Quantitative Approaches to the Genomics of Clonal Evolution

Sakellarios Zairis

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2018

ABSTRACT

Quantitative Approaches to the Genomics of Clonal Evolution

Sakellarios Zairis

Many problems in the biological sciences reduce to questions of genetic evolution. Entire classes of medical pathology, such as malignant neoplasia or infectious disease, can be viewed in the light of Darwinian competition of genomes. With the benefit of today's maturing sequencing technologies we can observe and quantify genetic evolution with nucleotide resolution. This provides a molecular view of genetic material that has adapted, or is in the process of adapting, to its local selection pressures. A series of problems will be discussed in this thesis, all involving the mathematical modeling of genomic data derived from clonally evolving populations. We use a variety of computational approaches to characterize over-represented features in the data, with the underlying hypothesis that we may be detecting fitness-conferring features of the biology.

In Part I we consider the cross-sectional sampling of human tumors via RNA-sequencing, and devise computational pipelines for detecting oncogenic gene fusions and oncovirus infections. Genomic translocation and oncovirus infection can each be a highly penetrant alteration in a tumor's evolutionary history, with famous examples of both populating the

cancer biology literature. In order to exert a transforming influence over the host cell, gene fusions and viral genetic programs need to be expressed and thus can be detected via whole transcriptome sequencing of a malignant cell population. We describe our approaches to predicting oncogenic gene fusions (Chapter 2) and quantifying host-viral interactions (Chapter 3) in large panels of human tumor tissue. The alterations that we characterize prompt the larger question of how the genetics of tumors and viruses might vary in time, leading us to the study of serially sampled populations.

In Part II we consider longitudinal sampling of a clonally evolving population. Phylogenetic trees are the standard representation of a clonal process, an evolutionary picture as old as Darwin's voyages on the Beagle. Chapter 4 first reviews phylogenetic inference and then introduces a certain phylogenetic tree space that forms the starting point of our work on the topic. Specifically, Chapter 4 describes the construction of our projective tree space along with an explicit implementation for visualizing point clouds of rescaled trees. The Chapter finishes by defining a method for stable dimensionality reduction of large phylogenies, which is useful for analyzing long genomic time series. In Chapter 5 we consider medically relevant instances of clonal evolution and the longitudinal genetic data sets to which they give rise. We analyze data from (i) the sequencing of cancers along their therapeutic course, (ii) the passaging of a xenografted tumor through a mouse model, and (iii) the seasonal surveillance of H3N2 influenza's hemagglutinin segment. A novel approach to predicting influenza vaccine effectiveness is demonstrated using statistics of point clouds in tree spaces.

Our investigations into clonal processes may be extended beyond naturally occurring genomes. In Part III we focus on the directed clonal evolution of populations of synthetic RNAs *in vitro*. Analogous to the selection pressures exerted upon malignant cells or viral

particles, these synthetic RNA genomes can be evolved against a desired fitness objective. We investigate fitness objectives related to reprogramming ribosomal translation. Chapter 6 identifies high fitness RNA pseudoknot geometries capable of inducing ribosomal frameshift, while Chapter 7 takes an unbiased approach to evolving sequence and structural elements that promote stop codon readthrough.

Contents

List of Figures

List of Tables

Acknowledgements

I thank my loving and supportive family, without whom I could never have embarked on such an exciting journey. I also thank Raul, Francesco, Hossein, Andrew A., Andrew B., Kevin, Jan-Willem, and Chris for traveling the road with me and catalyzing my scientific growth.

In honor of my parents, Ellie and Ignatios.

Chapter 1

Introduction

The guiding theme of this work is the characterization of clonally evolving populations under strong selection using molecular sequence data. Clonal evolution is also referred to as vertical evolution or asexual evolution in the biology literature, but for the sake of consistency we will limit ourselves to the former term. Our understanding of genetic determinants of fitness is immature even for the modest genomes of viruses, to say nothing of the more complex genomes of human cancer cells. Still, it stands to reason that deeply profiling the sequences of clonally evolving populations might reveal genetic regularities, recurrent patterns of variation among individuals. A natural interpretation of regularities in genetic mutation is the conferral of a fitness advantage against the selective pressures of the environment. In this thesis we will be interested both in the particular genetic alterations observed in molecular sequence data as well as the larger patterns of mutation over time. In our discussion of human tumors it is understood that the genome in question is the nuclear, double-stranded DNA of the malignant cell lineage. When we shift to the analysis of seasonal influenza, it is understood that the genome in question is the single-stranded RNA hemagglutinin segment of the virus. As we conclude with the directed evolution of short RNAs, it is understood the genome in question is the randomized RNA construct undergoing iterative *in vitro* selection.

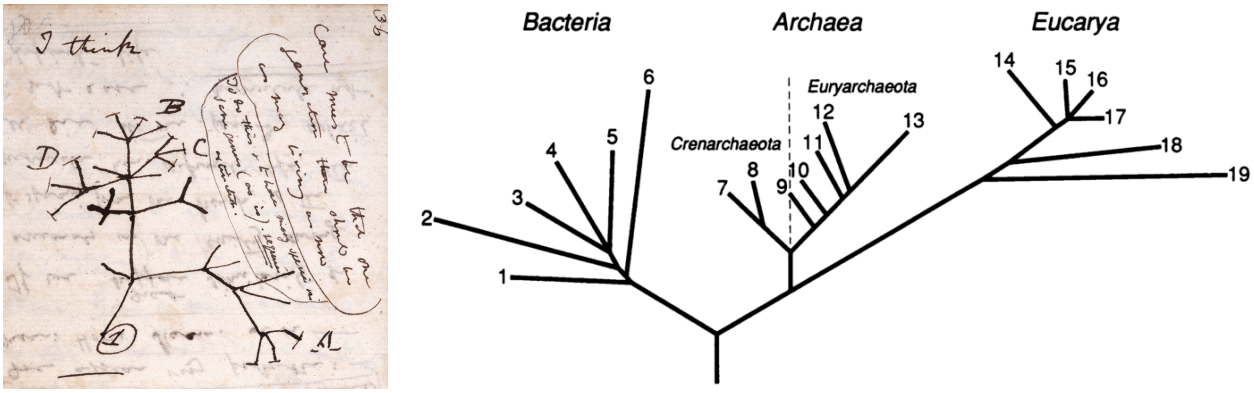


Figure 1.1: **Phylogenetic trees of life:** (Left) Sketch from Darwin's "notebook B" (1837-1838). (Right) Three domains of life arranged in a phylogenetic tree inferred from ribosomal RNA sequencing, image reproduced from [223].

1.1 Heritable Alterations and Common Phylogeny

Charles Darwin's picture of a single tree to describe the relation between species carries a profound corollary, that all extant life shares a common ancestry (Figure 1.1). This idea may represent the most general and useful constraint for the development of theory in biology [117]. Comparisons across taxa are valid and interpretable under the assumption of a common origin, and theoretical biology would be far more difficult if life indeed had myriad independent starting points. While the initial conditions of biology are not known, Darwin bequeathed to us a theory in which we understand phenotypic evolution from any starting point via the action of reproduction and selection.

Reproduction and selection are certainly essential for population adaptation, however they are not sufficient. Without mutation a population's survival would be determined at the outset for a fixed selection environment. Mutation is the final ingredient for true evolution, and it is the concept that was historically hardest to understand until the first rigorous experiments on heredity. In roughly the same era of history as Darwin, Gregor Mendel was making quantitative investigations about what seemed a narrow corner of the biological world, frequencies of inheritance of phenotypic traits in pea plants. In contrast with the poorly defined theories of heritability of the time, which included Darwin's "pangenesis" and notions of "blending inheritance," Mendel's sharp observations of segregation

and independent assortment of genes were a conceptual leap forward. Mendelian inheritance patterns gave phenotypic traits a discrete nature and pointed to an underlying granularity of the hereditary information. In 1952 Alfred Hershey and Martha Chase elegantly settled the debate as to whether protein or nucleic acid acts as the genetic material, at a time when most assumed protein would mediate heredity because of its greater combinatorial complexity (20 monomers rather than 4). Stochastic mutation in nucleic acid, and its subsequent fixation in a population, is the molecular mechanism of Darwin's transformative idea a century earlier.

Modifications to the genetic material are typically qualified with the terms "somatic" or "germline" when discussing the human genome. This distinction is necessary because only a privileged subset of cells (the gametes) in the human organism will transmit the heritable traits. Germline mutation is crucial when studying human evolution and human population genetics, but less central to the study of tumor evolution and tumor population genetics. In the latter case the individual element in the population is the malignant cell, which reproduces asexually. Cancers typically arise on the background of DNA damage in the somatic cells comprising our bodies, thus the ubiquitous term "somatic mutation." Of course, the terms "somatic" or "germline" are not applicable when our discussion turns to the propagation of RNA viruses or synthetic RNA genomes later in the thesis.

1.2 Nucleic Acid Sequencing as an Evolutionary Read-Out

Once it was settled that nucleic acid was indeed the polymer of heredity, sequencing of DNA became a technical challenge of paramount importance. The reason is simply that DNA bears the scars of its evolutionary past, so one must have access to this genetic information in order to understand biological adaption. It comes as a surprise to many that the monomeric composition of proteins (Edman degradation of the mid 1960's) could be deduced more than a decade before nucleic acids could be sequenced.

