Linux pthreads, and `MPIThread` implemented with the Message Passing Interface (MPI). The pthread and the MPI versions of the application are compiled separately as different dependencies are required (MPI programs are executed with `mpiexec`). The pthread version is intended for single machines with multiple cores, with each thread running on one processor core. In the MPI implementation, the zero rank (root) process acts as a delegator, passing data to the other worker nodes with nonzero rank to search each block for subsequence matches.

The `GenomeScanner` program generates an abundant number of threads. In order to avoid the overhead of frequent creation and destruction of threads, the thread pool design pattern has been implemented in the `ThreadPool` class to allow the reuse of thread objects (Figure 3.4).

As the `GenomeScanner` is reading viral subsequences from viral genomes and blocks

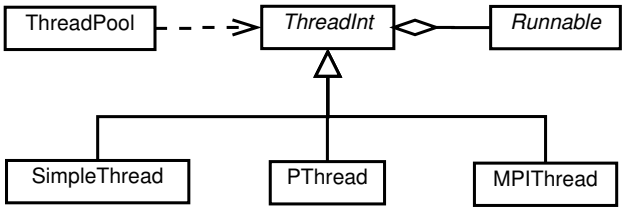

Figure 3.4: UML diagram of Threading hierarchy.

of genomic data from host genomes, it associates each subsequence i and block j with a thread from the thread pool and invokes a search with the appropriate search algorithm. However, it is not sufficient to search for the viral subsequence alone but also the complement. Sequences are saved in FASTA or Genbank files just as they would be read in the 5' to 3' direction. In other words, the way they would be read from mRNA molecules during translation. This allows researchers to search for start and stop codons, to find expressed sequences or pseudogenes within the genome. In the case of endogenous viral elements, it cannot be known whether the mRNA or the viral genome itself was inserted into the host genome. Thus, each viral subsequence actually results in two search threads, one for the 5' to 3' direction, and another for 3' to 5' (complementary). Figure 3.5 provides an example sequence.

To ensure that there are no dependencies between running search threads, the threads should not need to report their results back to the master node. To that end, each thread is responsible for writing its own data to disk. The `StringSearch` class is serializable to the JSON file format by implementing the `Jsonizable` interface class. When a thread finishes its search and if it has found any matches, it will generate

```
5'-GGTCTCTCTGGTTAGACC-3'
3'-CCAGAGAGACCAATCTGG-5'
```

Figure 3.5: Example of Complementary Viral Subsequences.

its own file name with the `JsonFile` method and write its contents to it with the `Jsonize` method. An example of the output is given in Figure 3.6.

The `GenomeScanner` also maintains a log file inspired by the `log4j` logger so that the process can be monitored while it is running. The logs and data are written to the file path specified by the `-fp` argument provided to the `GenomeScanner` when the executable is launched.

```
{
    "matches": [6407392],
    "block": 1,
    "begin": 1,
    "end": 100311,
    "dir": "5",
    "textID": "NC_004354_chrX",
    "pattID": "NC_001498",
    "pattern": "CCGAAGTTGGCCTTGTCG"
}
```

Figure 3.6: JSON output from a match within the X chromosome of *Drosophila melanogaster* and the Measles morbillivirus.

All of the match data for this project were collected by running the `GenomeScanner` across all five viral species and twelve host genomes on a Beowulf cluster with four nodes. Command lines for each host, virus pair (60 pairs) were generated and executed. The data were logged to the JSON flat file database.

A string search algorithm benchmark was performed on the Beowulf cluster with the parallel version of the `GenomeScanner`. The host input file was the human chromosome 22 and the viral input file the HIV-1 virus. Chromosome 22 consists of about 50 million base pairs and the FASTA file is divided into six blocks of 100K lines. The HIV-1 genome is a little more than nine thousand base pairs. There are ten SEVE matches from HIV-1 within chromosome 22. The five algorithms included

in the benchmark were brute force, Knuth-Morris-Pratt, Boyer-Moore, Rabin-Karp, and Backward Oracle Match. The brute force algorithm was implemented using the `find` method of the `string` class in the C++ standard template library.

Table 3.2: String Algorithm Benchmark Results

Algorithm	Abbreviation	Time (min)
Brute Force	BF	32.470
Rabin-Karp	RK	30.080
Knuth-Morris-Pratt	KMP	26.600
Boyer-Moore	BM	10.972
Backward Oracle Match	BOM	10.965

The results are summarized in Table 3.2. The Boyer-Moore and Backward Oracle Match algorithms have approximately the same running time and are both well ahead of the other string matching algorithms.

3.2.3 MatchDatabase

The third and final component of this software framework is the `MatchDatabase`. Not quite as structured as the other two packages, it is a collection of data processing scripts implemented in the Python programming language. This decision was motivated by the desire to take advantage of the excellent `Biopython` bioinformatics package developed by Cock et al. [50], and also the `scipy` and `pandas` scientific computing packages.

The first priority was to cross reference the output from the `GenomeScanner` with the NCBI database to determine if the SEVE matches are contained within any significant genes or noncoding regions within each host genome. This requires performing a BLAST [28] search for every subsequence match. The first attempt at this was to perform an online search using the `NCBIWWW.qblast` method from

the `Biopython` package. This proved prohibitively slow due to the large number of matches.

NCBI databases can be downloaded using the `update_blast_db` command that comes bundled with the BLAST+ toolkit [51]. Using this function, the `nt` (nucleotide), `nr` (non-redundant), and `refseq` databases were downloaded to the filesystem on the Beowulf cluster. The executables of the BLAST+ package were built locally on the system as well. Python code was developed to wrap the calls to the `blastn` command and capture the output. The results are returned in XML format, and additional code was developed to convert the XML data into JSON format and update the files in the `MatchDatabase`. The `mpi4py` package is an MPI implementation for the Python language [52]. The resulting `seveBlasterMPI` program was capable of running multiple BLAST searches in parallel for match, filter the data to those entries that pertained to the specific host and virus, and update the JSON files in the `MatchDatabase`.

BLAST search results for viral sequences contained the viral isolate from which the sequence was derived but lacked the actual viral genes. The `viralGenomeReader` program was developed to read the viral entries from the `MatchDatabase` and the corresponding Genbank file for each virus. Combining the annotations from the Genbank file with the match data allowed the actual location of the match within the viral genome to be determined, such as the nucleoprotein (NP) or glycoprotein (GP). Once the `MatchDatabase` entries were completed, the data could be analyzed.

The `matchDataWriter` program is the last major component of the `MatchDatabase` Python package. The role of this program is to read all of the SEVE matches, viral gene Genbank data, and BLAST search results from the database and write the results to one comprehensive data table. The only currently supported format is

comma separated values (CSV). This program also calculates the GC ratio of the sequence and can integrate the results with chromosomal band locations from the UCSC database.

CHAPTER 4

RESULTS

4.1 Overview

The final version of the `MatchDatabase` contains 47,480 total records. These consist of the verified exact matches of 18 base pair in length, derived from the five viral genomes across the twelve eukaryotic genomes. Since the step size parameter was set to 3 base pairs in the `step` parameter passed to the `GenomeScanner`, these data represent one third of all possible SEVE sequences from each virus.

The reduced sampling was deemed necessary, as even the smallest virus (HIV-1) contains over 9,000 base pairs requiring a total of 18,000 subsequence searches through every host genome to consider all possible sequences. Reducing the sample size enabled the collection of data across several viruses and numerous hosts. The total run time required to scan the entire human genome for every possible 18 bp SEVE match from the HIV-1 virus (step size of one) with the existing algorithm is 45 hours, so a step size of three reduces that time to 15 hours. The step size of one results in 2,745 unique SEVE sequences while the step size of three produces 1,450. This means that one third of the run time yields nearly one half of all sequences.

The data are first summarized by the number of matches from each virus within each species. In order to be properly compared, the data must first be normalized by the size of the host genome, as it is much more probable for a primate genome

with six billion nucleotides to share homology with a random viral sequence than the genome of a haploid yeast with only 120 million.

The haploid yeast *S. cerevisiae* genome contained no viral matches, and the *E. coli* bacteria only two, one from the *Ebolavirus* and another from SIV. Therefore, these two organisms will be omitted from the following summary figures and tables.

The number of SEVE matches in each host species are normalized by length (18 in this case). The host genome sizes are normalized by million of base pair (Mbp). In this way, the number of matches can be compared between genomes.

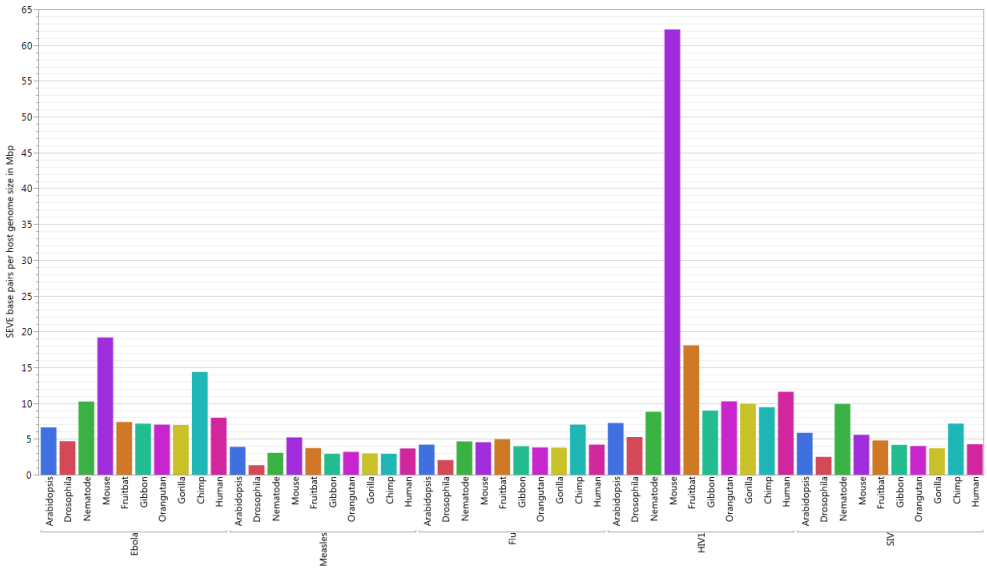

Figure 4.1: Ratio of SEVE sequences to host genome sizes by host and virus species.

The normalized results are summarized by the bar chart in Figure 4.1. The data set contains an obvious outlier. The mouse (*Mus musculus*) genome and the HIV-1 virus have a SEVE homology ratio of nearly 65, well above the next closest ratio of close to 20 between the mouse and HIV-1 virus. Further data mining revealed that the sequence `5'-AGAGAGAGACAGAGACAG-3'` alone accounts for 5,716 SEVE matches

between the mouse and HIV-1 genomes. BLAST search data confirm that this sequence is very prevalent in the mouse genome. The complementary sequence `3'-TCTCTCTCTGTCTCTGTC-5'` represents another 2,873 of the 9,672 total matches between the mouse and HIV-1 genomes. The two sequences together comprise 88% of the matches. Further analysis of the sequences will be included later, but for now they will be omitted so that the mouse and HIV-1 data are more comparable to the other sets.

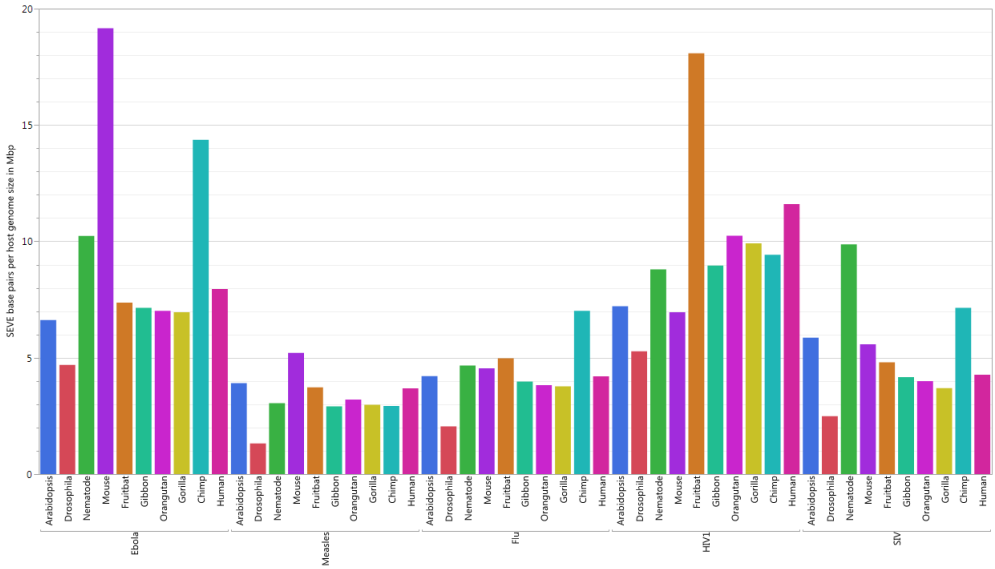

Figure 4.2: Ratio of SEVE sequences to host genome sizes by host and virus species with Mouse / HIV-1 outlier excluded.

The normalized data with the Mouse / HIV-1 outlier excluded are represented in Figure 4.2. After filtering the anomalous sequence, the data become more readily comparable. The noteworthy ratios are the Ebolavirus within the chimpanzee and mouse genomes, HIV-1 within the fruitbat, and SIV within the nematode genome. The SEVE matches will be further subdivided and analzed by the viral genes that contain them in the following sections.

4.2 Ebolavirus

4.2.1 Overview

The *Zaire ebolavirus* is a filovirus (*filo* meaning filamentous) named for a tributary of the Congo river known as the Ebola. The virus is 970 nm long and 80 nm in diameter. The genome encodes seven viral proteins in the order described in Table 4.1.

Table 4.1: Ebolavirus Gene Products

Name	Product	Size (bp)
NP	Nucleoprotein	2220
VP35	Polymerase cofactor	1023
VP40	Matrix protein	981
GP	Glycoprotein precursor	1095
VP30	Transcription cofactor	867
VP24	Membrane protein	756
L	RNA polymerase	6639

The lifecycle of the *Ebolavirus*, like many viruses, consists of seven stages. These are attachment to the host cell membrane, gaining entry to the cell cytoplasm, transcription of the viral genome into messenger RNAs, translation of mRNAs into viral proteins, replication of the viral genome, the assembly of genomes into proteins into new virions, and finally the exit of mature virions from the cell.