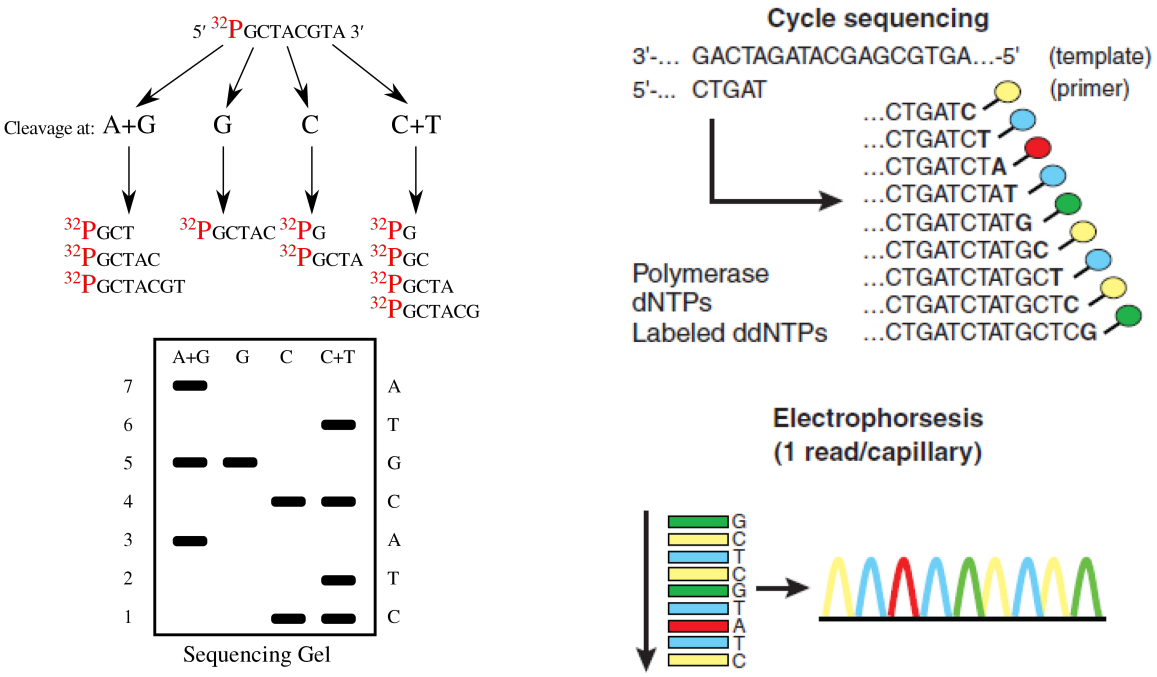


Figure 1.2: **First generation sequencing technologies:** (Left) Maxam-Gilbert sequencing, image reproduced from [141]. (Right) Sanger's method of sequencing by dideoxynucleotide chain termination, image reproduced with permission from [188].

The era of "first-generation" DNA sequencing began in the late 1970's with the discoveries of Allan Maxam and Walter Gilbert on the one hand, and Frederick Sanger on the other (Figure 1.2). Maxam-Gilbert sequencing, which has now fallen out of favor, treats ^{32}P 5' radiolabeled double-stranded DNA with different sets of chemical conditions that are specific for the cleavage of either: purines (A+G), guanine (G), pyrimidines (C+T), or cytosine (C). Cleaved products are run on a high resolution polyacrylamide gel for size separation and autoradiography reveals the nucleotide identities with the band furthest down the gel as the most 5' nucleotide on the input sequence. Sanger sequencing, which is still extremely popular for low-throughput applications, requires no hazardous chemicals or radioactivity but does require single-stranded input DNA as well as a primer. In its original incarnation there would be four distinct lanes on a gel, one for each dideoxynucleotide (A,C,G,T) reaction, although today dye terminators are more commonly used to generate a single capillary electrophoresis chromatogram encoding the sequence.

The era of "next-generation" sequencing has now firmly taken root, with many platforms competing on read length, read depth, speed, accuracy, and cost to name a few parameters. We highlight two specific next-generation technologies in Figure 1.3. Sequencing by synthesis is the most widely deployed method for modern high-throughput applications. Single stranded fragments of a DNA library are anchored to a glass slide (termed a flow cell) and are iteratively exposed to fluorescent nucleotides that compete for incorporation into a strand complementary to the template fragment. Optical recordings of color flashes along the flow cell determine the sequences for the millions of anchored fragments. The typical size of a sequencing read from Illumina machines implementing this method is 50-150 base pairs. A promising, though less developed, alternative is nanopore sequencing, in which single stranded DNA is threaded through a protein pore embedded in an electrically insulating membrane. A voltage difference across the membrane drives diffusion of the DNA through the pore, and the degree to which the nucleobases occlude ambient ionic current flow is used to sequence the strand. Nanopore sequencing can produce read lengths in the kilobases and provides sequencing data in real-time rather than at the termination of a long step-wise chemistry. We have only highlighted two of the many exciting platforms for modern-day high-throughput sequencing, but these two examples represent the ends of the technical spectrum in some sense. Illumina's sequencing by synthesis has become ubiquitous because the primary application area is medically relevant studies of human or model organism tissue. In such scenarios a reference genome is already known, mitigating the potential weakness of short sequencing reads via alignment of reads to the reference.

We take the time to familiarize the reader with the basics of DNA sequencing as a tool because it represents our most reliable read-out of evolutionary processes, a development as foundational to genetics as the invention of the microscope was to cytology. Pathogenesis and progression of cancer occur on long time scales compared with the lifetimes of individual malignant cells. Immune escape or vaccine resistance in viral populations occur on long time scales compared to the infectious cycles of individual virions. Genomes can be

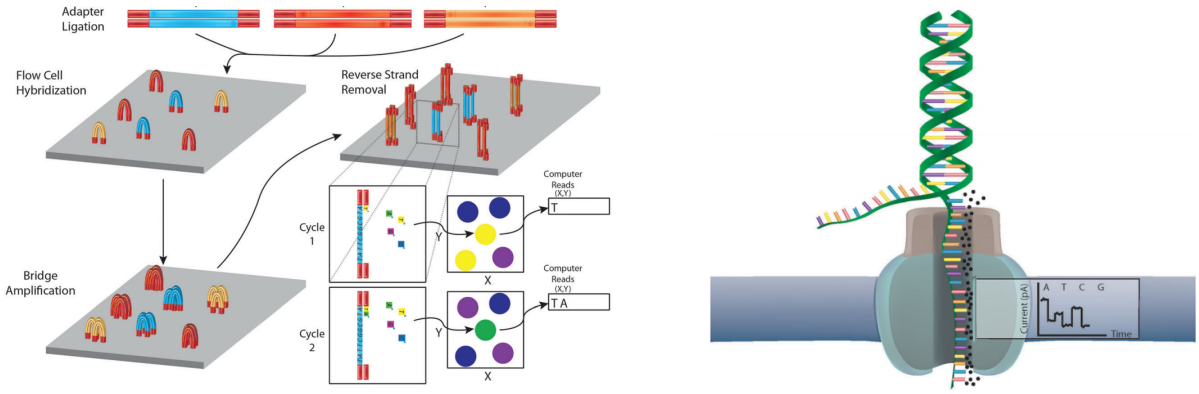


Figure 1.3: **Next generation sequencing technologies:** (Left) Reversible dye terminator based sequencing by synthesis, image reproduced with permission from [37]. Fragmented pieces of DNA are bound to, and amplified upon, a glass flow cell to form clusters of identical single-stranded fragments. A step-wise chemistry then incorporates fluorescent nucleotides into the synthesis of complementary strands, and recorded patterns of color flashes across the glass flow cell are converted into sequencing calls. (Right) α-Hemolysin based nanopore sequencing by voltage-driven diffusion, image reproduced with permission from [37]. Nucleobases occupying the narrowest section of the pore determine the degree of occlusion and thus resistance to ionic current flow.

highly dynamic in such settings of evolution under strong selection, and we would ideally like to sample the population densely in time. However, even with the falling cost and rising throughput of DNA sequencing, we are often limited to collecting single genetic cross sections in time. Part I of this thesis deals exclusively with cross-sectional sequencing data from cancer patient samples, implementing computational strategies for detecting recurrent alterations in the tumor transcriptome. Longitudinal sampling of cancer along the therapeutic time course is just now becoming feasible, thus in Part II we move from characterizing specific genetic alterations to characterizing patterns of genetic mutation across clinical time points. Influenza databases offer the gold standard in longitudinal genetic data, and Part II also illustrates a link between statistics of clonal evolutionary patterns in the viral hemagglutinin segment and the seasonal influenza vaccine effectiveness. Part III, which considers the clonal evolution of synthetic RNAs, we circumvent the common issue of haplotyping genetic variants detected via short read sequencing. The RNA genomes are designed to fit completely within a single Illumina read, thus allowing the depth of sequencing to deeply

probe the composition of the evolved populations.

1.3 Role of Computation in Interpreting Genetic Evolution

There are many ways to approach the role of computing in genetics (and biology in general). An immediate need certainly exists today for data analysis infrastructure, especially given the rate of adoption of high-throughput sequencing platforms. The maturing niche of bioinformatics seeks to provide new methods grounded in software engineering and applied statistics. And though it is tempting to stand in awe of the technical advances made in DNA sequencing, we should resist the urge to view this revolution as having achieved its ultimate goal. It is sobering to consider that today's technologies are still incapable of cheaply and exhaustively capturing all haplotypes present in populations of cellular organisms. The needs for computation in genetics often stem from inherent weaknesses of our measurement devices: short reads rather than long, bulk DNA rather than single cell, noise limitations on depth of sequencing, etc. Computational methods in the life sciences are often relied upon to provide robust inference as compensation for imperfect physical measurement.

To frame the role of computational biology in such limiting terms though, is to ignore the larger theoretical issues that persist even if our ability to read or write DNA were hypothetically perfected. Issues of predicting how sequences evolve against particular fitness landscapes, or understanding the time and length scales across which genetic alterations manifest, would not be trivially answered with further advances in measurement capability. Biology may be inundated with data in the present day, but statistical summaries of data should not be conflated with general model building [117]. Arguably the most important role for computational life scienstists is to investigate new mathematical ideas and techniques for parsimoniously modeling biological interactions.

1.4 Organization of the Thesis

This thesis is organized into three parts based on the type of high-throughput sequencing data analyzed and the particular biological setting being interrogated.

Part I deals with single-time-point (cross-sectional) observations of large sets of tumor transcriptomes, with the goal of detecting penetrant alterations such as gene fusion or oncovirus infection. Chapter 2 introduces the Pegasus pipeline, which casts the prediction of oncogenic gene fusions as a supervised learning task. The main results of Chapter 2 are the superior performance of Pegasus compared with existing tools and the successful detections of novel genomic translocations in peripheral T-cell lymphomas. The novel fusions we highlight are TRAF1-ALK in anaplastic large cell lymphoma and VAV1 with various fusion partners in angioimmunoblastic T-cell lymphoma or peripheral T-cell lymphoma not-otherwise-specified. Material presented in Chapter 2 is published, wholly or in part, in [1, 5]. Chapter 3 introduces the Pandora pipeline, which quantifies viruses present in human RNA-seq data and supports the exploration of genetic or microenvironmental host-viral interactions. The main results of Chapter 3 are the superior performance of Pandora compared with existing tools and the viral characterization of hematologic malignancies including Burkitt lymphoma and peripheral T-cell lymphoma. A novel EBV latency program is reported in endemic Burkitt lymphoma in addition to an anti-correlation between TCF3/ID3 mutation and EBV presence in Burkitt lymphoma more generally. In peripheral T-cell lymphoma Pandora detects a landscape of human herpesvirus infections and supports the likely B-cell origin of EBV in angioimmunoblastic T-cell lymphoma. Material presented in Chapter 3 is published, wholly or in part, in [4] and will be published in the manuscript "Viral landscape and microenvironment of peripheral T-cell lymphomas" (in preparation).

Part II considers multiple-time-point (longitudinal) observations of tumor and viral populations undergoing clonal evolution. Chapter 4 begins by reviewing much of the existing theory concerning phylogenetic inference and also describes the Billera-Holmes-Vogtmann (BHV) space of phylogenetic trees that forms the foundation for our work. The main results

of Chapter 4 are twofold. First, we introduce a projective BHV tree space along with an explicit implementation for visualizing point clouds of re-scaled trees. Second, we define a method for dimensionality reduction of large phylogenies, useful in analyzing long genomic time series, and prove its stability. Chapter 5 considers the particular longitudinal data sets derived from (i) cancer biopsies sequenced along the therapeutic time course, (ii) cancer xenograft models sequenced at different animal passages, and (iii) seasonal sequencing of H3N2 influenza hemagglutinin segments. These scenarios all represent clonal evolution under strong selection, and serve to illustrate the utility of our tree space methods. The main result of Chapter 5 is the inverse association between the variance of phylogenetic trees built from seasonal H3N2 hemagglutinin sequences and future influenza vaccine effectiveness. The material presented in Chapters 4 and 5 is published, wholly or in part, in [233, 234] and will be published in the manuscript "Phylogenetic dimensionality reduction" (in preparation).

Part III focuses on an engineering approach to clonal evolution, wherein large populations of RNAs are clonally evolved *in vitro* against a desired fitness objective. The fitness objectives we investigate both relate to altering ribosomal translation, first with programmed ribosomal frameshift and second with stop codon readthrough. Chapter 6 details the discovery of RNA pseudoknot compositions for high-efficiency induction of ribosomal frameshifting. The main result of Chapter 6 is the identification of an optimal frameshifter genotype in the evolved population of RNAs, and its use in constructing ligand-responsive riboswitches. The material presented in Chapter 6 is published, wholly or in part, in [9]. Chapter 7 takes an unbiased approach to evolving 3' RNA elements that induce eukaryotic stop codon readthrough. The main result of Chapter 7 is the development of a classifier of stop codon readthrough based on position-specific sequence information and structural features of predicted hairpins. The trained model is applied to the estimation of readthrough efficiency across all human 3'UTRs, with subsequent *in vivo* validation performed. The material presented in Chapter 7 will be published in the manuscript "Large scale profiling of stop codon readthrough RNA signals and their identification in human 3'-UTRs" (in preparation).

Part I

Cross Sectional View of Fitness Drivers in Cancer

Chapter 2

Gene Fusions as Penetrant Oncogenic Events

Gene fusions are the result of genetic aberrations (translocations, deletions, amplifications and inversions) involving the juxtaposition of two genes that can generate a single hybrid transcript. Since 1960 gene fusions have been known to play a major role in tumorgenesis. The BCR-ABL1 gene fusion, resulting in the Philadelphia chromosome (t(9;22)(q34;q11)), was the first case of a chromosomal translocation associated with the development of a cancer, namely chronic myelogenous leukemia [151]. In this fusion, the N-terminus oligomerization domain of BCR and the tyrosine kinase domain in ABL1 are essential in promoting oncogenic activity [236]. Among the gene fusions associated with tumor development, it is worth mentioning TMPRSS2-ERG, a gene fusion occurring in 40-80% of cases of prostate cancer [204, 38] and fusions involving ALK gene with different partners in various malignancies [34], such as NPM1-ALK in anaplastic large cell lymphoma (ALCL) [138] and ELM4-ALK in non-small-cell lung cancer [195].

Material presented in this chapter is published, wholly or in part, in: [1] in collaboration with F. Abate, E. Ficarra, A. Acquaviva, C.H. Wiggins, V. Frattini, A. Lasorella, A. Iavarone, G. Inghirami, R. Rabadan; [5] in collaboration with F. Abate, A.C. da Silva-Almeida, J. Robles-Valero, L. Couronne, H. Khiabanian, S.A. Quinn, M.Y. Kim, M.A. Laginestra, C.S. Kim, D. Fiore, G. Bhagat, M.A. Piris, E. Campo, I.S. Lossos, O.A. Bernard, G. Inghirami, S. Pileri, X.R. Bustelo, R. Rabadan, A.A. Ferrando, T. Palomero

Discovering the relationship between gene fusions and cancer is gaining significant momentum thanks to advances in next generation sequencing (NGS) technology, particularly RNA-seq paired-end data [131]. Recently, the application of this technology allowed the discovery of new chromosomal rearrangements of the CIITA gene with various promiscuous partners in the lymphomagenesis of primary mediastinal B cell lymphomas [197]. In Singh et al. [190], the analysis of RNA-seq paired-end data led to the discovery of the highly oncogenic fusion protein FGFR3-TACC3 in 3% of patients diagnosed with glioblastoma multiforme (GBM). Even though FGFR3-TACC3 occurs at low frequency, the efficacy of FGFR inhibitors in the treatment of these tumors opens the door to personalized therapies for this deadly disease. Moreover, the FGFR3-TACC3 fusion has been found in other cancers such as bladder [220] and lung [132]. These recent discoveries underscore the power of high throughput genomics for the identification of targetable gene fusions, opening the door to personalized cancer therapies.

Several bioinformatics tools have are now established for the detection of candidate fusion events. Most of the fusion detection tools exploit an algorithmic strategy based on RNA-seq paired-end data. Generally, the detection of read pairs that discordantly map to two distinct genes generates a first set of gene fusion candidates. Subsequently, the exact fusion junction is determined for each candidate by searching for reads spanning the breakpoint, i.e. reads that partially map to both genes. FusionSeq [183] and deFuse [134] are the earliest examples of software based on this strategy. Detection tools differ in the type and number of cascading filters they apply to reduce the large number of false positive fusions. Chimerascan [97] implements a variation of the algorithm based on trimming reads to increase the fusion detection sensitivity. Bellerophontes [2] uses TopHat [206] and Cufflinks [207] to identify gene fusions involving truly expressed genes, and applies a set of modular cascading filters based on an accurate gene fusion model [55]. A comprehensive comparison of fusion detection tools has recently been performed in [31].

The methods adopted by fusion detection tools to shrink the list of candidates lead to

increased specificity but reduced sensitivity. As reported in the comparative analysis performed by Abate *et al.* [2], because of the heterogeneity of the applied filters the resulting set of detected fusions poorly overlaps. To obtain a more complete list of relevant fusions, the union set of candidate fusions detected by all the detection tools should be considered for further experimental validation. A problem arises however, since the number of putative gene fusions might be on the order of hundreds of candidates per single RNA-seq sample. This is largely due to the presence of readthrough events, reverse transcriptase template switching artifacts, and different systematic errors in the analysis of the reads [159]. The naÃŕve approach of considering all candidates quickly overwhelms the capacity of experimental validation procedures, and highlights the need to focus on a reduced number of select biologically relevant fusions driving the oncogenic progression of disease.

Driver mutations are understood to be genetic alterations that confer a fitness advantage to malignant cells, and historically they have been identified simply by their high prevalence across cancer patients. Passenger mutations are alterations that do not have appreciable impact on the malignant phenotype. The classification of gene fusions into driver and passenger events is a complex problem that has not been fully explored yet. To address this issue, several databases have collected hundreds of chromosomal translocations involved in cancer cases and reported in the biomedical literature. For instance, Mitelman [136], TICdb [150], and ChimerDB2.0 [110] are manually curated repositories of known gene fusions along with detailed information such as chromosomal breakpoints, reported tissue types, and fusion sequences. New computational approaches to nominate biologically relevant fusions from high-throughput data have been proposed. ConSig assesses driver gene fusions by combining copy number variations (CNV), ontologies and interactomes based on the assumption that fusion events are more likely to arise from genes with similar biological functions [217]. Wu *et al.* have proposed a network based approach relying on relative co-occurrence of protein domains and domain-domain interactions, and location of the gene fusion in a gene network [225]. Recently, Oncofuse has improved the computational analysis with a machine

learning approach based on a Naive Bayes classifier applied to preserved domains after chromosomal rearrangement [189]. Compared to the previous methods, Oncofuse introduces a level of detail considering only the domains that are effectively maintained on the resulting gene fusions. The domain analysis should be extended, however, taking into account all the possible transcript isoforms as well as the reading frame, which plays a crucial role since frameshifted fusions imply a loss of the 3'-gene domains. Moreover, Oncofuse relies on a NaÃŕve Bayes classifier that makes a restrictive assumption on the class conditional independence of all features. Taking the FGFR3-TACC3 gene fusion as an example, however, the acquired coiled-coil domain of the TACC3 gene cooperates with tyrosine kinase functionality of FGFR3 to produce the dramatic oncogenic effect [190]. This example illustrates the limitations of a model assumption that ignores interactions between functional protein domains.

In this chapter we aim to discern oncogenic driver fusions from the background of passenger events and artifacts by combining 1) functional domain annotation based on accurate fusion sequence analysis and 2) a binary classification algorithm using gradient tree boosting. The implementation of this methodology is Pegasus, a new framework for the functional characterization of RNA-seq gene fusion candidates and quantification of their oncogenic potential. Pegasus runs on top of multiple state of the art fusion detection tools in order to maximize detection sensitivity and consider the largest possible set of fusion candidates.