The coding regions are flanked on either side by the 3'-OH leader and the 5' trailer respectively, as it is a minus strand (-ssRNA) virus. The virus is characterized by having two or three gene overlaps of VP35/VP40, GP/VP30, or VP24/L [53].

The glycoprotein (GP) facilitates attachment to the host receptors DC-SIGN and DC-SIGNR [54]. Phosphatidyl serine on the viral membrane surface then binds to the HAVCR1 cell receptor, inducing the cell to initiate apoptic mimicry by signal transduction, and permitting the virion to enter the cell via macropinocytosis.

The RNA polymerase from the L gene binds to the 3'-OH leader of the viral RNA transcribes it into an mRNA complete with 5' cap and 3' poly-A tail. The glycoprotein is cleaved by the furin enzyme in the host cell into GP1 and GP2 proteins. The furin enzyme, also known as the paired basic amino acid cleaving enzyme (PACE), is a calcium-dependent serine endoprotease expressed in the cells of many tissue types, including neuroendocrine, liver, gut, and brain. The human FURIN gene is found on chromosome 15 [55].

GP1 promotes fusion of the viral membrane with the vesicle membrane by interacting with host NPC1, allowing the ribonucleocapsid to enter the cytoplasm. NPC1 (Niemann-Pick disease, type C1) is a transmembrane protein responsible for mediating cholesterol transfers in mammalian cells. The human version is located on chromosome 18.

Viral genome replication can begin once sufficient nucleoproteins have been translated to encapsulate the newly produced genomes. Viral genes are typically organized by the quantity of a protein product required. In Ebola, for example, many more copies of the NP protein are required than of the L polymerase. Therefore, the NP gene is the first after the 3' leader and the L gene is the last before the 5' trailer.

The VP35 protein is a polymerase cofactor, involved in host immune system evasion, specifically by inhibiting the RIG-I-like receptors of the host cells [56]. The encapsulated virions then interact with the matrix protein (VP40), and exit the cell by budding with the aid of host ESCRT protein complexes [57].

The VP30 zinc-binding protein is necessary for activation of viral transcription and is associated closely with the nucleocapsid complex [58]. VP24 promotes viral survival by suppressing the the production of alpha/beta interferon (IFN-α/β) [59]. The role of the 5' trailer sequence of the Ebola genome is not completely understood. However, transcripts with the 5' trailer deleted have been shown to be deficient in replication, indicating that the trailer is important for viral genome replication [60].

4.2.2 Ebola SEVEs by Species

Grouping the number of ebolavirus matches by host chromosome and normalized by the host chromosome size (Mbp) as a measure of relative homology (Figure 4.3) reveals that the chimpanzee and even more so the mouse contain a relatively high degree of homology compared to the other species.

According to a study of ebolavirus infections in 47 different mouse lineages by Rasmussen et al., the mice displayed a range of symptoms from full hemorrhagic fever to none at all [61]. Those that developed the lethal fever exhibited low levels of activity for the Tie1 and Tek genes that increased the permeability of their membranes and resulted in significant inflammatory responses. Importantly for this study, the mouse adapted version of the ebolavirus (MA-EBOV) does not lead to the fatal syndrome in lab mice. That includes GRCm38.3 genome from which these data are produced. Similarly, not all humans are susceptible to the ebolavirus hemorrhagic fever.

A study by Bermejo et al. indicates that both chimpanzees (*Pan troglodytes*) and western gorillas (*Gorilla gorilla*) are susceptible to infection by the ebolavirus and indeed the disease has resulted in sharp declines in their numbers in Gabon and the

Congo [62]. The mortality rates could be as high as 95% for gorillas and 77% for chimpanzees.[1]

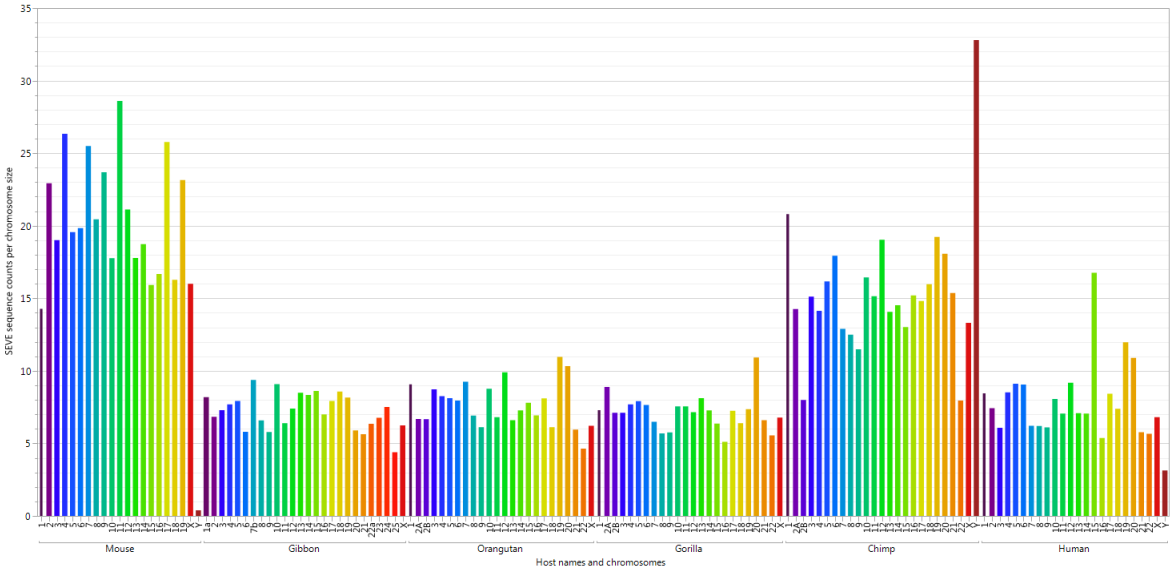

Figure 4.3: *Zaire ebolavirus* SEVE match count by host name and chromosome.

The most frequently occurring sequences in the chimpanzee are on chromosome 1 and the Y chromosome. The top two sequences in chromosome 1 are `5'-AAAAATTTAAA-AATAAAT-3'` and `3'-TTTTTAAATTTTTATTTA-5'`, which occur 206 times and 466 times, respectively. The two sequences are also complementary. The 5' to 3' sequence is located within the 5' header of the Ebola genome. BLAST searches returned no significant results for the 5' to 3' sequence, however the 3' to 5' sequence was found in the CAMK2N1 gene for the calcium/calmodulin-dependent protein kinase II inhibitor 1 protein on chromosome 1. The sequence can be found on every chromosome in the *Pan troglodytes* genome.

[1] http://www.projetogap.org.br/en/noticia/ebola-killed-third-gorillas-chimpanzees-world

The same sequences are also the most frequent Ebola matches in the human genome, with 92 matches for the 5' and 252 matches for the 3' to 5' sequence. `3'-TTTTTAAATTTTTATTTA-5'` can be found in the CECR2 gene in humans, a cat eye syndrome candidate. The gene encodes a protein containing a bromodomain involved in chromatin remodeling that may also play a role in DNA damage response [63].

Another often occurring sequence is `5'-TATTGAGCAGTATTGAAA-3'` with 57 matches on human chromosome 15. The sequence comes from the VP24 gene for the Ebola membrane protein. BLAST searches indicate that the sequence occurs only in non-coding regions.

Other notable Ebola SEVE matches in the human genome include `5'-ATTATTTAA-AATTCTTTC-3'` in the TRIM37 motif, `5'-AAAACAAAACTGATCTTT-3'` in the GRIN2B glutamate receptor on chromosome 12, and `3'-TTACAAAGATGGCCTTAG-5'` in the PKC-potentiated PP1 inhibitory protein (*PPP1R14A* gene) on chromosome 19.

The most frequent match for the chimpanzee Y chromosome is `5'-TCAACCACCACCT-GGACC-3'` with 14 matches. The sequence is located within the VP35 polymerase cofactor gene in the Ebola genome. The sequence is also found on the human Y chromosome, within the inverted repeat IR2 of the Y palindromes P1, P2, and P3.

The now familiar `5'-AAAAATTTAAAAATAAAT-3'` and `3'-TTTTTAAATTTTTATTTA-5'` sequences also make numerous appearances within the mouse genome, occurring 40 and 157 times each. The 3' sequence is found in the *RRNAD1* gene for ribosomal RNA adenine dimethylase located on chromosome 1.

The SEVE sequence `5'-TGAGTTCCAGGCCAGCCT-3'` from the Ebola VP35 protein occurs an incredible 1,807 times in the mouse genome. BLAST searches find it to be located within the *LOC105246594* noncoding RNA region. Another commonly matched sequence with 100 total matches is `3'-CTGCTCCTGCTGCTCCTG-5'` from the

Ebola nucleoprotein gene, also located within the *CXCL14* gene that encodes the chemokine (C-X-C motif) ligand 14. This protein is involved in immunuregulatory and inflammatory responses, and is related to the Akt signalling pathway [64]. The human version is located on chromsome 5.

An additional sequence from the nucleoprotein is `5'-GAAAAAGAGGCCATGAAT-3'`, found 68 times in the mouse genome. BLAST results located it within the MAP3K13 gene that codes for a member of the serine/threonine protein kinase family that can activate MAPK8 or MAP2K7 MAP kinase cascades, which indicates a likely role in the JNK signalling pathway [65].

4.2.3 Ebola SEVEs by Viral Gene

Having considered the SEVE matches from the host perspective by chromosome, it is also useful to group by viral gene. The number of matches in *D. melanogaster*, *A. thaliana*, and *C. elegans* are negligible. The majority of the Ebolavirus SEVE matches occur within the 5' trailer at the end of the genome. In the *M. musculus* genome, the sequence `5'-TGAGTTCCAGGCCAGCCT-3'` discussed in the preceding section accounts for the large number of matches from the VP35 polymerase cofactor protein. The role of the 5' genome header in viral infectivity is not well understood.

4.3 Human Immunodeficiency Virus 1

4.3.1 Overview

The *Human immunodeficiency virus* is a retrovirus with a +ssRNA genome complete with 5' cap and 3' poly-A tail. The virus is conical to spherical in shape, 80-100 nm in diameter, and contains over 1500 capsid proteins in the mature form. The viral

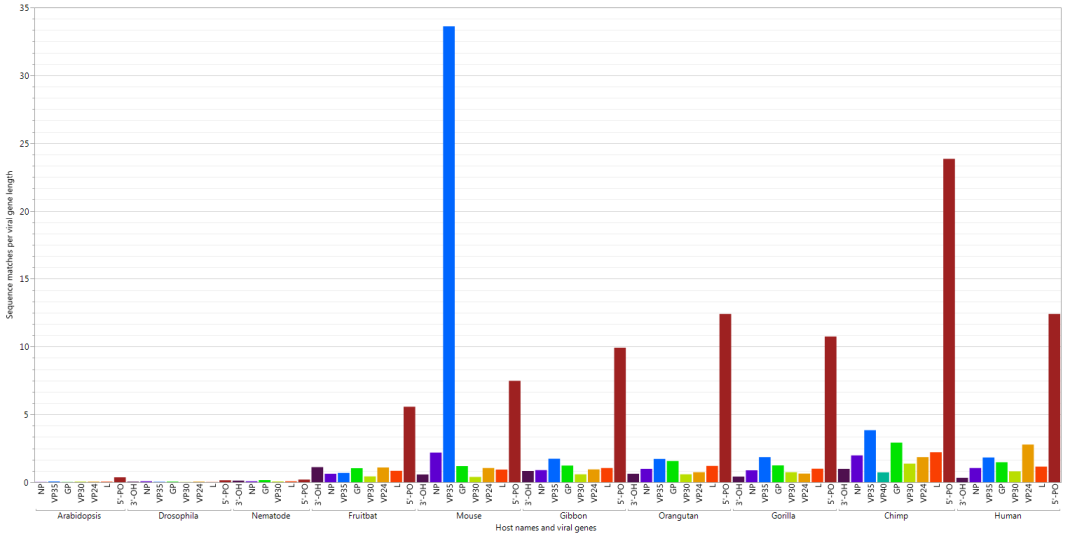

Figure 4.4: *Zaire ebolavirus* SEVE match count per viral gene and normalized by gene size.

proteins encoded by the genome are summarized in Table 4.2. The protein coding section is flanked on either side by long terminal repeat (LTR) regions 600 bp in length. The 5' end also contains a primer binding site (PBS) and the 3' end includes a polypurine tract (PPT).

Table 4.2: HIV-1 Gene Products

Name	Product	Size (bp)
gag-pol	Protease, RT, RNaseH, integrase	4308
gag	Group-specific antigen	1503
vif	Viral infectivity factor	579
vpr	Viral protein R	291
tat	Transcriptional activator	2595
rev	Regulator of expression	2685
vpu	Viral protein U	249
env	Envelope surface glycoprotein	2571
asp	Antisense protein (unknown)	570
nef	Negative factor	621

Lentiviruses such as HIV-1 are characterized by long term illnesses with extensive

incubation periods. The infection process begins with attachment by the gp120 glycoprotein to the host cell surface receptors DC-SIGN [19], Heparan Sulfate Proteoglycan [66], and the CD4 receptors of the helper T cells [19]. Host cell entry is mediated via clathrin-dependent endocytosis with the transmembrane glycoprotein gp4 facilitating dynamin-dependent fusion with the endosome. The envelope spike encoded by the *env* gene consists of three copies of gp120 and gp41 to form a trimer of heterodimers.

Once the nucleocapsid enters the cytoplasm, the +ssRNA viral genome is transcribed into linear dsDNA by the viral reverse transcriptase (RT) enzyme. The dsDNA must be transported to the host nucleus along with the viral integrase encoded by the *pol* gene. The integrase enzyme randomly integrates the viral DNA into the nuclear host DNA to form a provirus, accomplished by hijacking the DNA repair mechanisms of the host cell [67]. The provirus may become latent, awaiting later activation, or be transcribed immediately into new viral genomes.