The main innovative steps introduced by Pegasus are as follows: Common interface between several fusion detection tools; Chimeric transcript assembly: a key feature since fusion detection tools do not report whole transcript sequences; Reading frame identification and accurate domain annotation, including both preserved and lost protein domains within the assembled chimeric transcript; Prediction of fusion oncogenic potential: high performance ensemble learning technique trained on a feature space of protein domain annotations; Automated workflow that would otherwise require massive effort if manually executed by the scientist.

We assess the trained Pegasus model's prediction accuracy by applying it to a set of recently discovered gene fusions where it compares quite favorably with the current state of the art, Oncofuse. The term "prediction" in this context is understood to mean the mapping of input descriptors of a gene fusion event to a numeric score representing the likelihood of the event being a driver. Assessment of prediction accuracy is always performed on out-of-sample data points, and the data used for benchmarking with Oncofuse is, in fact, more recent than our training corpus, thus representing a true prediction into future data for Pegasus. Beyond curated datasets, we report the results of Pegasus on real RNA-seq data from two distinct patient cohorts: public GBM samples from TCGA and non-public ALCL samples. We successfully identify driver gene fusions in both cancers and demonstrate the utility of coupling our algorithm with experimental analysis in cancer research.

2.1 Pegasus: Chimeric Transcript Annotation and Scoring

In order to first motivate our feature engineering, we briefly review the main mechanisms hitherto identified in oncogenic gene fusions (Figure 2.1). Fusion transcripts can broadly lead to three scenarios: i) enhanced overexpression of an oncogene ii) deregulation of a tumor suppressor gene iii) formation of a new, aberrant protein. Enhanced overexpression of an oncogene is exemplified by the famous IgH-MYC fusion and is the main reason for our explicit annotation of oncogene status and interactions with known oncogenes in our feature space representation of fusion transcripts. In other cases, deregulating properties can be associated with the fused transcript, such as insertion of one or two nucleotides across the junction breakpoint introducing a shift of the reading frame. This scenario is illustrated in the PPP2RA-CHEK2 fusion [101] where the introduced frameshifted sequence prevents the formation of the CHEK2 protein that is a known tumor suppressor gene. Here we see the motivation for our explicit annotation of tumor suppressor status and interactions with

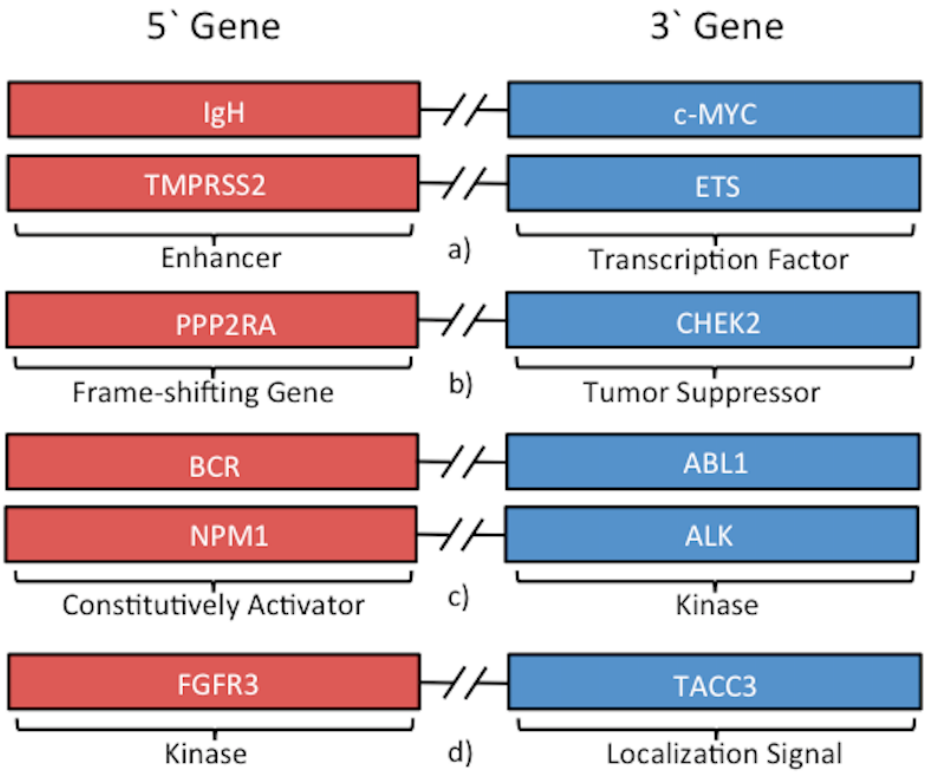


Figure 2.1: **Famous oncogenic gene fusions:** Different mechanisms of gene fusions in cancer are shown. (a) Shows the mechanisms of oncogenic transcription factor activation by means of an endogenous enhancer. (b) Shows an example of disrupting gene fusion where the 5' gene leads to the deregulation of the 3' tumor suppressor gene. Finally, (c) and (d) depict the BCR-ABL1, NPM1-ALK, and FGFR3-TACC3 chimeras where a completely new protein is produced.

known tumor suppressors in the feature space, as well as the need for computing reading frame of each candidate fusion. Finally, fusion transcripts can also yield a completely new chimeric protein. BCR-ABL1 [151] and NPM1-ALK [138] are well studied examples of such in-frame fusions. The new protein is generally larger than the kinase involved and causes an increase of the tyrosine kinase activity. Moreover, in the recently discovered FGFR3-TACC3 gene fusion, the acquired coiled-coil domain of TACC3 gene drives the localization of the fusion protein to the mitotic spindle through a mechanism that is dependent on tyrosine kinase functionality [190]. It seems that a reasonable feature space representation for predicting the oncogenic properties of such novel chimeric proteins should maintain knowledge of both preserved and lost functional domains in the partner genes.

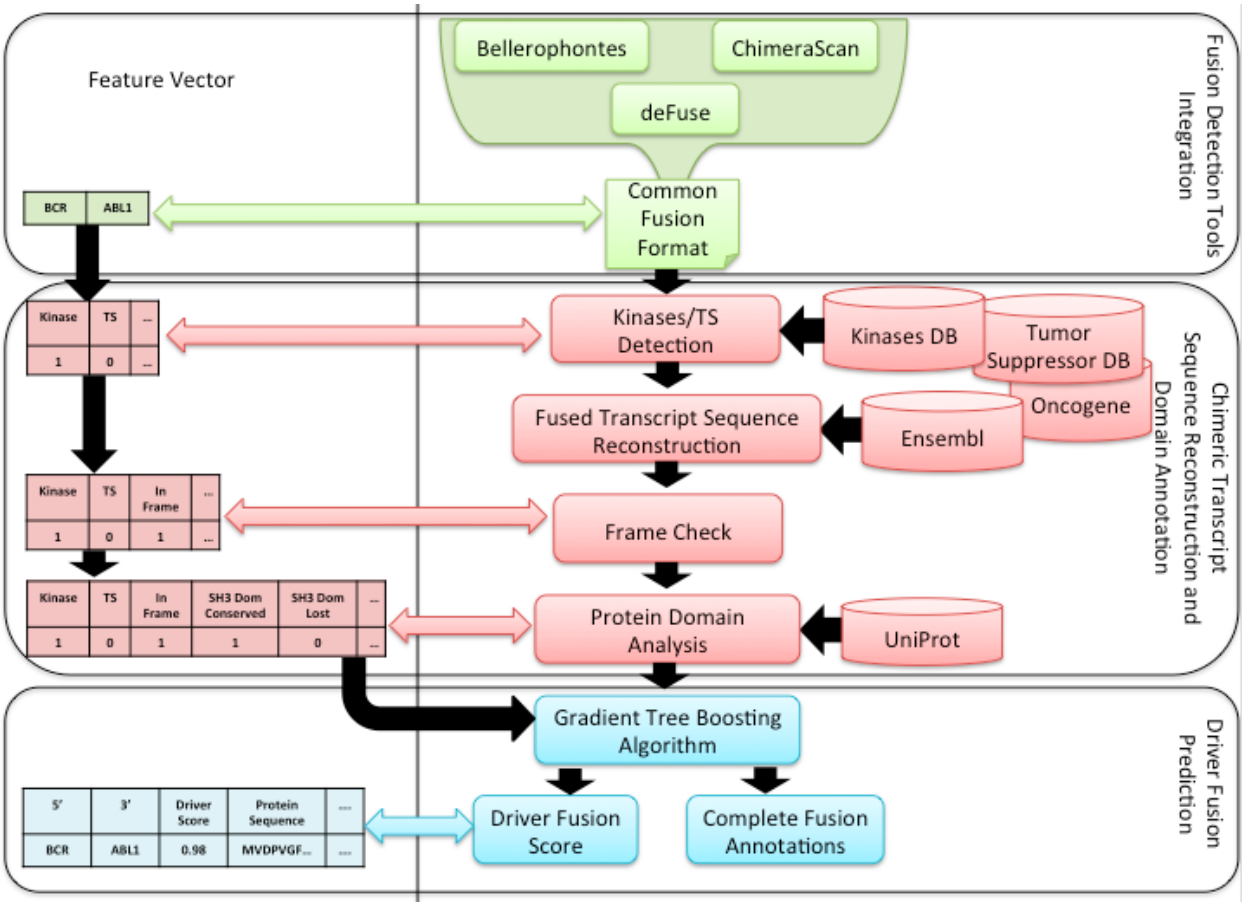


Figure 2.2: **Pegasus pipeline architecture:** For each phase, the figure shows how the feature vector is constructed on the left side of the panel. In Fusion Detection Tools Candidates Integration step, report files from several fusion detection tools are loaded in a unique fusion database. In Chimeric Transcript Sequence Reconstruction and Functional Analysis phase, the fusion transcript is assembled according to the fusion breakpoint coordinates, the reading frame is checked and the protein domain annotation is performed on the resulting fused sequence. Finally, the Driver Fusion Prediction applies machine learning techniques to determine prediction scores.