The 5'-LTR of the provirus contains promoter elements that are bound by the RNA polymerase II enzyme of the host to begin transcription. Some of the transcripts will be unspliced and others will be spliced by post-transcriptional modification in the spliceosome. The unspliced transcripts will either become future RNA genomes or be translated after the transcripts are exported from the nucleus. The spliced transcripts will be immediately translated to produce Tat, Rev, and Nef proteins.

The Rev protein contains both a nuclear localization sequence (NLS) to remain in the nucleus. The unspliced transcripts bind to the rev response element (RRE), located immediately downstream of the *env* gene. RRE binding results in a conformational change that exposes the nuclear export sequence (NES), allowing the unspliced transcripts to be shuttled to the cytoplasm by exportin proteins. Upon releasing the cargo, the NLS sequence is exposed and importins bring the Rev protein back into

the nucleus for another cycle.

The unspliced transcripts are then translated into Env, Gag, and Gag-pol polyproteins. Cleavage of the Env proteins by the viral protease yields the envelope proteins TM and SU, as well as the accessory proteins Vif, Vpu, and Vpr. New virions are assembled and the genomes packaged at the host plasma membrane. The virions are released via exocytosis by budding. The precusor polyproteins translated from the unspliced transcripts are cleaved by the viral protease to form mature virions.

The Tat protein (Trans-Activator of Transcription) is a kinase that greatly increases transcription rate of viral dsDNA by phosphorylating cell factors [68]. Tat can also be absorbed by nearby uninfected T cells, inducing apoptosis, and accelerating the demise of the host immune system [69].

The Nef protein is a negative regulatory factor that helps active T cells to increase the likelihood of infection. It acts as an enzyme to lower the activation energy of CD4+ lymphocytes. The T cell receptor response (TCR) renders the cells susceptible to infection by other virions [70]. Nef expression is not strictly required for HIV infection to occur.

The Vif protein, or viral infectivity factor, inhibits the antiviral activity of the APOBEC3G protein by marking it for degradation via ubiquitination. APOBEC3G is a cytidine deaminase that mutates viral mRNAs by deaminating the cytosine nucleotides into uracil. Vif is necessary for viral replication because otherwise the deaminase will enter the budding virions and scramble their genomes before they reach the next target cell [71].

The Vpr protein (Viral Protein R) is involved in the regulation of nuclear import of the pre-integration complex, including the reverse transcribed dsDNA and the integrase enzyme. Vpr is required for viral replication within post-mitotic macrophages

and can also suspend dividing in the G2 phase leading to apoptosis [72].

The Vpu protein (Viral Protein Unique) induces the degradation of the CD4 viral receptor in the endoplasmic reticulum, resulting in a downregulation of CD4 expression. This results in the prevention of unintentional CD4-Env binding in the ER to faciliate the proper formation of virions within the cell. The Vpu protein itself is not packaged into new virions [73].

The role of the *asp* (antisense protein) remains unclear, though recent evidence suggests it may be involved in the process of autophagy [74].

4.3.2 HIV-1 SEVEs by Species

In Figure 4.5, the data indicate that the least significant degree of homology exists between the *Mus musculus* genome and the HIV-1 virus. According to Zheng et al., lab mice do not exhibit symptoms when infected with HIV due to a post-transcriptional block [75]. The mouse version of the protein (*mp32*) is a nuclease that actively cleaves HIV mRNA transcripts. The human version (*p32*) does not cleave the transcripts. When the human *p32* protein is introduced to their genomes, the mice become susceptible to infection.

The primate species have similar levels of relative homology, with exceptions in the gibbon chromosome 24, orangutan chromosome 19, chimpanzee Y chromosome, and human chromsomes 20 and 21. A pair of complementary sequences from the *rev* gene are the cause of these high values and will be discussed in the next section.

4.3.3 HIV-1 SEVEs by Viral Gene

Like the *Ebolavirus*, the number of matches in *D. melanogaster*, *A. thaliana*, and *C. elegans* are not significant. The majority of the HIV-1 SEVE matches occur within

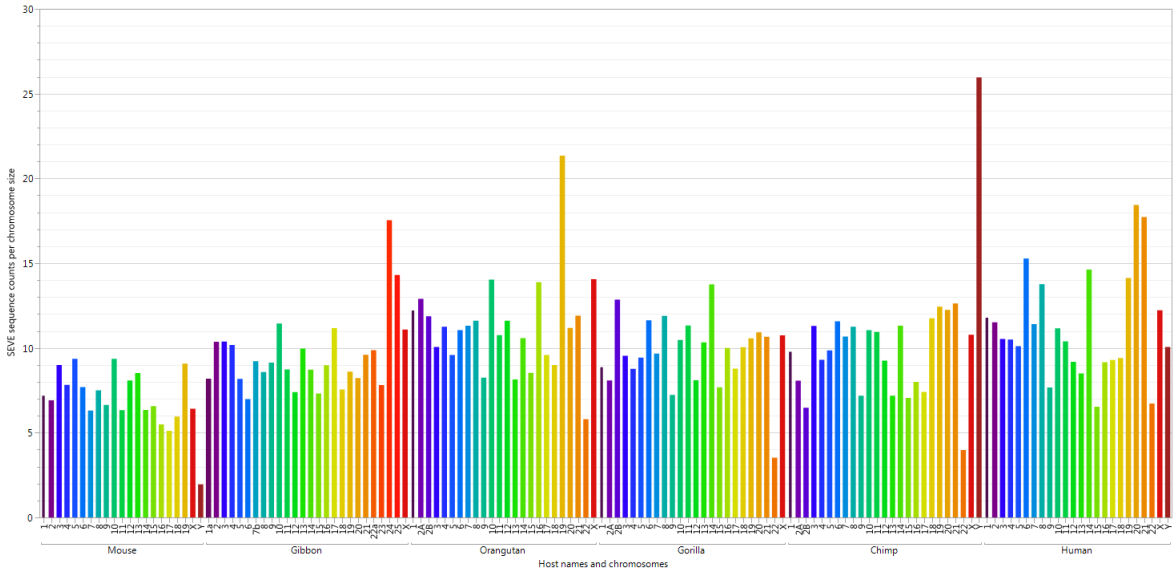

Figure 4.5: *Human immunodeficiency virus 1* SEVE match count by host name and chromosome.

the *rev* gene, with an exceptionally high ratio in the *P. alecto* genome. The large number of matches from the HIV-1 *rev* gene in the `MatchDatabase` are due to two complementary sequences, `5'-AGAGAGAGACAGAGACAG-3'` having 8,715 matches and `3'-TCTCTCTCTGTCTCTGTC-5'` with 4,877 matches. These sequences alone account for 28.6% of the SEVE matches in the entire database and can be found in eight of the twelve host genomes. The 5' to 3' sequence is found in the *C7orf34* human gene, or chromosome 7 open reading frame (ORF) 34. The 3' to 5' sequence is found in the human ribosomal RNA gene (rRNA) intergenic spacer, downstream of the 47S coding region based on BLAST results. It is also found in the *Rbfox1* mouse gene. The human version *RBFOX1* is located on chromosome 16, from the Fox-1 family of RNA binding proteins. The *RBFOX1* gene is highly conserved across evolution and is believed to play a role in neuronal development [76]. The sequence also occurs in the chimpanzee gene *SLC9A3*, a solute carrier cation proton antiporter channel. The

human version is located on chromosome 5.

Since the fruitbat genome has not been annotated, the chromosome data is unavailable. However, these same two sequences, `5'-AGAGAGAGACAGAGACAG-3'` and `3'-TCTCTCTCTGTCTCTGTC-5'` occur 739 and 515 times respectively. Clearly these are the most significant sequences from the HIV-1 virus in host genomes.

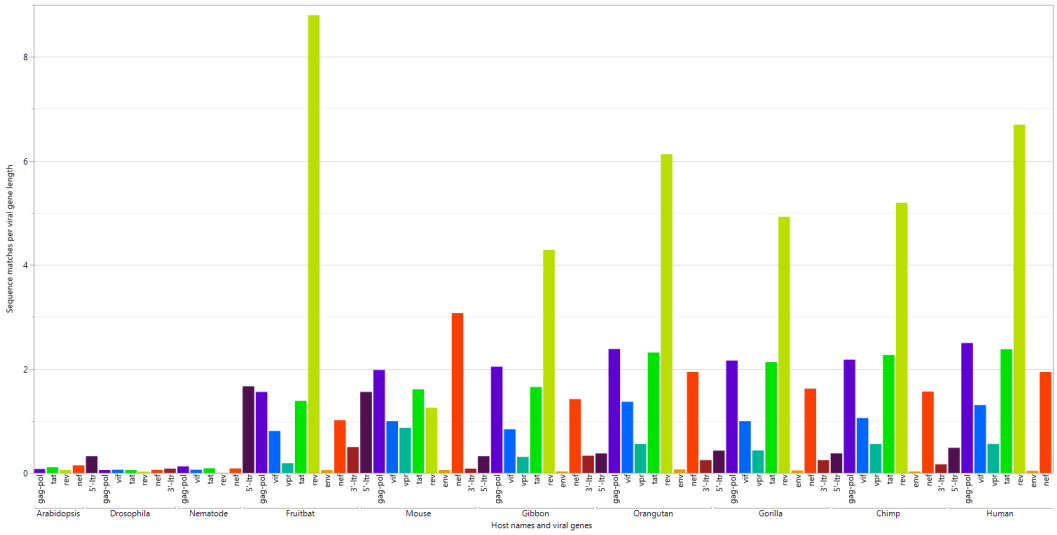

Figure 4.6: *Human immunodeficency virus 1* SEVE match count per viral gene and normalized by gene size.

4.4 Simian Immunodeficiency Virus

4.4.1 Overview

The *Simian immunodeficiency virus* is similar to the human version as it is also a lentivirus. SIV is also spherical in shape with a diameter of 80-100 nm. The genome is somewhat simpler and encodes fewer viral proteins (Table 4.3). SIV is known to infect up to 45 primate species in an evolutionary history spanning at least 32,000 years.

Multiple lineages of SIV exist, including SIVcpz (chimpanzee), from which HIV-1 evolved, SIVsm (sooty mangabey) from which HIV-2 evolved, SIVagm (African green monkey), and several others. The data presented here are from the SIVcpz genome, which is only known to naturally infect *Pan troglodytes* [77]. Unlike HIV in humans, SIV in primates is not always pathogenic, but can lead to the fatal Simian AIDS (SAIDS).

Table 4.3: SIV Gene Products

Name	Product	Size (bp)
gag	SIV2 Glycoprotein 1	1554
pol	Protease, RT, RNaseH, integrase	3033
vif	Viral infectivity factor	639
vpx	Viral protein X	300
vpr	Viral protein R	306
tat	Transcriptional activator	2574
env	Envelope surface glycoprotein	2601
nef	Negative factor	558

The SIV life cycle is similar to the HIV-1, as the two viruses are closely related. Like HIV, attachment is mediated via the gp160 glycoproteins in the viral envelope binding to the CD4 molecules in the T cell membranes [78]. The glycoproteins are encoded by the *gag* gene.

The *pol* gene encodes four proteins: the protease (*prot*), the reverse transcriptase (*p51*), the RNase (*p15*), and the integrase (*p31*). The *gag*, *pol*, *vif*, *vpr*, *tat*, *env*, and *nef* proteins perform similar roles in SIV as in HIV-1 (described in the previous section). One exception is the SIV *vpx* protein found in HIV-2, but not HIV-1. The *vpx* protein is similar to *vpr* in that it exploits cellular machinery by ubiquitylating specific cellular proteins and marking them for destruction [79].

4.4.2 SIV SEVEs by Species

The chimpanzee genome contains the greatest number of SEVE matches of any of the primates in the study. This could be significant as only chimpanzees are susceptible to SIV (the SIVcpz strain specifically). Indeed it is suspected by Worobey et al. that HIV-1 is a derivative of SIVcpz that crossed the species barrier due to bushmen hunting chipanizees for food [80]. The SEVE matches in the chimpanzee genome are well distributed with each only occurring only a few times. The exceptions to this are `3'-TGGTGTTTGGTTTTTCGT-5'`, `5'-AAAGAAAGGAAAATAGAA-3'`, and `3'-GTGTTAAAATTTTCTTTT-5'`, each occurring 26, 22, and 16 times, respectively. These results indicate that a high degree of viral homology does not confer immunity to the host.

Mice are unlikely to be infected by SIV if they are not naturally infected by HIV [81]. Figure 4.7 contains the SEVE matches by host name and chromosome. The unexpectedly high degree of homology in the mouse Y chromosome is due to two sequences, `3'-GTTTATTGTGTATAAGAA-5'` from the *tat* gene with 87 matches, and `'5-GATGGTGAATTTTTTAGG-3'` from the *gag* gene with 52 matches. Neither sequence occurs within a protein coding region. There are also a significant number of matches (52) within the mouse chromosome 12. Most of the matches are within noncoding regions except for `3'-TCTTTGAGGTTTCTTCCC-5'`. That sequence is contained within the immunoglobulin heavy chain region of the genome of *Mus musculus* strain 129S1 [102].

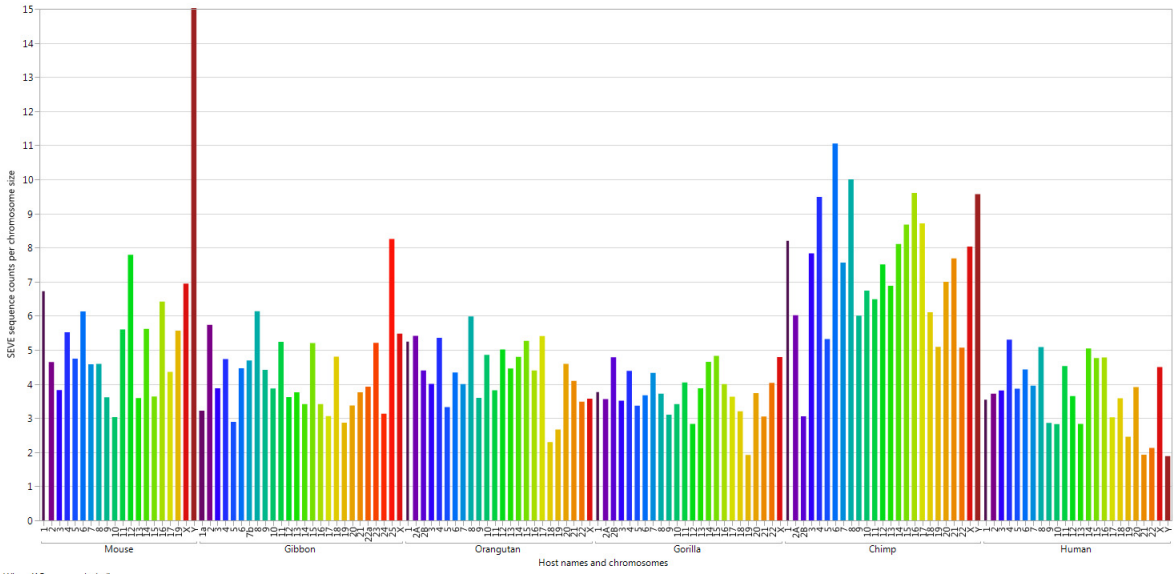

Figure 4.7: *Simian immunodeficiency virus* SEVE match count by host name and chromosome.

4.4.3 SIV SEVEs by Viral Gene

Most of the SEVE matches from the SIV virus occur within the chimpanzee genome, particularly from the *pol*, *tat*, and *vif* viral genes (Figure 4.8). Of the matches within the *pol* gene, there are three protein coding genes. The SEVE sequence `5'-AAAGAAGGGAAAGCAGGA-3'` is contained within the *KLHL33* gene for kelch-like family member 33 from chromosome 14. SEVE `5'-TTGTGGTATAACCTGTTG-3'` resides in the *GTF2A1* gene also on chromosome 14 that encodes the general transcription factor TFIIA. Lastly, the sequence `5'-AGAGACCAAGCAGAGAAA-3'` is found in the *THSD7A* gene on chromosome 7, encoding the thrombospondin type I glycoprotein necessary to create blood platelets (thrombocytes).

From the SEVEs in the HIV-1 *tat* gene, the sequence `5'-CAAGACTATCCATGTGGG-3'` is contained within the *ZP1* gene on chromosome 11 that encodes the zona pellucida

glycoprotein 1, a sperm cell surface receptor. Another is `3'-ATTGACTGTTATACTGTC-5'`, contained within the *ADAM18* gene on chromosome 8, encoding the disintegrin and metalloproteinase domain 18. Only one sequence from the SIV *vif* SEVEs is significant, `3'-TATATTCAAGTGTTTGAT-5'`, contained within the *HOOK3* chimpanzee gene on chromosome 8. *HOOK3* codes for the hook microtubule-tethering protein 3.

There are also a few significant matches from the *tat* gene within the mouse genome. One is SEVE seqence `3'-GTCCAGGGTACCTTCTTT-5'` from chromosome 16 for the *LOC105246101* long noncoding RNA. Another is `3'-TGATACTTCTCGTTGGTT-5'` located within the *Gnpda2* gene on chromosome 5 for the glucosamine-6-phosphate deaminase 2 protein. Finally, there is sequence `3'-TACATTGTCTTTACAAGA-5'` from the *Gpr126* gene for G protein coupled receptor (GPCR) 126 on chromosome 10.

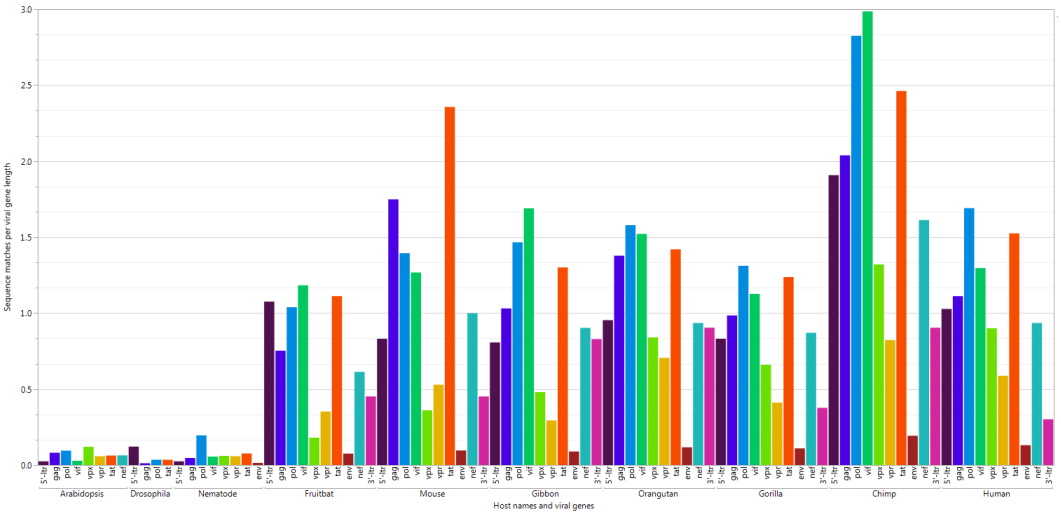

Figure 4.8: *Simian immunodeficency virus* SEVE match count per viral gene and normalized by gene size.

4.5 Measles Morbillivirus

4.5.1 Overview

The *Morbillivirus*, commonly known as the measles, has a spherical capsid of 150-300 nm with a -ssRNA genome that is 15-16 kb in size and encodes eight proteins as described in Table 4.4.

Table 4.4: Measles Morbillivirus Gene Products

Name	Product	Size (bp)
N	Nucleocapsid	1581
P/V/C	Phosphoprotein	1524
M	Matrix protein	1008
F	Fusion protein	1653
H	Hemagglutinin	1854
L	RNA Polymerase	6552

The *Morbillivirus* life cycle is similar to other group V viruses like the *Ebolavirus*. Attachment occurs through the hemaglutinin (H) protein on the viral surface to the cell surface receptors. Three such receptors for the measles H protein have been identified in humans. One is the CD46 complement regulatory protein (cluster of differentiation 46), an inhibitory receptor encoded by the *CD46* gene. Another is the signalling lymphocyte activation molecule (SLAM) encoded by the *SLAMF1* gene. The third is the Nectin-4 cellular adhesion molecule encoded by the *PVRL4* gene [82]. All three genes are located on chromosome 1.

Following the binding of the receptor by the H protein, the F protein trimer conformation changes, allowing fusion with the plasma membrane to occur [83]. The ribonucleocapsid is then released into the cytoplasm via endocytosis.

The viral RNA polymerase (L) binds to the viral genome at the 3'-OH leader and sequential transcription begins. The polymerase adds the 5' cap and 3' polyadenyla-

tion to form mature mRNA transcripts. The gene that encodes the phosphoprotein P also contains two overlapping genes for the V and C proteins. The mRNA for the V protein is an edited version of the P mRNA and the C protein is a result of leaky scanning. The process of leaky scanning involves a weak start codon (e.g., `ACG`) and a small upstream open reading frame (uORF), allowing the ribosome to occasionally skip the weak codon and translate multiple proteins from one mRNA [84].

Replication begins when sufficient nucleoproteins have been translated. The nucleocapdsid (N) binds to the matrix protein (M) near the plasma membrane. The P protein is a polymerase cofactor that binds the N proteins and helps position them for assembly. The V and C proteins are viral infectivity factors that are not strictly required for propagation [85]. Exocytosis is facilitated by host ESCRT proteins (endosomal sorting complex for transport), and the virion is released through budding [86].

4.5.2 Measles SEVEs by Species

There are no known animal hosts for the measles virus (MeV), though it is believed to have evolved from the rinderpest virus of cattle [87]. Other viruses belonging to the *Morbillivirus*, such as distemper, can infect dogs, cats, and cetaceans (Figure 4.9). Recent research indicates that although the instances of infection are rare in the wild, viruses such as measles and influenza from humans are capable of crossing species and infecting apes and monkeys, including chimpanzees [103].

The species displaying the greatest level of homology with the measles virus is *Mus musculus*. Mice are not naturally susceptible to infection by the measles virus, as their cells lack the CD46 membrane receptor protein that human cells have. Transgenic mice modified such that their dendritic cells express the CD46 cofactor and their

CD150 interferon (IFN) pathways disrupted have been engineered to study measles infections in mouse models [104].

There are four SEVE sequences that account for most of the matches between measles and mice. The `3'-GGGGTGATTGGGAGGAGT-5'` sequence from the matrix protein (M) gene occurs 68 times, frequently in chromosomes 1, 3, 8, 10, and X. Sequence `5'-CAGCAACTGCATGGTGGC-3'` from the hemagglutinin protein (H) accounts for 73 matches, mostly within chromosomes 1, 2, and 7. SEVEs `5'-AAGAAAAGGAGATCAAGG-3'` and `5-TAGCAACAGTGTACTCAT-3` from the polymerase (L) contribute 63 and 35 matches, respectively, the former from chromosomes 1, 6, X and the latter from 7, 10, and 17. None of these appeared within protein coding regions.

The Y chromosome of the chimpanzee appears as an outlier in Figure 4.9 because it contains 14 SEVE matches in a very small chromosome, three of which occur three times each.

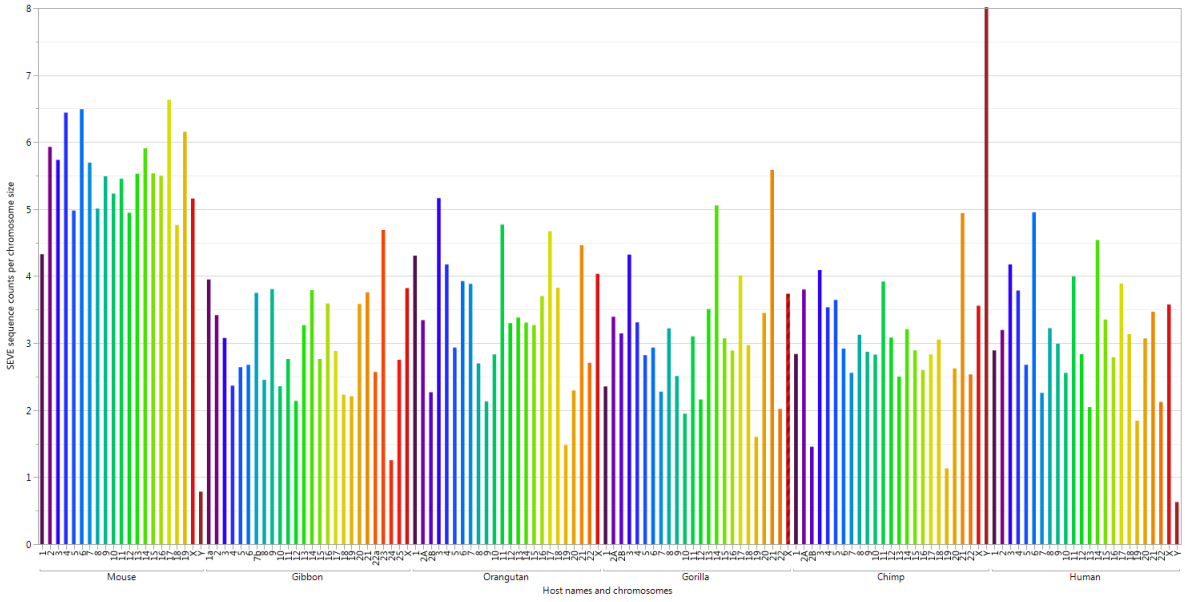

Figure 4.9: *Measles morbillivirus* SEVE match count by host name and chromosome.

4.5.3 Measles SEVEs by Viral Gene

Most of the Measles SEVE matches occur within the 3'-OH Leader of the viral genome by a significant margin (Figure 4.10). That could be significant since the 3'-OH leader is where the L polymerase binds at the beginning of transcription. Only one of the 3'-OH matches occurs a significant number of times, with `3'-TTGTCCCAGCCCCTCTTC-5'` having twelve appearances.

Two of the sequences do appear in protein coding genes in *Homo sapiens* and other primates. The `3'-TTGGATCCTAACGACTTT-5'` sequence occurs within the *NUCB2* gene on chromosome 11, encoding the nucleobindin 2 regulator for glucose transporter 4 (*GLUT4*) [105]. Another interesting sequence is `3'-TTGTCCCAGCCCCTCTTC-5'`, residing within the *ARRB2* gene on chromosome 17, responsible for coding arrestin β 2 protein believed to play a role in the agonist-mediated desensitization of GPCRs [106].

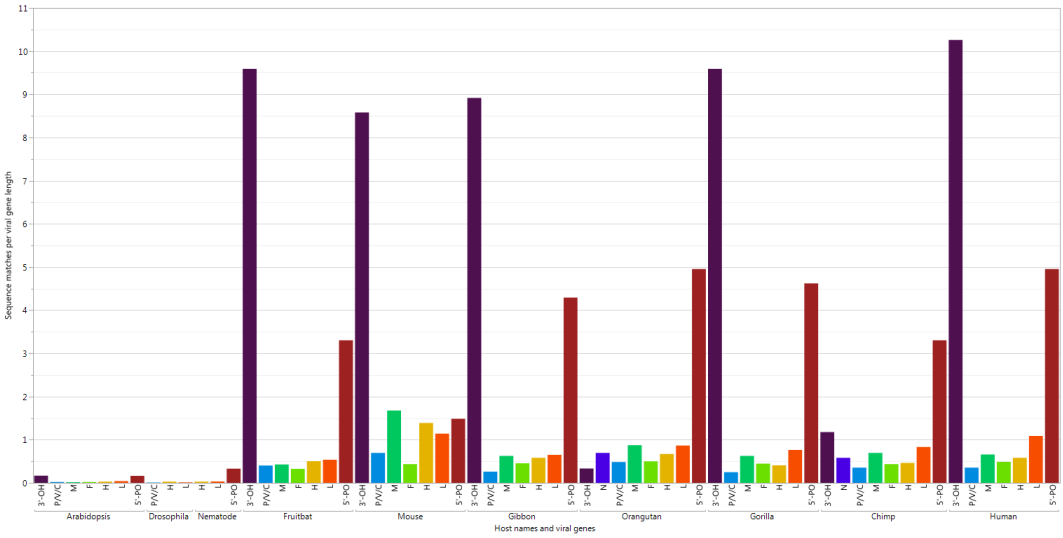

Figure 4.10: *Measles morbillivirus* SEVE match count per viral gene and normalized by gene size.