The methodology of Pegasus is composed of three phases (Figure 2.2): a) integration of candidates from fusion detection tools b) chimeric transcript sequence reconstruction and domain annotation c) classifier training and driver prediction. The first phase, Fusion Detection Tools Integration, involves pooling the entire set of unique gene fusion candidates detected by any of the fusion detection tools. The second phase, Chimeric Transcript Sequence Reconstruction and Domain Annotation, includes two steps: i) reconstruction of the chimeric transcript using the genomic breakpoint coordinates and the partner gene annotations ii) annotation of the assembled sequence to provide information on the fusion frame and to generate a report of all the protein domains conserved or lost in the gene fusion. The final phase frames Driver Fusion Prediction as a binary classification task and fits an ensemble of decision tress via the gradient boosting algorithm.

The Fusion Detection Tools Integration is the repository of the entire set of fusion candidates detected by any of the fusion detection tools. Several fusion detection algorithms are supported in Pegasus: Bellerophontes, deFuse, and ChimeraScan. Each tool adopts a private formalism for reporting fusion information with different levels of detail. However, some chimeric fusion features are common to all the fusion detection tool reports (e.g. genes involved in the fusion, genomic breakpoint coordinates, number of reads encompassing and spanning the fusion breakpoint, etc.). Thus, the internal database structure of Pegasus provides a unique point of access for all the information needed to fully describe a gene fusion candidate. Furthermore, experimental analysis might involve the comparison of several RNA-seq samples per case study. To this end, the common repository embedded in Pegasus provides an organized overview of all the fusions occurring in the entire sample set. This feature allows comparison and the recurrence analysis of the fusion candidates within both the same experimental dataset (samples of the same disease) and within different experimental datasets (samples across different diseases).

For each gene fusion candidate, the entire chimeric transcript sequence is first assembled according to publicly available gene annotations and the fused gene breakpoint coordinates.

This is the most computationally intensive step in the methodology. For each gene fusion candidate, Pegasus assembles the chimeric sequence based on the possible isoforms and splicing junctions of each gene, as well as the genomic breakpoint coordinates (Figure 2.3). It is worth specifying that Pegasus reconstructs the fusion sequence exclusively on the basis of gene annotation and fusion breakpoint, and it does not exploit the sequenced reads because they are not an input to the program. Therefore the reconstructed sequence may not reflect the actual sequences especially in case of alternative splicing events. We adopt the annotation file from ENSEMBL database [69]. Since several distinct isoforms might be available for a specific gene, Pegasus considers the combinations of all possible isoforms reported in the annotations of those genes involved in the fusion (Figure 2.3). The chimeric transcript sequence is therefore reconstructed combining the 5' gene isoform sequence (from the isoform start codon to the genomic breakpoint) and the 3' gene isoform sequence (from the genomic breakpoint to the isoform stop codon). Different gene isoforms allow for different protein domains to be retained or disrupted during the fusion. If this scenario occurs, Pegasus considers the union of all possible domains that are retained and lost and as input features for downstream classification. Furthermore, the fusion breakpoint can fall in either the coding region (exon-exon junction boundaries), or in non-coding regions (exon-intron or intron-intron junction boundaries). Pegasus takes the latter scenario into account and if the fusion breakpoint falls in an intron, the intronic sequence is retained.

After sequence reconstruction we assess the preservation of the reading frame in the chimeric transcript, which enables a great deal of our downstream feature engineering (Figure 2.4). If the gene fusion introduces or deletes a nucleotide in one of the codons, the entire reading frame is shifted and the corresponding amino acid sequence changes. Consequently, the resulting protein sequence is different from the one encoded by the gene involved in the fusion. The gene fusion encodes a protein sequence that either corresponds to a completely unknown protein or contains a premature stop codon (the presence of a premature stop codon in the chimeric sequence interrupts the protein translation resulting in the truncation

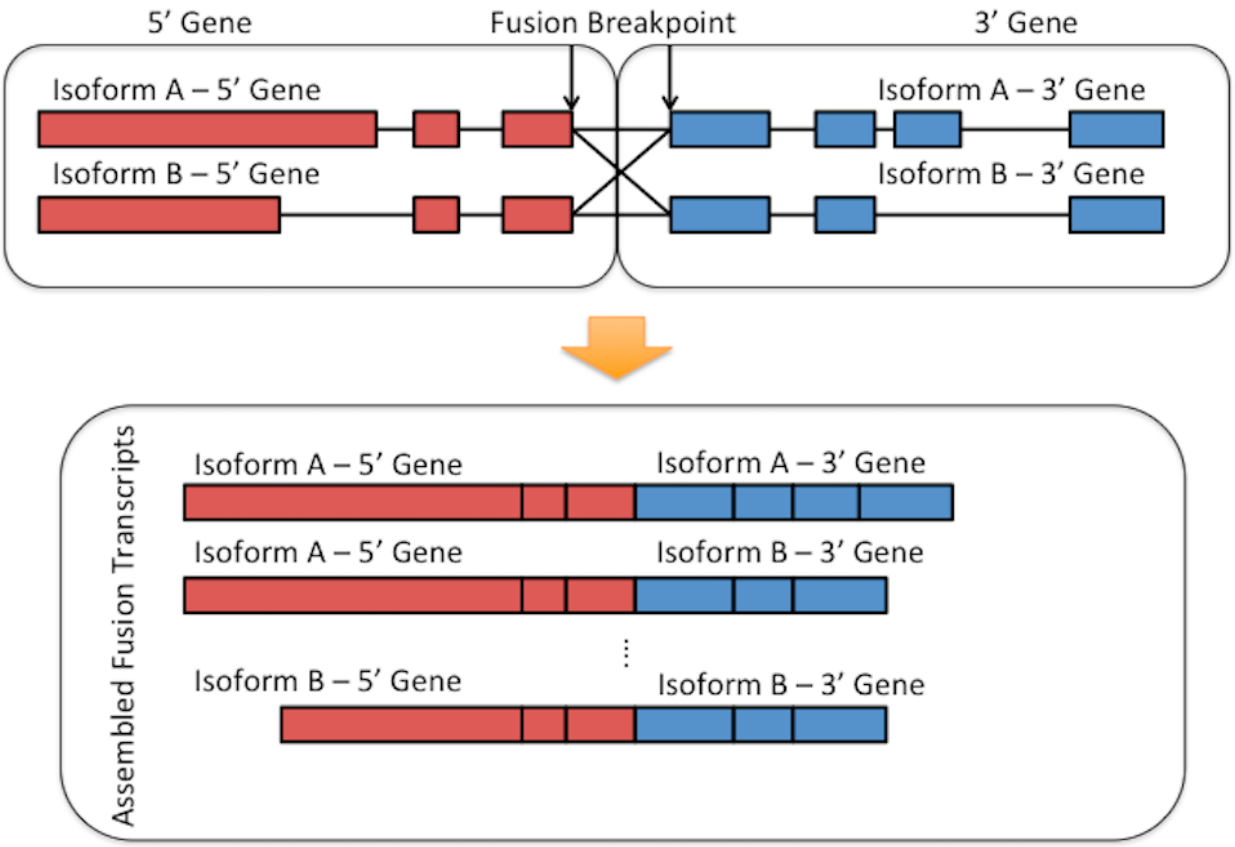


Figure 2.3: **Chimeric transcript reconstruction:** Both for the 5' and the 3' gene the annotated isoform sequences are retrieved. All the possible combinations between the isoforms of the 5' and 3' genes are considered as putative fusion transcripts.

of the protein encoded by the 5' fused gene). This class of mutations is functionally similar to nonsense point mutations that play a role in many cancers and might imply the loss of functionality of the 5' fused gene. The sequence is labeled as in-frame if the number of nucleotides composing the fusion sequenced is a multiple of three and no premature stop codons are introduced in the chimeric sequence.

The annotation of the preserved and lost protein domains is essential in order to capture the oncogenic potential of a translated chimeric transcript. The nucleotide fusion sequence assembled in the previous step is translated into an amino acid sequence. Subsequently, the UniProt web service [39] is queried for all available annotations of the putative protein encoded by the two genes involved in the fusion (Figure 2.4). Leveraging the reading frame information and fusion breakpoint, Pegasus determines the conserved and lost domains as-

sociated with both 5' and 3' genes. It is worth emphasizing that both conserved and the lost domains are valuable features of a fusion transcript, with the former more likely to discriminate oncogene related fusions and the latter more likely to discriminate tumor suppressor related fusions.

The domain annotation permits the creation of a detailed feature space for the fusion transcripts, a pre-requisite step for posing the ensuing machine learning task. In Pegasus the feature space is composed of:

- Binary information about reading frame and breakpoint region (if the breakpoint falls in coding regions, introns and UTRs).
- Presence or absence of ∼1000 protein domains from UniProt. Our selection was based on the domains occurring in the training set from ChimerDB2.0.
- Number of oncogenic or tumor suppressor domains, as defined by association with the keywords "tumor suppressor" or "oncogene" in the UniProt database.
- Number of protein-protein oncogenic interacting domains. We check if one or more domains of the fusion interact with both oncogenic and tumor suppressor domains.