4.6 Influenzavirus A

4.6.1 Overview

The *Influenzavirus* is an enveloped virus with a spherical or filamentous capsid of 80-120 nm in diameter with a -ssRNA genome that is 13-14 kb in size and encodes twelve proteins in eight segments, as described in Table 4.5. Unlike the other group V viruses, the influenza genome is segmented, with segments ranging in length from 890 to 2,340 nucleotides. Influenza viruses are classified by the hemagglutinin and neuraminidase receptor proteins expressed in the envelopes, H7N9 in this case. The viruses are also categorized by the natural host, such as avian, bovine, or swine (e.g., H1N1). According to the Centers for Disease Control (CDC), researchers have identified 11 neuraminidasae subtypes and 18, for a total of 198 possible influenza combinations.

Table 4.5: Influenzavirus A H7N9 Gene Products

Name	Product	Segment	Size (bp)
PB2	PB2 Polymerase	1	2280
PB1	PB1 Polymerase	2	2274
PB1-F2	Apoptotic factor	2	273
PA	PA Polymerase	3	2151
PA-X	PA-X protein	3	760
HA	Hemagglutinin	4	1683
NP	Nucleocapsid protein	5	1497
NA	Neuraminidase	6	1398
M2	Matrix protein 2	7	982
M1	Matrix protein 1	7	760
NEP	Nuclear export protein	8	838
NS1	Nonstructural protein 1	8	654

Attachment occurs between the sialic acid receptor of the host cell membrane with the hemagglutinin (HA) protein in the viral envelope. Sialic acid is a derivative of

neuraminic acid that is prevalent in animal tissues, particularly in the human brain, where it is involved in synaptogenesis [88]. HA proteins also cause red blood cells to agglutinate. The virion enters the cell via clathrin mediated endocytosis and the endosome releases the RNA segments into the cytoplasm.

The RNA segments are encapsidated to form ribonucleoproteins (RNPs). The viral proteins contain nuclear localization signals (NLS) allowing them to bind to importins that carry them through the nuclear pore complexes (NPCs) as part of the Ran-GTP pathway [89]. Once inside the nucleus, the RNPs disassemble and the RNA segments are transcribed by the viral RNA polymerases (PB1, PB2, and PA) to create one mRNA per segment. Rather than generating their own 5' methyl-G caps, the viral polymerase cleave them from cellular mRNAs in a process called cap snatching [90]. The mRNAs are polyadenlyated by polymerase stuttering [91].

The PA polymerase and PA-X protein are both translated from the third segment due to a ribosomal frameshift. The PA-X protein is an endonuclease that acts by inhibiting host immune system response and encouraging viral growth [92]. The PB1 polymerase and PB1-F2 protein are translated from segment two by leaky scanning. The PB1-F2 protein invades the mitochondrial inner membrane via Tom40 channels, repressing cellular innate immunity and ultimately leading to apoptosis [93]. Matrix proteins M1 and M2 are translated from the segment 7 mRNA as a result of alternative splicing. M1 forms the matrix boundary between the viral genome and the envelope. M2 is a proton selective ion channel that is activated by low pH and is necessary for viral replication [94]. Segment 5 encodes the nucleoprotein. The NS1 and NEP proteins are translated from segment 8, also products of alternative splicing. NS1 is a nonessential accessory protein with suspected roles in preventing cellular mRNA polyadenylation and inhibiting interferon production [95].

Once sufficient levels of NP and M1 proteins have been produced, the nuclear export protein (NEP) triggers binding to exportins that transport the virion components back to the cytoplasm, where they migrate to the plasma membrane for assembly and exocytosis via budding.

4.6.2 Influenza A SEVEs by Species

The H7N9 influenza A variant included in this study is a form of avian flu that normally only circulates in bird populations, especially agriculturally grown animals, such as chickens and turkeys. Only rarely does an avian virus infect other species as the H7N9 did in China during the year 2013, resulting in a mortality rate of 30% [107]. The strain that wreaked such havoc was likely a result of recombination between wild and domesticated bird populations, infecting humans that came in contact with them. Unfortunately, there are no representatives from class Aves in the match database, but of the species included the chimpanzee apparently has the greatest homology with *Influenzavirus A* (Figure 4.11).

Most of the SEVE matches are evenly distributed across the chimpanzee genome, with two notable exceptions. The sequence `5'-CAAGGGATTCTCATACCT-3'` from the neuraminidase (*NA* gene) is repeated 26 times, with four of these repeats from chromosome 2A. Finally, the SEVE `3'-TTCCTTTTCTTCTTCCTC-5'` from the *PB2* polymerase also appears 26 times, four times each within chromosomes 1, 10, and 14.

4.6.3 Influenza A SEVEs by Viral Gene

The chimpanzee *Pan troglodytes* has a large number of matches for this virus, as discussed in the previous section. The nucleoprotein (NP) in *Mus musculus* and PA in the orangutan and gorilla also appear significant (Figure 4.12). The number of

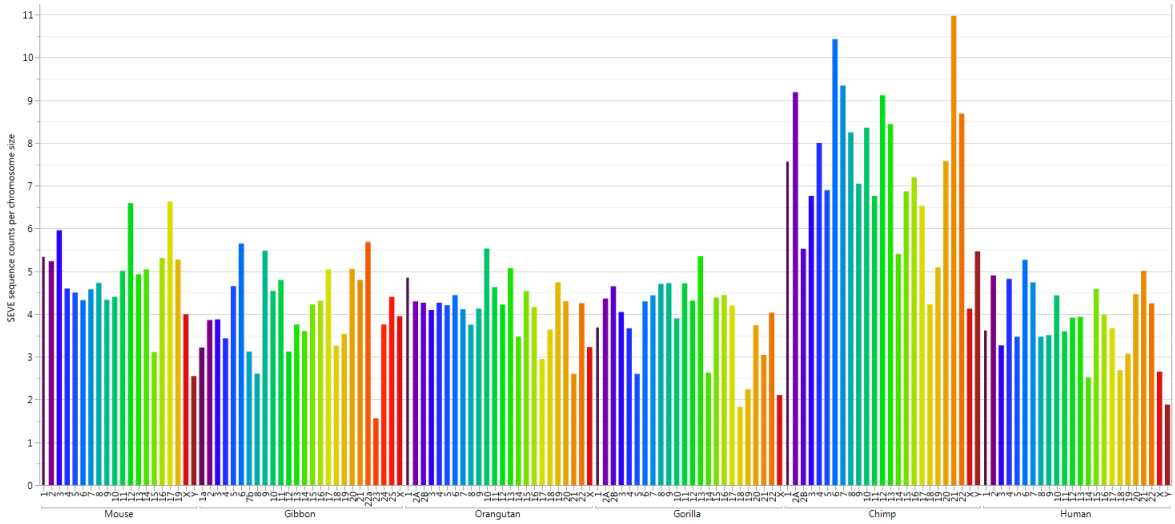

Figure 4.11: *Influenzavirus A* SEVE match count by host name and chromosome.

matches in the human genome are relatively high for all of the viral genes. Note that there are no 3'-OH header or 5'-PO trailer for the influenza virus as the genome is segmented.

There are no particularly frequent sequence matches between the influenza genome and the human, but there are some occurrences within protein coding regions. The sequence `3'-TTCCTTTTCTTCTTCCTC-5'` from the viral *PB2* polymerase gene is found within the human *CDC42BPA* gene that encodes the CDC42 binding protein kinase α. SEVE `5'-ATGCTGTGGATGTTGACG-3'` from the viral matrix protein *M2* is contained within the *RNF14* gene for ring finger protein 14 on chromosome 5. The *SLC7A2* gene for the Y+ cationic amino acid transporter on chromosome 8 contains sequence `5'-TATATGAACACTCAAATC-3'` from the PA polymerase. The *CSAD* gene for the human cysteine sulfinic acid decarboxylase enzyme contains the SEVE `3'-CACTGCACCACCTTGTCT-5'` from the *PB2* polymerase. Finally, the *VOPP1* gene for the vesicular prosurvival protein located on chromosome 7 that is overexpressed

in cancer cells contains the sequence `5'-AGGACAGGTCAGCGTTCA-3'` from the viral nucleoprotein (*NP*) gene.

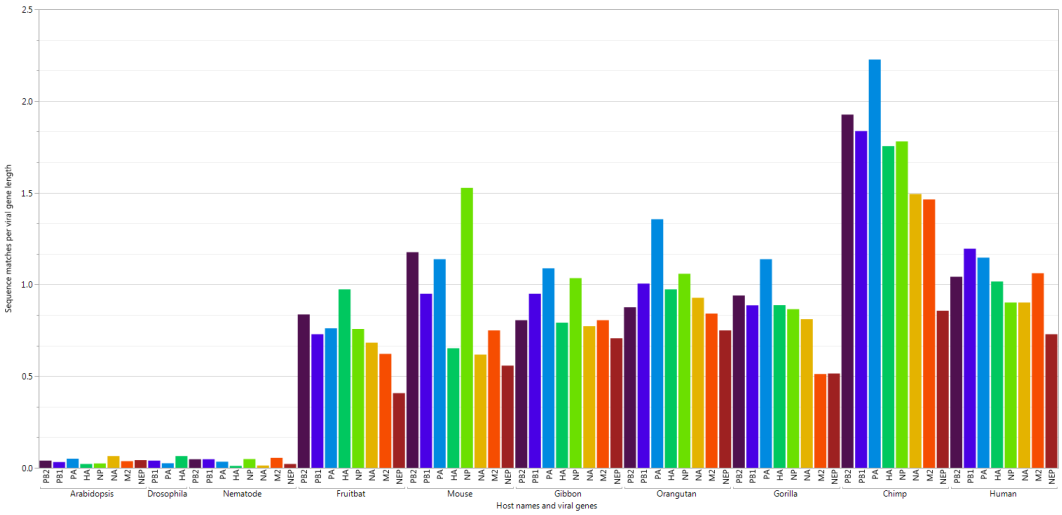

Figure 4.12: *Influenzavirus* A SEVE match count per viral gene and normalized by gene size.

4.7 SEVEs in miRBase

The miRBase database contains 28,645 miRNA entries from 219 eukaryotic species and was last updated in June 2014. The data can be retrieved in two forms. The hairpin form of the miRNA is the initial state following transcription that folds into a double-stranded RNA hairpin structure. The mature form is the RNA molecule that remains after the miRNA has been cleaved by the RNase III Dicer enzyme and the RNA-induced silencing complex (RISC). The miRBase hairpin sequences were compared to the SEVE entries in the MatchDatabase for homology. *Pteropus alecto*, the black flying fox, and *Nomascus leucogenys*, the gibbon, were absent from the miRBase and so could not be compared.

Only one SEVE sequence had a match, `TTCTCCTCCTCCTCCACC`. The sequence exists in chromosomes II, III, and IV of the *Arabidopsis thaliana* genome, and within the gene that encodes the Nef protein in HIV-1. The Nef protein name stands for "negative factor" and it is one of the virulence factors responsible for promoting survival and reproduction of the virus. The miRNA in Arabidopsis was identified by Breakfield et al. in 2012 [96].

The same sequence occurs several times in the match database, including the parathymosin pseudogene (PTMS) RNA sequence in *Pan troglodytes*,[2] and *Nomascus leucogenys*. The sequence is also in the interleukin 1 receptor antagonist (IL1RN) gene. There is one more appearance within the *ACHE* gene that encodes the acetylcholinesterase (Yt blood group) in *Nomascus leucogenys*.

4.8 Randomly Generated Genome

Having assembled a database of nearly 50,000 SEVE sequences in twelve host genomes, and with the remaining potential for tens of thousands more, begs the question of whether these matches are simply the result of chance. To help answer that question, a random eukaryotic organism with a single chromosome 1 Gbp in length has been generated. The `GenomeScanner` has been tested against this *Randomus organismus* with the same five viral genomes as the other twelve actual organisms.

Revisiting the data from Figure 4.2, with the addition of the data from *Randomus organismus*, one can infer that the observed homology implied by the number of SEVE matches is not merely random. The ratio of match count to host genome size is low for the randomly generated genome even when compared to the plant and

[2]NCBI Reference Sequence XR_675047.1

invertebrate data (Figure 4.13). The match data for the random organism are also much more evenly distributed than for the naturally occurring ones.

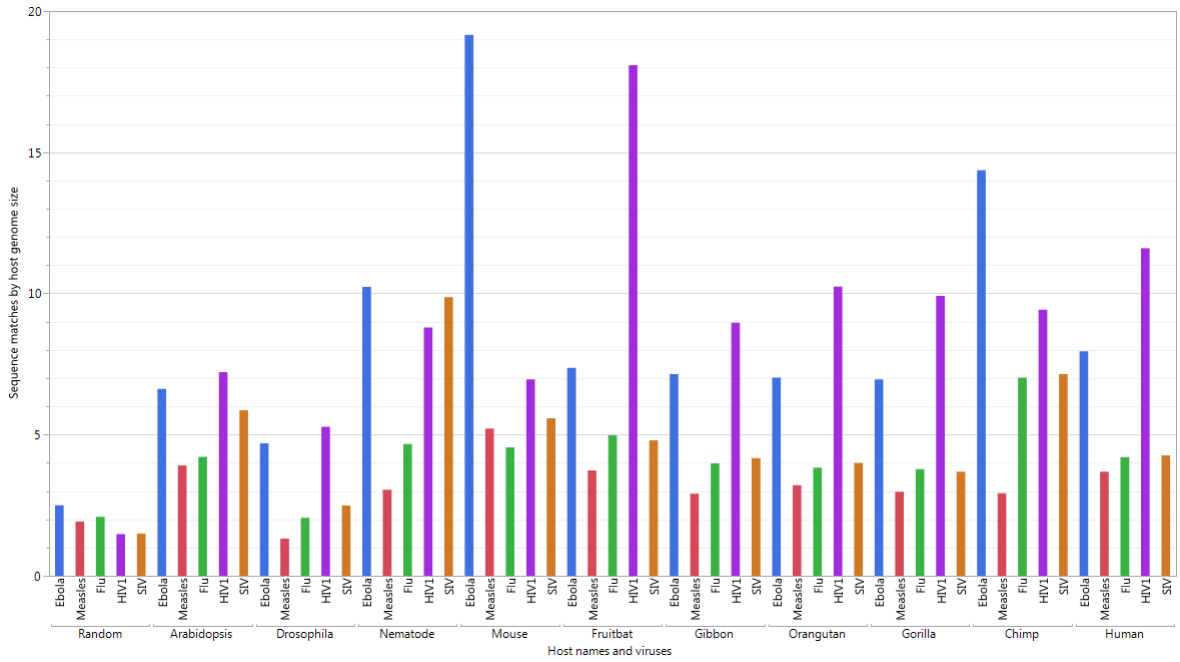

Figure 4.13: Ratio of SEVE sequences to host genome sizes by host and virus species with random organism included.