2.2 Boosting Framework for Supervised Learning

We aim to fit a model that can identify oncogenic fusions from the background of passenger events and artifacts. More precisely, we aim to learn a mapping $f : X \mapsto y$ from the fusion transcript feature space X to a label $y \in \{0, 1\}$ representing oncogenic driver status. Since we desire a biologically interpretable model that is also capable of capturing interactions between features, the decision tree is a natural choice. On the other hand, high dimensional feature spaces predispose to overfitting, and previous driver fusion prediction studies [189] focused a great deal on upfront dimensionality reduction for this reason. A single, large decision tree classifier is not likely to generalize beyond some training depth. An ensemble of shallower decision trees, if learned in a boosting framework, can guard against overfitting

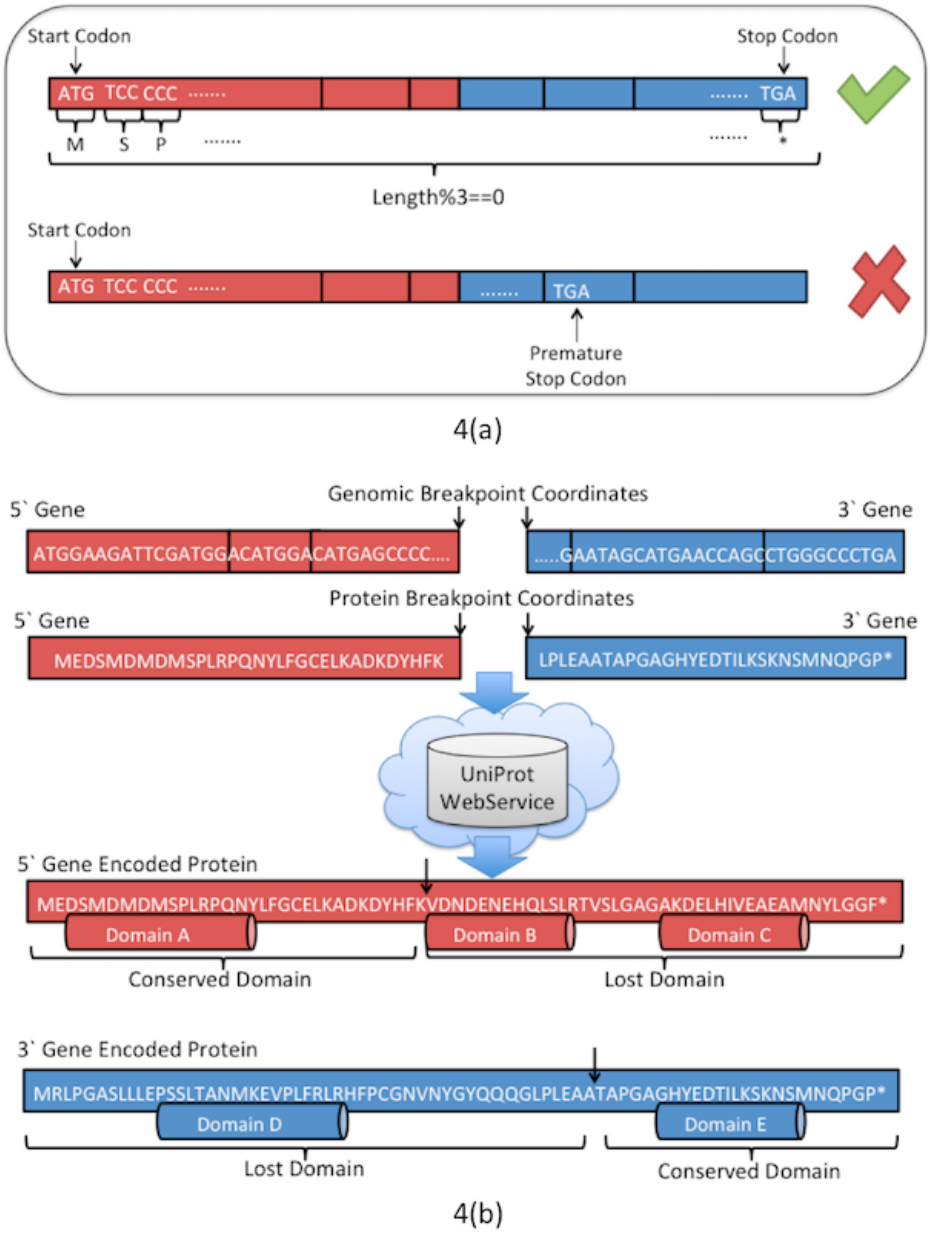


Figure 2.4: **Chimeric frame and domain annotation:** (a) The length of the fusion transcript, from the start to the stop codon, must be multiple of three (three nucleotides per single encoded codon). If the length of the sequence modulo 3 is non-zero, the fusion sequence is frameshifted. A premature stop codon can be introduced in the protein sequence. (b) The nucleotide sequence resulting from the fusion of the 5' and 3' gene is translated into amino acid sequence. Similarly, the genomic breakpoint coordinates are translated into protein amino acid coordinates. UniProt Web Service is queried and the list of the available domains for both the gene is retrieved. On the basis of the protein domain sequence and protein breakpoint, the list of both conserved and lost domains is reported.

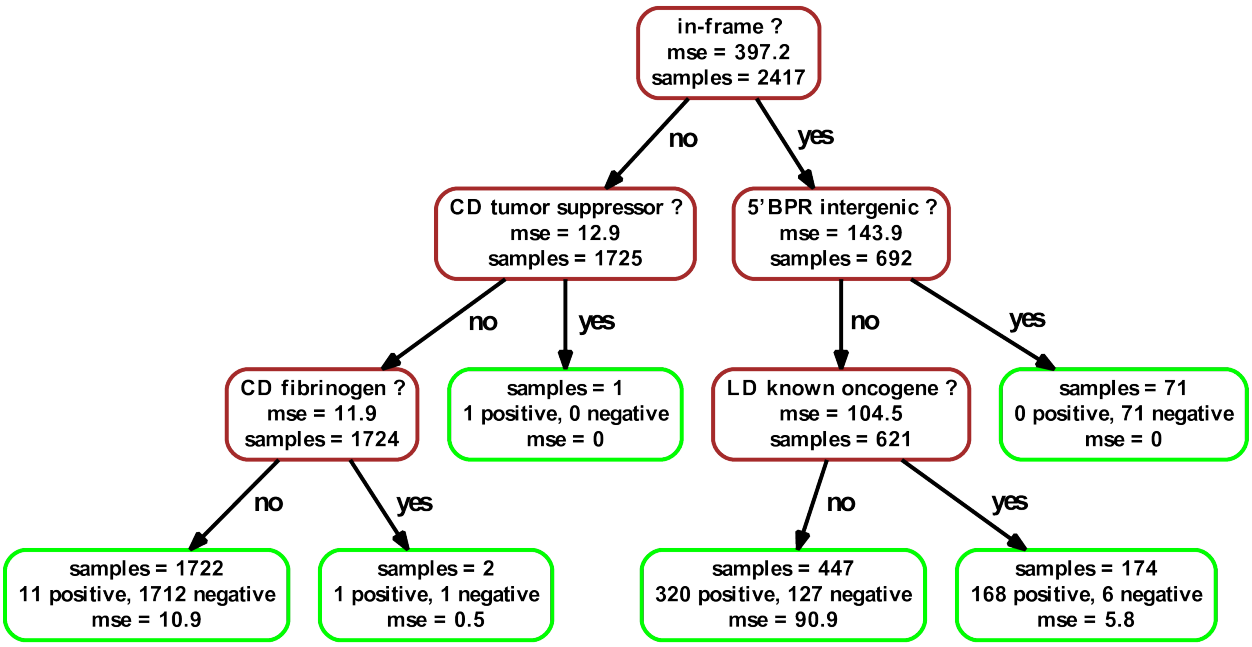


Figure 2.5: **Decision tree base learner:** The base learner in the boosted classification model is the decision tree, which defines a recursive partitioning of the feature space into disjoint regions each modeled by a constant. Decision nodes are colored brown while leaf nodes are colored green. At the leaf nodes, the samples have been partitioned into disjoint sets with a higher degree of label homogeneity than at the root.

because of the iterative nature of the learning and the additive structure of the model [75]. Therefore the balance we strike between the expressive power of decision trees and robustness to overfitting comes in the form of stagewise additive modeling. In Pegasus we employ an additive model $f(x) = \sum_m \alpha_m h_m(x)$ composed of weighted decision trees, an instance of which is shown in Figure 2.5, and fit via gradient boosting [73]. There is no manual reduction of features or feature space dimension in this strategy, unlike the manual selection of 6 enriched functional categories in the Oncofuse framework [189].

Thus we require neither upfront dimensionality reduction schemes nor the restrictive assumption of class conditional feature independence in the NaiÌĹve Bayes model. Gradient tree boosting is an ensemble learning technique, wherein decision trees are used as base learners and the final model is expressed as an expansion in these basis functions. Figure 2.5 depicts a sample regression tree that would be added to the ensemble in a single round of boosting. Although the base learners are performing regression, appropriate choice of loss

Algorithm 1 Gradient Tree Boosting Algorithm.

$f_0(x) \leftarrow arg\,min_\gamma \mathcal{L}(y_i, \gamma)$
for $m = 1, 2, \ldots, M$ **do**
    for $i = 1, 2, \ldots, N$ **do**
        Compute $r_{im} = -[\frac{\partial \mathcal{L}(y_i, f(x_i))}{\partial f(x_i)}]_{f=f_{m-1}}$
    end for
    Fit a regression tree to the targets r_{im} giving terminal regions R_{jm}, with $j = 1, 2, \ldots, J_m$
    for $j = 1, 2, \ldots, J_m$ **do**
        $\gamma_{jm} = arg\,min_\gamma \sum_{x_i \in R_{jm}} \mathcal{L}(y_i, f_{m-1}(x_i) + \gamma)$
    end for
    Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} \mathbb{1}(x \in R_{jm})$
end for
return $\hat{f}(x) = f_M(x)$

function for gradient boosting yields a classification task. Here we use the binomial deviance loss, and enforce a maximum depth of 5 nodes in the individual decision trees.

Adaptive boosting, first proposed by Freund and Schapire in the 1990s, has gained recognition as one of the most powerful algorithms in modern machine learning. More recent work by Friedman in 2000 generalized the adaboost algorithm to arbitrary differentiable loss functions and established gradient boosting. The gradient tree boosting algorithm [74], originally published in 2000, is outlined below (Algorithm 1) and the implementation we use can be found in the scikit-learn python library [163]. We denote the number of training examples by N, the number of boosting rounds by M, and the loss function by $\mathcal{L}$:

2.3 Model Training and Validation Performance

This section highlights the performance of Pegasus in detecting driver gene fusions. First, we examine the performance of the classifier on the training data and compare its effectiveness to a recently published tool, Oncofuse, on a separately curated validation dataset. Next, we run Pegasus on two experimental datasets and demonstrate its role in reducing the search space of potential oncogenic drivers by accurately ranking fusion transcripts from a vast

set of putative candidates. The first is the publicly available RNA-seq data of GBM from TCGA.The second is a non-public set of 23 RNA-seq samples from a cohort of patients with ALCL, with 2 out of the 23 samples reporting the NPM1-ALK fusion. We analyze these datasets with ChimeraScan or deFuse and apply Pegasus to the entire set of detected fusions. It is worth specifying that in the reported results about chimeric transcript annotations, if two or more fusions share the same junction breakpoint coordinates, they are counted as a single fusion. The rationale is that according to the Pegasus fusion domain analysis, if two genes fuse in different samples with the same breakpoint they also share exactly the same domain. Conversely, if two genes occur in different samples with different junction breakpoint coordinates, the domain analysis accordingly changes.