4.9 Chromosome Bands

Exhaustive SEVE sequence match data of length 18 has been generated for the human and mouse genomes compared to the HIV-1 virus, meaning that the step size of the `GenomeScanner` was set to one (`-step=1`). The original data set collected with the step size of 3 produced 2,077 total matches for the human and 9,672 matches for the mouse. The exhaustive data set includes 5,694 matches from the human genome and 23,666 matches from the mouse. The ratio for the two data sets are 2.75 and 2.5,

respectively. These are nearly 3:1 ratios, indicating that the original data set is a representative sample of all SEVEs.

Merging the comprehensive data with the cytological bands allows the creation of chromosome maps for a given organism. Figure 4.14 provides a plot of SEVE clusters, meaning the number of sequences within each band on the y-axis and the chromosome number along the x-axis.

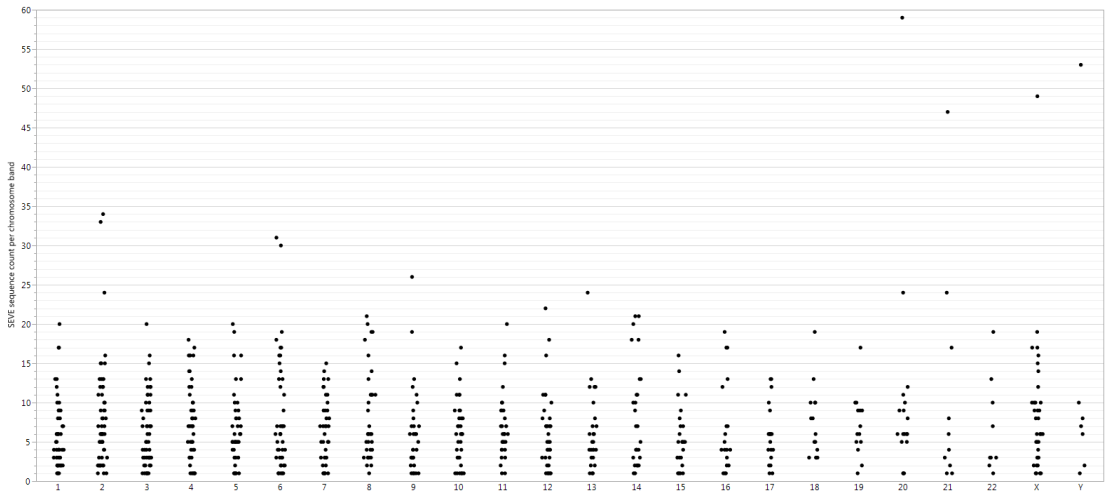

Figure 4.14: HIV-1 SEVE sequence matches by human chromosome bands.

Delving more deeply into the data from organism to chromosome level, the majority of the chromosome 2 matches occur either at the top of the short (p) arm in band 2p25.3 or at the bottom of the long (q) arm at 2q37.3. The X chromosome also contains most matches at the top of the small arm within the Xp22.33 band, perhaps implying that small endogenous viral elements are more likely to be inserted at the ends of chromosomes, or migrate to the telomeres over time. See Figure 4.15 for the chromosome 2 example. Chromosome diagrams courtesy of NCBI.[3]

[3]http://ghr.nlm.nih.gov/chromosome

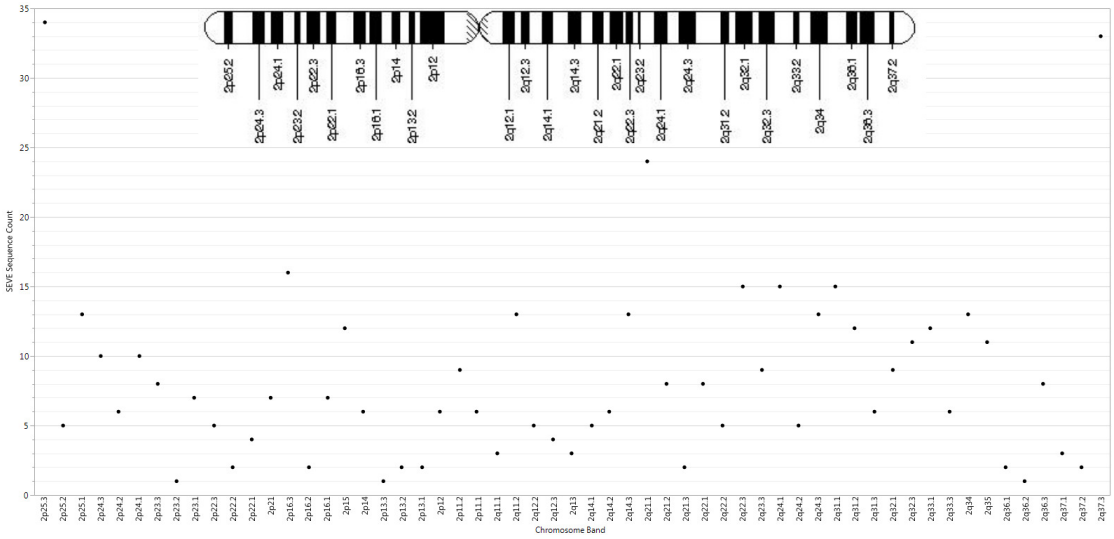

Figure 4.15: HIV-1 SEVE sequence matches in human chromosome 2 bands.

4.10 Most Frequent SEVE Sequences

Although there are 47,480 SEVE matches in the `MatchDatabase`, a remarkable 18,867 of those matches are copies of the most frequently occurring sequences. This means that nearly 40% of the SEVE matches are duplications of the same twelve sequences.

The top two are the complementary pair from the HIV-1 *rev* gene, `5'-AGAGAGAGAC-AGAGACAG-3'` and `3-TCTCTCTCTGTCTCTGTC-5`. The pair was discussed at the beginning of this chapter with respect to the pervasiveness in the *Mus musculus* genome. However, the sequences are also present to varying degrees in all ten of the multicellular organisms.

The third sequence with 1,823 matches is `5'-TGAGTTCCAGGCCAGCCT-3'` from the Ebola *VP35* gene. This sequence is only found in the mouse and the primates human, gorilla, gibbon, chimp, but not in the orangutan genome. Interestingly, the complementary sequence is absent from the database.

Another pair of complementary sequences from the 5' header of the ebolavirus genome, `5'-AAAAATTTAAAAATAAAT-3'` and `3'-TTTTTAAATTTTTATTTA-5'`, are next with 655 and 1,600 matches, respectively. These sequences can be found in all ten multicellular genomes with the exception of the fruitfly *D. melanogaster*. Interestingly, the sequence `TTAAAA` in the center of this SEVE is a common pre-insertion site motif in the human genome [97].

Sequence `5'-GTTCCAGGCCAGCCTGGC-3'`, also from the Ebola *VP35* gene, occurs 369 times in the mouse and primate genomes. Refer to Figure 4.16 for the remaining six sequences.

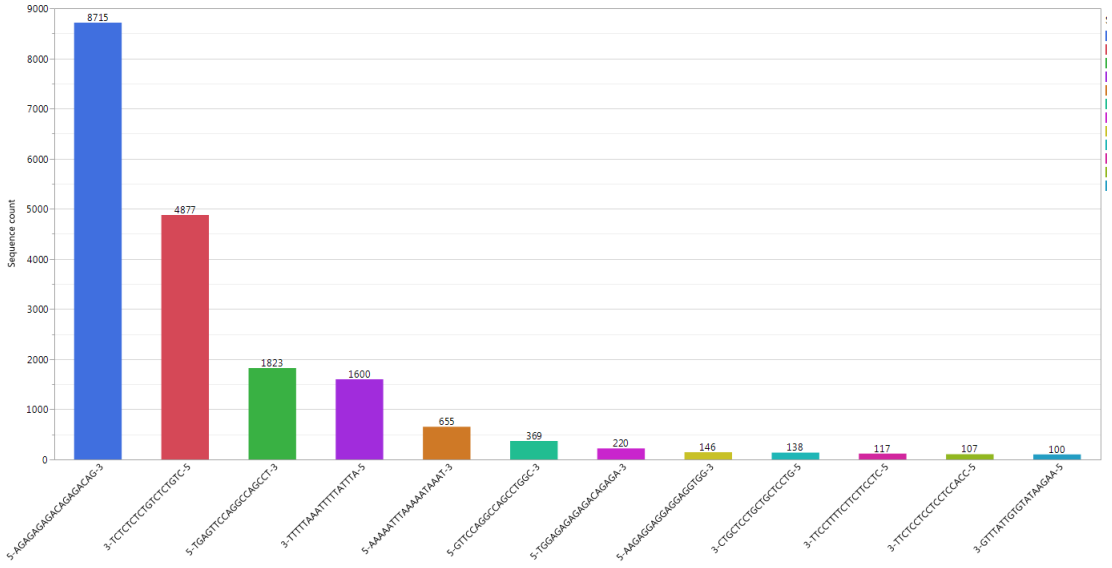

Figure 4.16: Most frequent SEVE sequences in the `MatchDatabase`.

4.11 Scalability and Efficiency

Any parallel architecture or algorithm should scale well with increasing data size (n). Among the variable input parameters to the `GenomeScanner` are the chromosome sizes of the host genomes. One chromosome is typically stored per file, so there is a direct relation to the input size. Timing data was collected while scanning the human, orangutan, and mouse host genomes for all possible SEVEs from the HIV-1 viral genome (i.e., with a step size of one).

Plotting these data with run time on the y-axis, file size on the x-axis, and performing a best fit through the points produces the graph in Figure 4.17. The clear linear relationship indicates that the algorithm run time scales linearly with input file size. These benchmark data were generated using the Boyer-Moore string search algorithm.

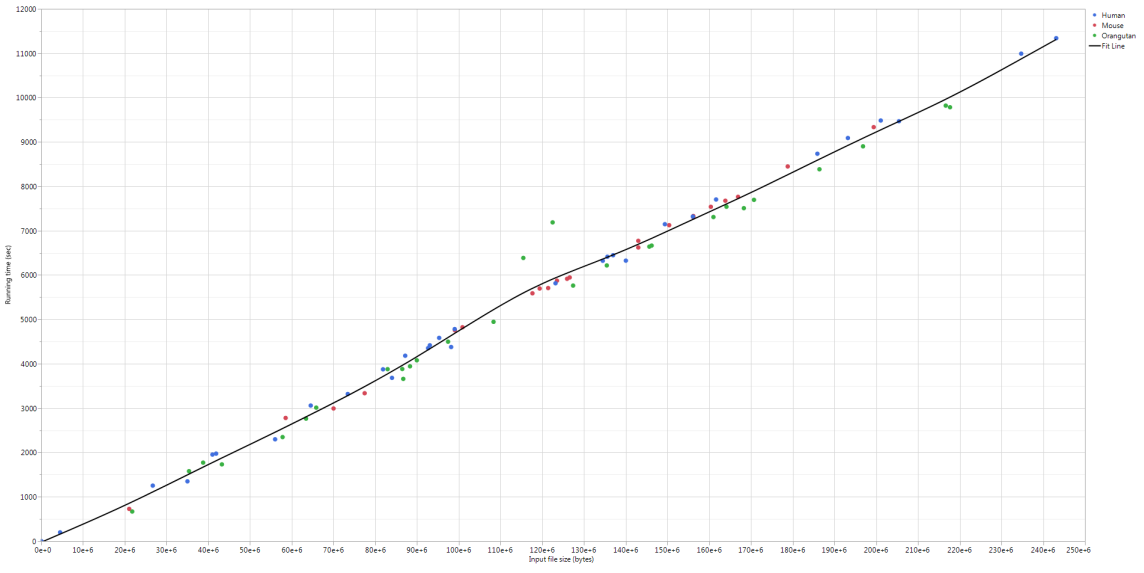

Figure 4.17: `GenomeScanner` scalability graph including the file sizes from three input genomes (Human, Mouse, and Orangutan) versus running time, indicating a clear linear relationship.

Another input parameter that can be varied for a given experiment is the SEVE sequence length (k). Timing data were collected by ranging k from 10 to 30 bp, while comparing human chromosome 22 to the HIV-1 virus with a step size of one. These are the same conditions used to generate the string search algorithm comparison from Table 3.2.

These data indicate that the algorithm runs comparatively longer for smaller k as it results in more subsequences, and therefore a greater number of searches. Larger k values result in fewer subsequences and so less overhead time is required for thread management. Figure 4.18 represents a plot of these results with a vertical line at k=18, the SEVE length used to populate the `MatchDatabase`. The algorithm becomes more efficient near that point as the overhead ceases to be dominant, and the overall runtime grows more linear.

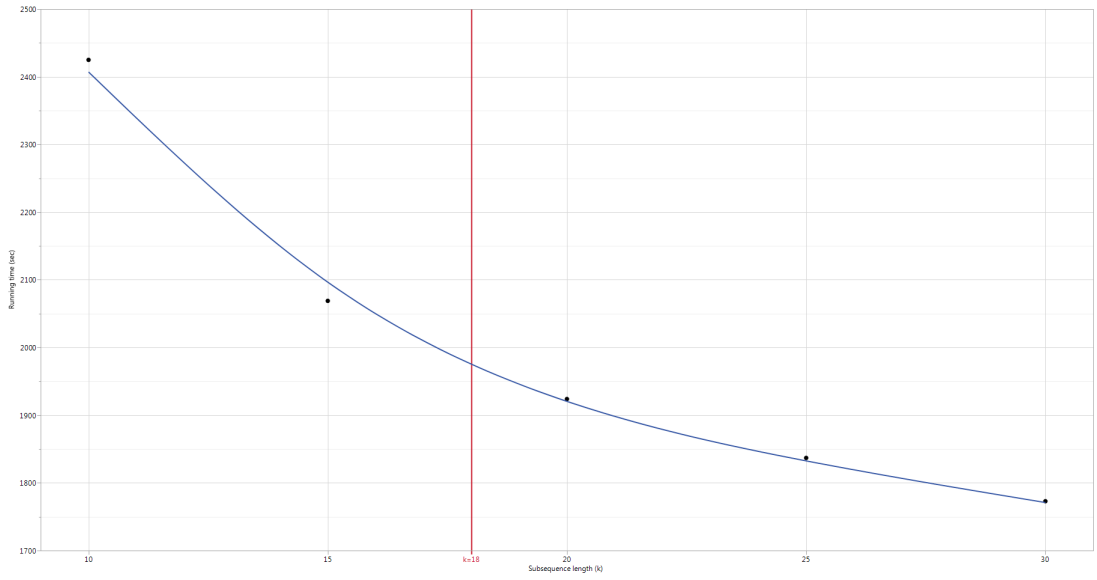

Figure 4.18: `GenomeScanner` efficiency graph of subsequence size k fit against running time using human chromosome 22 and the HIV-1 virus.

CHAPTER 5

CONCLUSIONS

5.1 Future Work

There are a number of ways that this research could be expanded. First and foremost is the exhaustive collection of every SEVE match in each organism by decreasing the nucleotide step size passed to the `GenomeScanner` from three to one. These data have already been collected for the HIV-1 virus.

The set of viruses under study could be expanded to other Baltimore groups than V and VI, for example a dsDNA virus (Group I) such as the herpesvirus, a ssDNA virus (Group II) such as parvovirus, a dsRNA virus (Group III) such as rotavirus, a +ssRNA virus (Group IV) such as rhinovirus (common cold), and a dsDNA retrovirus (Group VII) such as hepadnavirus (hepatitis B). Plans to study the *Zikavirus*, a group IV member of the Flavirvirus family, are also being developed [108].

The host range could be expanded as well. Including, for example, a plant virus such as tobacco mosaic could be searched for within plant genomes, or bacteriophage sequences within bacteria. Additional animal genomes such as the dog (*Canis lupus familiaris*) and cat (*Felis catus*) could also be considered. The sooty mangabey genome (*Cercocebus atys*) has now been sequenced as well (although not annotated).[1]

[1] https://www.hgsc.bcm.edu/non-human-primates/sooty-mangabey-genome-project

Mosquito species such as *Aedes aegypti* are frequent viral hosts and would be interesting to consider as well [109], due to their importance as disease vectors.

The `GenomeScanner` software could be modified to support CUDA string matching, such as Kouzinopoulos and Konstantinos have done [99]. Other bioinformatics projects implemented in CUDA include the `G-BLASTN` implementation of the BLAST algorithm by Zhao and Chu [100], and `GAMUT`, a GPU accelerated microRNA analyzer developed by Wang et al. [101]. Support for additional string searching algorithms could be added, such as the Backward Nondeterministic DAWG Matching algorithm (BNDM) [110] or the Zhu-Takaoka algorithm [111]. The scanner could also be expanded to include other alphabets, such as amino acids or written text.

The SEVE `MatchDatabase` could be compared to other RNA databases, such as Rfam. The Rfam database of RNA families is another online repository of noncoding RNA genes, cis-regulatory elements, and self-splicing molecules. The repository is derived from covariance models similar to the hidden Markov models employed by the related Pfam database for protein families [98].

Rather than simply storing the `MatchDatabase` results in a flat JSON database, the data could be stored in a full featured JSON database such as MongoDB or a distributed database like Cassandra. Other examples of NoSQL databases in the field of bioinformatics are the `LNCipedia` for lncRNA transcripts [112] and the `BIGNAsim` database for nucleic acid simulation data [113].

With the data stored in a proper database, the results could be served on the web with an interface for browsing results, comparing to other databases or even submitting jobs, similarly to the BLAST, BLAT, Ensembl, or UCSC search tools.

5.2 Summary

The coevolution of viruses and cells likely traces back to the origins of life. The complex interplay between dueling nucleic acids through evolutionary time leaves evidence of its passage in the genomes of the organisms that survived to pass hereditary information to their offspring. Endogenous viral elements have been identified across all branches of life and all classes of viruses.

The work of identifying and understanding these common sequences is here extended with the development of a parallel computing system capable of locating all exact genetic matches of a given length between the genomes of the viral invaders and the defending hosts. The software package also provides the ability to cross reference those matches with the NCBI databases via BLAST searches to determine which, if any, coding regions the SEVEs fall within. The sequences can also be precisely located within chromosomal bands by interfacing with the cytological data from the UCSC genome browser.[2]

There are insufficient data to decisively conclude whether the presence of SEVEs within a host genome confers immunity to the host, assistance to the virus, or both. However, considerable quantities of data have been collected and numerous significant sequences have been identified. It is a striking indication of the conservation of

[2]https://genome.ucsc.edu/

genetic information across species that so much of the homology observed across three kingdoms and five viruses can be attributed to a dozen small sequences.

Additional data collection and data mining are required to draw any definitive conclusions. Some of these SEVE sequences may be significant, and others may be random, but the tools have now been developed to continue searching for their meaning.

REFERENCES

[1] Lander, Eric S., et al. "Initial sequencing and analysis of the human genome." *Nature* 409.6822: 860-921, 2001.

[2] R. Belshaw, V. Pereira, A. Katzourakis, G. Talbot, J. Paes, A. Burt, and M. Tristem. "Long-term reinfection of the human genome by endogenous retroviruses." *Proc. Natl. Acad. Sci. USA*, 101(14): 4894-4899, 2004.

[3] L. D. Ward and M. Kellis. "Evidence of abundant purifying selection in humans for recently acquired regulatory functions." *Science* 337.6102: 1675-1678, 2012.

[4] P. Aiewsakun and A. Katzourakis. "Endogenous viruses: Connecting recent and ancient viral evolution." *Virology* 479: 26-37, 2015.

[5] V. A. Belyi, A. J. Levine and A. M. Skalka. "Unexpected Inheritance: Multiple Integrations of Ancient Bornavirus and Ebolavirus/Marburgvirus Sequences in Vertebrate Genomes." *PLoS Pathogens* 6(7): e1001030, 2010.

[6] S. Kumar and S. Subramanian. "Mutation rates in mammalian genomes." *Proc. Natl. Acad. Sci. USA* 99: 803808, 2002.

[7] Y. Li, J. Lu, Y. Han, X. Fan, and S. W. Ding. "RNA interference functions as an antiviral immunity mechanism in mammals." *Science*, 342(6155): 231-234, 2013.

[8] E. B. Chuong, N. C. Elde, and C. Feschotte. "Regulatory evolution of innate immunity through co-option of endogenous retroviruses." *Science* 351.6277: 1083-1087, 2016.

[9] A. L. Boi, A. iber, and R. Podgornik. "Statistical analysis of sizes and shapes of virus capsids and their resulting elastic properties." *Journal of biological physics*, 39(2): 215-228, 2013.

[10] J. Zipprich, K. Winter, J. Hacker, D. Xia, J. Watt, and K. Harriman "Measles outbreakCalifornia, December 2014February 2015." *MMWR Morb Mortal Wkly Rep*, 64(6): 153-154, 2015.

[11] J. S. Welsh. "Contagious cancer." *The oncologist*, 16(1): 1-4, 2011.

[12] C. J. Konstantoulas, and S. Indik. "Mouse mammary tumor virus-based vector transduces non-dividing cells, enters the nucleus via a TNPO3-independent pathway and integrates in a less biased fashion than other retroviruses." *Retrovirology*, 11(1): 1-15, 2014.

[13] A. van Dijk, E. V. Makeyev, and D. H. Bamford. "Initiation of viral RNA-dependent RNA polymerization." *Journal of general virology*, 85(5), 1077-1093, 2004.

[14] A. Katzourakis and R. J. Gifford. "Endogenous Viral Elements in Animal Genomes". *PLoS Genetics*, Vol. 6, Issue 11; e1001191, 2010.

[15] A. Lee, A. Nolan, J. Watson and M. Tristem. "Identification of an ancient endogenous retrovirus, predating the divergence of the placental mammals." *Phil Trans R Soc B*, 368: 20120503, 2013.

[16] M. Horie, Y. Kobayashi, Y. Suzuki and K. Tomonaga. "Comprehensive analysis of endogenous bornavirus-like elements in eukaryote genomes." *Phil Trans R Soc B* 368: 20120499, 2013.

[17] A. Aswad and A. Katzourakis. "Paleovirology and virally derived immunity." *Trends in ecology and evolution*, 27(11): 627-636, 2012.

[18] S. Kalyana-Sundaram, C. Kumar-Sinha, S. Sunita, et al. "Expressed Pseudogenes in the Transcriptional Landscape of Human Cancers." *Cell*, 149(7): 16221634, 2012.

[19] B. M. Curtis, S. Scharnowske, and A. J. Watson. "Sequence and expression of a membrane-associated C-type lectin that exhibits CD4-independent binding of human immunodeficiency virus envelope glycoprotein gp120". *Proc. Natl. Acad. Sci. USA*, 89 (17): 835660, 1992.

[20] D. J. Taylor, R. W. Leach, and J. Bruenn. "Filoviruses are ancient and integrated into mammalian genomes." *BMC Evolutionary Biology*, 10(1): 193, 2010.

[21] C. E. Shannon. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27 (3): 379423, 1948.

[22] E. Zuckerkandl, L. Pauling. "Molecules as documents of evolutionary history." *J. Theor. Biol.* 8: 357366, 1965.

[23] J. R. Jungck, R. M. Friedman, "Mathematical tools for molecular genetics data: An annotated bibliography." *Bull. Math. Biol.* 46: 699744, 1984.

[24] M. Dayhoff Schwartz. "A model of evolutionary change in proteins." *Atlas of protein sequence and structure.*, 1978.

[25] S. Henikoff and J.G. Henikoff. "Amino acid substitution matrices from protein blocks." *Proc. Natl. Acad. Sci. USA* 89.22: 10915-10919, 1992.

[26] T. F. Smith and M. S. Waterman. "Identification of common molecular subsequences." *Journal of molecular biology*, 147(1): 195-197, 1981.

[27] D. J. Lipman and W. R. Pearson. "Rapid and sensitive protein similarity searches." *Science*, 227(4693): 1435-1441, 1985.

[28] S. F. Altschul, W. Gish, and W. Miller. "Basic Local Alignment Search Tool." *Journal of Molecular Biology*, 215: 403-410, 1990.

[29] W. J. Kent. "BLATthe BLAST-like alignment tool." *Genome research*, 12(4): 656-664, 2002.

[30] R. Chenna, H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, and J. D. Thompson. "Multiple sequence alignment with the Clustal series of programs." *Nucleic acids research*, 31(13): 3497-3500, 2003.

[31] B. Yoon. "Hidden Markov models and their applications in biological sequence analysis." *Current genomics*, 10.6: 402-415, 2009.

[32] Y. Saeys, I. Inza, and P. Larraaga. "A review of feature selection techniques in bioinformatics." *Bioinformatics*, 23.19: 2507-2517, 2007.

[33] B. G. Hall. "Building phylogenetic trees from molecular data with MEGA." *Molecular biology and evolution*, mst012, 2013.

[34] D. Knuth, J. H. Morris,V. Pratt. "Fast pattern matching in strings." *SIAM Journal on Computing*, 6(2): 323350, 1977.

[35] R. S. Boyer, J. S. Moore. "A Fast String Searching Algorithm." *Communications of the ACM*, 20(10): 762772, 1977.

[36] R. M. Karp, M. O. Rabin. "Efficient randomized pattern-matching algorithms." *IBM Journal of Research and Development* 31(2): 249260, 1987.

[37] C. Allauzen, M. Rochemore , M. Affinot. "Factor oracle: a new structure for pattern matching." *SOFSEM99, Theory and Practice of Informatics, Lecture Notes in Computer Science*, 1725: 291306, 1999.

[38] S. Faro and T. Lecroq, "Efficient Variants of the Backward-Oracle-Matching Algorithm." *International Journal of Foundations of Computer Science*, 20(6): 967984, 2009.

[39] E. Ukkonen. "On-line construction of suffix trees." *Algorithmica*, 14(3): 249260, 1995.

[40] D. Baltimore. "Expression of animal virus genomes." *Bacteriological Reviews*, 35(3): 235, 1971.

[41] M. Horie, T. Honda, Y. Suzuki, Y. Kobayashi, T. Daito, T. Oshida, K. Ikuta, P. Jern, T. Gojobori, J. M. Coffin, and K. Tomonaga. "Endogenous non-retroviral RNA virus elements in mammalian genomes." *Nature*, 463: 7277-84, 2010.

[42] A. Kozomara, S. Griffiths-Jones. "miRBase: integrating microRNA annotation and deep-sequencing data." *Nucleic acids research*, gkq1027, 2010.

[43] S. Sagan and P. Sarnow. "RNAi, Antiviral After All." *Science*, 342: 207, 2013.

[44] B. Sumpter, R. Dunham, S. Gordon, J. Engram, M. Hennessy, A. Kinter, M. Paiardini, B. Cervasi, N. Klatt, and H. McClure. "Correlates of preserved CD4(+) T cell homeostasis during natural, nonpathogenic simian immunodeficiency virus infection of sooty mangabeys: implications for AIDS pathogenesis." *Journal of Immunology*, 178(3): 1680-1691, 2007.

[45] F. M. Boni et al. "Homologous recombination is very rare or absent in human influenza A virus." *Journal of Virology*, 82.10: 4807-4811, 2008.

[46] X. Pourrut et al. "Spatial and temporal patterns of Zaire ebolavirus antibody prevalence in the possible reservoir bat species." *Journal of Infectious Diseases*, 196.Supplement 2: S176-S183, 2007.

[47] C. Cowled et al. "Characterisation of novel microRNAs in the Black flying fox (Pteropus alecto) by deep sequencing." *BMC genomics*, 15.1: 1, 2014.