The corpus of labeled data used to train the classifier comes from two sources. Positive examples, meaning true oncogenic driver fusions, are drawn from ChimerDB2.0, which contains 501 curated driver fusions. 1500 negative examples are then drawn from an internal collection of reactive lymph node tissue in patients with no clinical history of malignancy. The negative examples contain passenger fusions as well as readthrough transcripts. We also supplement the negative training data with 416 deliberately frameshifted transcripts from ChimerDB2.0 such that the necessary driver domains are lost. In total there are 501 positive examples and 1916 negative examples in the training corpus. The rationale for augmenting the negative set with 416 frameshifted fusions from ChimerDB2.0 is to include the scenario of chimeric transcripts containing an oncogene at the 3' position that is frameshifted. Since such events occur at low frequency in normal lymph node tissue, this design choice improves the performance of the classifier. In summary, the 501 fusions from ChimerDB2.0 form the positive training set and provide mostly in-frame fusions involving oncogenes. The 1500 fusions from normal tissue contribute to the negative set and provide both in-frame and frameshifted fusions. The 416 deliberately frameshifted fusions from ChimerDB2.0 complete the negative set and provide frameshifted gene fusions mostly with an oncogene at the 3' position.

The classifier is trained for 100 rounds of boosting under 10-fold stratified cross validation (CV) and achieves a mean test split AUROC of 0.96 and an average precision of 0.91. As expected, in Figure 2.6 the loss on the train split monotonically decreases with increasing model complexity, while we see no sign of overfitting in the form of rising loss on the test split. As the binomial deviance falls most sharply in the early rounds of boosting, we specifically quantified the gain in precision achieved by more complex models in comparison to classifiers with shallow weak rules (decision trees) or small ensembles of decision trees. The marginal benefit to classifier precision when augmenting the depth of decision trees or the number of boosting rounds is illustrated in Figure 2.7, where we observe saturating performance after roughly 20 boosting rounds in accordance with Figure 2.6. Since each decision tree base learner implies a hierarchy of informative features, we can average over the boosting rounds to produce an aggregate view of the most important features in the classification task. Specifically, the relative feature importances in Figure 2.6 are computed via an ensemble average of the single decision tree feature importances as defined in Breiman *et al.* [20]. Let M denote the number of boosting rounds. Let I_k^2 denote the squared importance of feature k. Let T denote a decision tree with $J-1$ pairs $(t, \hat{i}_t^2)$ of (internal decision node, estimated improvement in squared risk). The squared importance of a feature within a boosting round can be defined as

$$I_k^2(T) = \sum_{t=1}^{J-1} \hat{i}_t^2 \mathbb{1}(t = k), \tag{2.1}$$

and therefore the squared feature importance over the ensemble of base learners is

$$I_k^2 = \frac{1}{M} \sum_{m=1}^{M} I_k^2(T_m). \tag{2.2}$$

We observe that the computationally expensive step of computing the fusion transcript reading frame is justified in the eyes of the classifier, as it is the single most informative feature. Looking a little further down the list we learn other transcript features that are highly informative of driver events, such as having breakpoints in the CDS and conserving domains shared with or interacting with known oncogenes. Despite strong performance

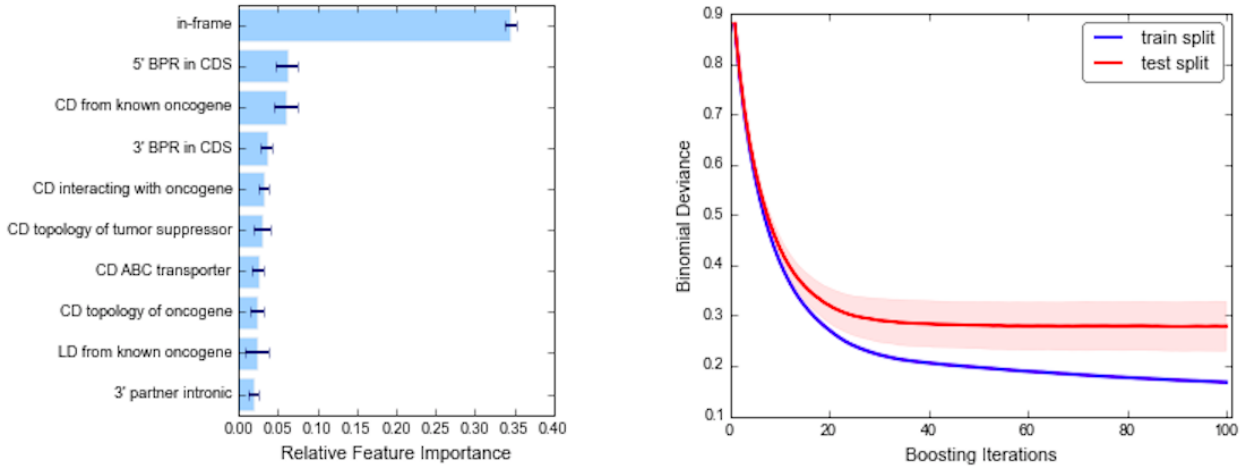


Figure 2.6: **Classifier training under stratified 10-fold cross validation:** Because of the imbalance in positive/negative training examples we stratify on the label and ensure that each 10% of data held out as a test set contains equal proportions of both labels. (a) Relative feature importances averaged over the 10 CV folds. (b) Binomial deviance loss function plotted against model complexity for both the train and test splits of the data.

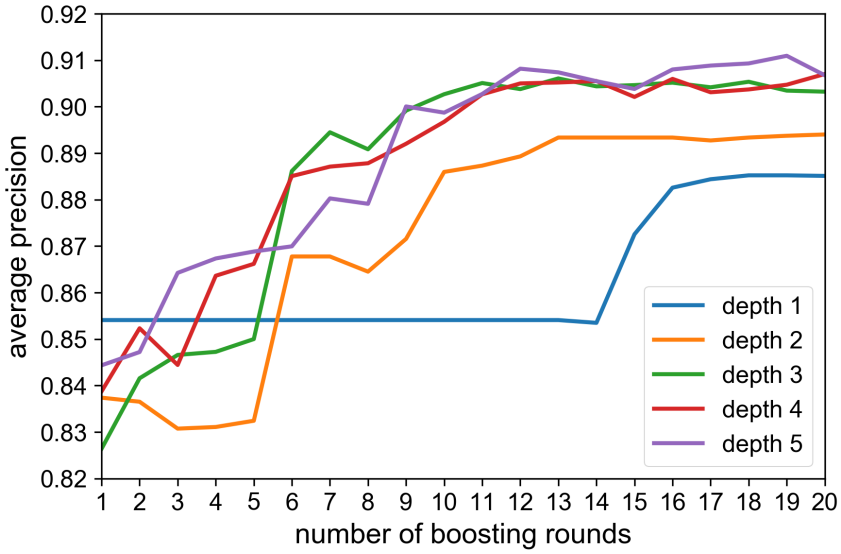


Figure 2.7: **Classifier precision as a function of model complexity:** Average precision is calculated as a function of number of boosting rounds and decision tree (our weak rules) depth. We observe saturating performance after roughly 20 boosting rounds, and our choice of depth 5 trees for Pegasus appears superior to models composed of simpler weak rules (shallower decision trees, decision stumps).

of the model under 10-fold CV on the training corpus, we are interested to see whether the classifier can generalize to new fusion transcripts that were unseen during the training phase. A list of 39 driver fusions, the majority of which are more recent than ChimerDB2.0, and corresponding to the validation set used in [189], is adopted as the positive validation set examples. To balance the label frequencies, we also select 39 transcripts from benign, reactive lymph node tissue as the negative validation set examples. None of the 78 validation examples are included in the training data. The negative examples are selected to contain at least one oncogene or tumor suppressor domain with the rationale that such transcripts more closely resemble driver fusions and would be most challenging for a classification function. In Figure 2.8 we demonstrate the favorable performance of the trained Pegasus classifier versus Oncofuse, the current state of the art in data-driven prediction of driver fusions. Since the ROC curve does not necessarily reveal how well separated the Pegasus scores are for the two class labels, we include the boxplot of Figure 2.8 to illustrate the remarkable resolution the classifier achieves between positive and negative examples. We also verify that Pegasus outperforms Oncofuse on randomly drawn sets of 39 non-oncogenic transcripts, though by a smaller margin. This is to be expected because the majority of non-oncogenic fusions are very easily classified, whereas our curated subset represents a more challenging task. Such robust performance on manually curated data sets naturally leads to the next proving ground, applying Pegasus to the enormous candidate lists generated from real RNA-seq samples.