[48] K. Lindblad-Toh, M. Garber, O. Zuk, M. F. Lin, B. J. Parker, S. Washietl, P. Kheradpour, J. Ernst, G. Jordan, E. Mauceli, et al. "A high-resolution map of human evolutionary constraint using 29 mammals." *Nature*, 478(7370): 476-82, 2011.

[49] S. Faro, T. Lecroq. "The Exact Online String Matching Problem: a Review of the Most Recent Results." *ACM Computing Surveys*, Vol. V, No. N, Art. A, 2013.

[50] P. J. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke and M. J. de Hoon. "Biopython: freely available Python tools for computational molecular biology and bioinformatics." *Bioinformatics*, 25(11): 1422-1423, 2009.

[51] C. Camacho et al. "BLAST+: architecture and applications." *BMC bioinformatics*, 10.1: 1, 2009.

[52] L. Dalcn et al. "MPI for Python: Performance improvements and MPI-2 extensions." *Journal of Parallel and Distributed Computing* 68.5: 655-662, 2008.

[53] J. H. Kuhn, S. Becker, H. Ebihara, T. W. Geisbert, K. M. Johnson, Y. Kawaoka, W. I. Lipkin, A. I. Negredo et al. "Proposal for a revised taxonomy of the family Filoviridae: Classification, names of taxa and viruses, and virus abbreviations." *Archives of Virology* 155(12): 2083103, 2010.

[54] C. P. Alvarez et al. "C-type lectins DC-SIGN and L-SIGN mediate cellular entry by Ebola virus in cis and in trans." *Journal of virology* 76.13: 6841-6844, 2002.

[55] A. M. Khatib, F. Sfaxi. "FURIN (furin (paired basic amino acid cleaving enzyme)." *Atlas Genet Cytogenet Oncol Haematol*, 16(9): 639-643, 2012.

[56] D. W. Leung et al. "Structural basis for dsRNA recognition and interferon antagonism by Ebola VP35." *Nature structural & molecular biology*, 17.2: 165-172, 2010.

[57] J. M. Licata et al. "Overlapping motifs (PTAP and PPEY) within the Ebola virus VP40 protein function independently as late budding domains: involvement of host proteins TSG101 and VPS-4." *Journal of virology*, 77.3: 1812-1819, 2003.

[58] J. Modrof, S. Becker, and E. Muhlberger. "Ebola virus transcription activator VP30 is a zinc-binding protein." *Journal of virology*, 77.5: 3334-3338, 2003.

[59] L. W. Leung et al. "Ebola virus VP24 binds karyopherin α1 and blocks STAT1 nuclear accumulation." *Journal of virology*, 80.11: 5156-5167, 2006.

[60] N. Biedenkopf et al. "Phosphorylation of Ebola virus VP30 influences the composition of the viral nucleocapsid complex impact on viral transcription and replication." *Journal of Biological Chemistry*, 288.16: 11165-11174, 2013.

[61] A. L. Rasmussen et al. "Host genetic diversity enables Ebola hemorrhagic fever pathogenesis and resistance." *Science*, 346.6212: 987-991, 2014.

[62] M. Bermejo et al. "Ebola outbreak killed 5000 gorillas." *Science*, 314.5805: 1564-1564, 2006.

[63] S. K. Lee, E. J. Park, H. S. Lee, Y. S. Lee, and J. Kwon. "Genome-wide screen of human bromodomain-containing proteins identifies Cecr2 as a novel DNA damage response protein." *Molecules and cells*, 34(1): 85-91, 2012.

[64] J. Lu, M. Chatterjee, H. Schmid, S. Beck, and M. Gawaz. "CXCL14 as an emerging immune and inflammatory modulator." *Journal of Inflammation*, 13(1): 1, 2016.

[65] A. V. Marakhonov, A. V. Baranova, and M. Y. Skoblov. "Antisense regulation of human gene MAP3K13: True phenomenon or artifact?." *Molecular Biology*, 42(4): 514-520, 2008.

[66] L. de Witte, M. Bobardt, U. Chatterji, G. Degeest, G. David, T. B. Geijtenbeek, P. Gallay. "Syndecan-3 is a dendritic cell-specific attachment receptor for HIV-1." *Proc. Natl. Acad. Sci. USA*, 104(49): 194649, 2007.

[67] J. A. Smith, R. Daniel. "Following the path of the virus: the exploitation of host DNA repair mechanisms by retroviruses." *ACS Chem Biol*, 1(4): 21726, 2006.

[68] S. Debaisieux, F. Rayne, H. Yezid, B. Beaumelle. "The ins and outs of HIV-1 Tat.". *Traffic*, 13(3): 35563, 2012.

[69] G. R. Campbell, E. Pasquier, J. Watkins, V. Bourgarel-Rey, V. Peyrot, D. Esquieu, P. Barbier, J. de Mareuil, D. Braguer, P. Kaleebu, D. L. Yirrell, and E. P. Loret. "The glutamine-rich region of the HIV-1 Tat protein is involved in T-cell apoptosis." *J. Biol. Chem.*, 279(46): 48197204, 2004.

[70] L. Abraham, O. T. Fackler. "HIV-1 Nef: a multifaceted modulator of T cell receptor signaling." *Cell Communication and Signaling*, 10(1): 39, 2012.

[71] J. H. Miller, V. Presnyak, and H. C. Smith. "The dimerization domain of HIV-1 viral infectivity factor Vif is required to block APOBEC3G incorporation with virions." *Retrovirology*, 4(1): 81, 2007.

[72] M. Bukrinsky, A. Adzhubei. "Viral protein R of HIV-1." *Rev. Med. Virol.*, 9(1): 3949, 1999.

[73] E. Estrabaud, E. Le Rouzic, S. Lopez-Vergs, M. Morel, N. Beladouni, R. Benarous, C. Transy, C. Berlioz-Torrent, and F. Margottin-Goguet. "Regulated degradation of the HIV-1 Vpu protein through a betaTrCP-independent pathway limits the release of viral particles." *PLoS Pathogens*, 3(7): e104, 2007.

[74] C. Torresilla, J. Mesnard, and B. Barbeau. "Reviving an old HIV-1 gene: the HIV-1 antisense protein." *Current HIV research*, 13.2: 117-124, 2015.

[75] Y. Zheng, H. Yu, and B. Matija Peterlin. "Human p32 protein relieves a post-transcriptional block to HIV replication in murine cells." *Nature cell biology*, 5.7: 611-618, 2003.

[76] B. L. Fogel, E. Wexler, A. Wahnich, T. Friedrich, C. Vijayendran, F. Gao, and D. H. Geschwind. "RBFOX1 regulates both splicing and transcriptional networks in human neuronal development." *Human molecular genetics*, dds240, 2012.

[77] M. Peeters, V. Courgnaud, B. Abela. "Genetic Diversity of Lentiviruses in Non-Human Primates." *AIDS Reviews*, 3: 310, 2001.

[78] M. Kim, B. Chen, R.E. Hussey, Y. Chishti, D. Montefiori, J. A. Hoxie, O. Byron, G. Campbell, S. C. Harrison, and E. L. Reinherz. "The stoichiometry of trimeric SIV glycoprotein interaction with CD4 differs from that of anti-envelope antibody Fab fragments." *J Biol Chem.*, 276(46): 42667-76, 2001.

[79] D. Ayinde, C. Maudet, C. Transy, C., and F. Margottin-Goguet. "Review Limelight on two HIV/SIV accessory proteins in macrophage infection: Is Vpx overshadowing Vpr?." 2010.

[80] M. Worobey et al. "Island biogeography reveals the deep history of SIV." *Science* 329.5998: 1487-1487, 2010.

[81] M. B. Gardner and P. A. Luciw. "Animal models of AIDS." *The FASEB Journal* 3.14: 2593-2606, 1989.

[82] G. Lu, G. F. Gao, and J. Yan. "The receptors and entry of measles virus: a review." *Chinese journal of biotechnology* 29.1: 1-9, 2013.

[83] S. Heidmeier et al. "A single amino acid substitution in the measles virus F 2 protein reciprocally modulates membrane fusion activity in pathogenic and oncolytic strains." *Virus research* 180 (2014): 43-48, 2014.

[84] X. Q. Wang and J. A. Rothnagel. "5'Untranslated regions with multiple upstream AUG codons can support lowlevel translation via leaky scanning and reinitiation." *Nucleic acids research* 32.4: 1382-1391, 2004.

[85] P. Devaux and R. Cattaneo. "Measles virus phosphoprotein gene products: conformational flexibility of the P/V protein amino-terminal domain and C protein infectivity factor function." *Journal of virology* 78.21: 11632-11640, 2004.

[86] E. Avota, S. Koethe, and S. SchneiderSchaulies. "Membrane dynamics and interactions in measles virus dendritic cell infections." *Cellular microbiology*, 15(2): 161-169, 2013.

[87] Y. Furuse, A. Suzuki, and H. Oshitani. "Origin of measles virus: divergence from rinderpest virus between the 11th and 12th centuries." *Journal of virology* 7: 52, 2010.

[88] B. Wang and J. Brand-Miller. "The role and potential of sialic acid in human nutrition." *European journal of clinical nutrition* 57(11): 1351-1369, 2003.

[89] E. C. Hutchinson and E. Fodor. "Transport of the influenza virus genome from nucleus to nucleus." *Viruses* 5(10): 2424-2446, 2013.

[90] E. Decroly, F. Ferron, J. Lescar, and B. Canard. "Conventional and unconventional mechanisms for capping viral mRNA." *Nature Reviews Microbiology* 10(1): 51-65, 2012.

[91] H. Zheng, H. A. Lee, P. Palese, P., and A. Garca-Sastre. "Influenza A virus RNA polymerase has the ability to stutter at the polyadenylation site of a viral RNA template during RNA replication." *Journal of virology* 73(6): 5240-5243, 1999.

[92] T. Hayashi, L. A. MacDonald, and T. Takimoto. "Influenza A virus protein PA-X contributes to viral growth and suppression of the host antiviral and immune responses." *Journal of virology* 89(12): 6442-6452, 2015.

[93] T. Yoshizumi, T. Ichinohe, O. Sasaki, et al. "Influenza A virus protein PB1-F2 translocates into mitochondria via Tom40 channels and impairs innate immunity." *Nature communications* 5, 2014.

[94] E. Alvarado-Facundo, Y. Gao, R. M. Ribas-Aparicio, et al. "Influenza virus M2 protein ion channel activity helps to maintain pandemic 2009 H1N1 virus hemagglutinin fusion competence during transport to the cell surface." *Journal of virology* 89(4): 1975-1985, 2015.

[95] B. G. Hale, R. E. Randall, J. Ortn, D. Jackson. "The multifunctional NS1 protein of influenza A viruses." *Journal of General Virology* 89(10): 2359-2376, 2008.

[96] N. W. Breakfield, et al. "High-resolution experimental and computational profiling of tissue-specific known and novel miRNAs in Arabidopsis." *Genome research* 22.1: 163-176, 2012.

[97] B. B. Dev, A. Malik, and K. Rawal. "Detecting motifs and patterns at mobile genetic element insertion site." *Bioinformation* 8.16: 777, 2012.

[98] S. W. Burge, et al. "Rfam 11.0: 10 years of RNA families." *Nucleic acids research* gks1005, 2012.

[99] C. S. Kouzinopoulos and G. M. Konstantinos. "String matching on a multicore GPU using CUDA." *Informatics* PCI'09. 13th Panhellenic Conference on, IEEE, 2009.

[100] K. Zhao and X. Chu. "G-BLASTN: accelerating nucleotide alignment by graphics processors." *Bioinformatics* 30.10: 1384-1391, 2014.

[101] S. Wang, et al. "GAMUT: GPU accelerated microRNA analysis to uncover target genes through CUDA-miRanda." *BMC medical genomics* 7.Suppl 1: S9, 2014.

[102] I. Retter, et al. "Sequence and characterization of the Ig heavy chain constant and partial variable region of the mouse strain 129S1." *Journal of Immunology* 179.4: 2419-2427, 2007.

[103] B. Rothschild. "Emerging infectious diseases and Primate Zoonoses." *Journal of Primatology* 4: e130, 2015.

[104] H. Takaki, et al. "Dendritic cell subsets involved in type I IFN induction in mouse measles virus infection models." *International journal of biochemistry and cell biology* 53: 329-333, 2014.

[105] T. Saito, et al. "Nucleobindin-2 is a positive regulator for insulin-stimulated glucose transporter 4 translocation in fenofibrate treated E11 podocytes." *Endocrine Journal* 61.9: 933-939, 2013.

[106] H. Zheng H, H. H. Loh, P.Y. Law. "β-Arrestin-Dependent μ-Opioid Receptor-Activated Extracellular Signal-Regulated Kinases (ERKs) Translocate to Nucleus in Contrast to G Protein-Dependent ERK Activation." *Molecular Pharmacology* 73 (1): 17890, 2008.

[107] Q. Li, L. Zhou, M. Zhou, et al. "Preliminary Report: Epidemiology of the Avian Influenza A (H7N9) Outbreak in China." *New England Journal of Medicine* 370(6): 52032, 2013.

[108] G. Kuno and G-JJ Chang. "Full-length sequencing and genomic characterization of Bagaza, Kedougou, and Zika viruses." *Archives of virology* 152.4: 687-696, 2007.

[109] V. Nene, et al. "Genome sequence of *Aedes aegypti*, a major arbovirus vector." *Science* 316.5832: 1718-1723, 2007.

[110] G. Navarro and M. Raffinot, "A Bit-Parallel Approach to Suffix Automata: Fast Extended String Matching." *Proceedings of the 9th Annual Symposium on Combinatorial Pattern Matching, Lecture Notes in Computer Science* 1448: 14-31, 1998.

[111] R. F. Zhu and T. Takaoka. "On improving the average case of the Boyer-Moore string matching algorithm." *Journal of Information Processing* 10(3):173-177, 1987.

[112] P. J. Volders, et al. "LNCipedia: a database for annotated human lncRNA transcript sequences and structures." *Nucleic acids research* 41.D1: D246-D251, 2013.

[113] P. Andrio, et al. "BIGNASim: a NoSQL database structure and analysis portal for nucleic acids simulation data." *Nucleic acids research* 44.D1: D272-D278, 2016.