2.4 Applications in Glioma

In order to demonstrate the effectiveness of Pegasus in predicting driver fusions, we analyze 15 samples of short-term glioblastoma stem cells freshly isolated from individuals with GBM. RNA-seq samples were first analyzed with ChimeraScan [97] and deFuse[134] for fusion detection. Next, we apply Pegasus to the set of gene fusion candidates and consider as

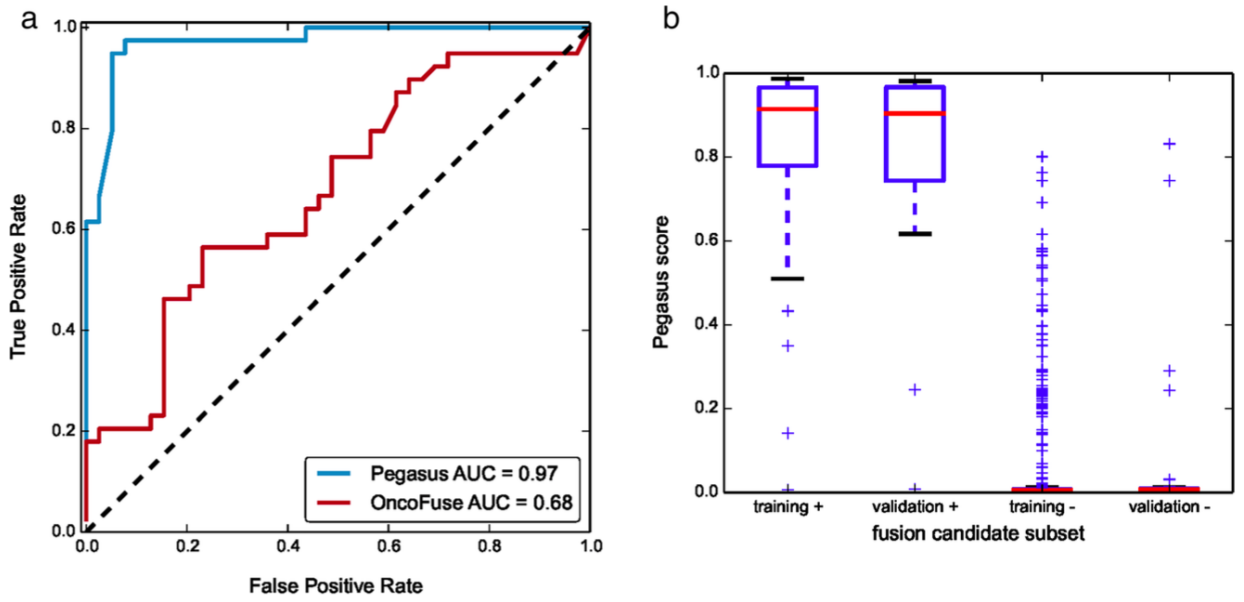


Figure 2.8: **Classification performance of trained model:** A curated set of 39 recently reported fusions that were not present in the ChimerDB2.0 training corpus are now used as an independent validation set. (a) Superior classification performance of Pegasus compared to Oncofuse on the validation dataset. (b) Boxplot demonstrating the high resolution of the Pegasus driver score in discerning the class boundary. Positive examples are consistently scored near 1 and negative examples consistently near 0.

driver events all those fusions having a number of supporting reads greater than 10 and a Pegasus Driver Score (PDS) greater than 0.8. As shown in Figure 2.8, a threshold of PDS > 0.8 promises a good trade-off between specificity and sensitivity. Table 2.1 reports the 4 detected driver fusions. All fusions have been validated with RT-PCR yielding a 100% rate of transcript validation. And while recurrence is often the surrogate measure of functional importance, the four unique driver candidates from this 15 patient sample still contain features associated with oncogenic gene fusions. In fact, CAND1-EGFR has been reported in [71] and the EGFR gene has been demonstrated to have an oncogenic role in GBM. Moreover, fusions involving MAPK1 and VOPP1 are reported as frequent in GBM with different gene partners [71, 21]. These results show that Pegasus can successfully detect relevant driver fusion candidates from RNA-seq data and that a threshold of PDS>0.8 and number of supporting reads greater than 10 provide a strong transcript validation rate.

As the most common and deadly primary brain cancer, GBM has recently undergone a

Table 2.1: **Pegasus predictions on 15 private GBM cases:** Pegasus top driver scores (PDS > 0.8) indicate 4 new driver fusions in GBM data.

5' Gene Partner	3' Gene Partner	Spanning Reads	Split Reads	Pegasus Driver Score	Validated
CAND1	EGFR	17	14	0.9437	Yes
MAPK1	FAM119B	145	96	0.9426	Yes
ADCK4	NUMBL	11	4	0.9426	Yes
VOPP1	IL22	48	35	0.8243	Yes

Table 2.2: **Pegasus predictions on 161 public GBM cases:** Pegasus top driver scores (PDS > 0.8) predicts 46% of known driver gene fusions in GBM data from TCGA cohort. Recurrence is assessed on the basis of gene fusion frequency reported in [71, 21].

5' Gene Partner	3' Gene Partner	Pegasus Driver Score	Recurrence
YEATS4	XRCC6BP1	0.9598	1
EIF4H	GTF2I	0.9440	1
ASH1L	C1orf61	0.9256	1
SEC61G	EGFR	0.9234	4
BCAN	NTRK1	0.9182	1
EGFR	VOPP1	0.9130	2
EGFR	SEPT14	0.9042	6
TDRD3	ESD	0.8959	1
TFG	GPR128	0.8880	4
PPP2R2B	CCT3	0.8699	1
TBC1D14	HTRA3	0.8697	1
FGFR3	TACC3	0.8442	3
LANCL2	SEPT14	0.8428	3

deep investigation by the multi-institutional consortium, the cancer genome atlas (TCGA). TCGA makes its collected RNA-seq data available to the larger scientific community, and we analyze a set of 161 samples from their GBM cohort. We first analyze the 161 RNA-seq samples with ChimeraScan (default parameters) [97], detecting a total of 9349 unique fusions across the entire dataset. Next, we apply Pegasus to the set of candidates and consider as driver events all fusion transcripts having a number of supporting reads greater than 10, Pegasus Driver Score (PDS) greater than 0.8 and recurrence greater than 1. As shown in the non-public RNA-seq data, these filtering thresholds provide a good validation rate by

RT-PCR. The application of these filters reduces the original list of 9349 candidates down to 13 high confidence fusions, making further functional analysis and validation tractable. Pegasus computes a score greater than 0.8 for both FGFR3-TACC3 and EGFR-SEPT14 gene fusions, which are already reported as driver translocation events in GBM [190, 71]. However, since TCGA biological material is not available, we are unable to perform further functional analysis of all predicted driver fusions with experimental procedures. Nonetheless, in order to validate Pegasus performance we compare PDS values with the frequencies reported in both [71] and [21] (Table 2.2). Of the 13 high confidence Pegasus predictions, 6 are recurrent in TCGA data suggesting a potential functional driver role in GBM tumorgenesis [198]. Some of the recurrent fusions involve the EGFR gene that is usually amplified in GBM, where it is known to activate STAT3 signaling and is thus a drug target. Particularly interesting is also the BCAN-NTRK1 gene fusion. In fact, NTRK1 is often translocated with different partners in cancers beyond just GBM [71].

2.5 Applications in Peripheral T-cell Lymphoma

Anaplastic large cell lymphoma (ALCL) is a form of peripheral T-cell lymphoma that is often associated with translocations of the ALK gene. In 23 non-public ALCL samples (∼450 million properly mapped reads) we detect a total of 5201 candidate fusion transcripts by means of deFuse [134] and ChimeraScan [97]. Beyond the two NPM1-ALK fusion transcripts (PDS = 0.98) that are already reported, Pegasus properly annotates and reveals 16 new biologically relevant fusions in these 23 samples. All 16 candidate driver fusions have been validated with RT-PCR, and 4 gene fusions have successfully undergone functional assays and in vivo validation. An example of Pegasus's effectiveness in functional domain analysis lies in the oncogenic role of TRAF1-ALK [61], a novel fusion in ALCL that Pegasus reports as driver.

The TRAF1-ALK fusion has been reported in three cases of ALCL (one in [61] and two

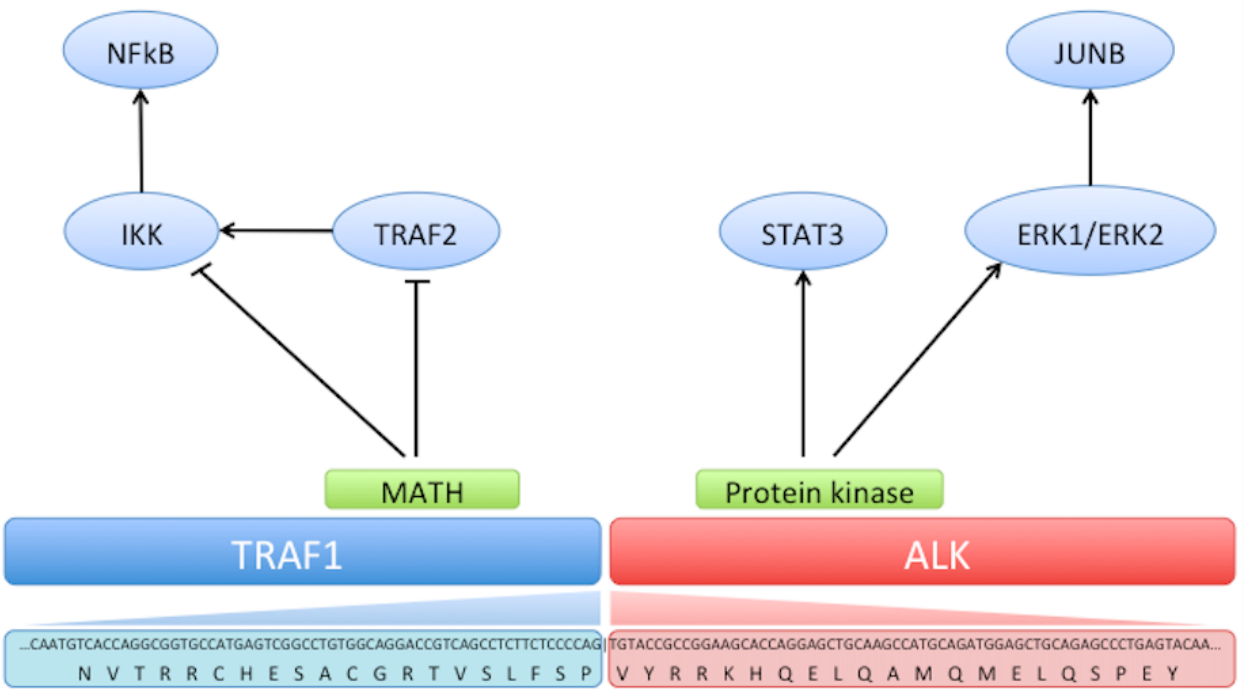


Figure 2.9: **Novel driver TRAF1-ALK gene fusion in ALCL:** A graphical representation of the Pegasus annotation on the TRAF1-ALK fusion in ALCL. Conserved domains are reported according to the junction breakpoint.

in [3]) suggesting a driver role. Pegasus accurately assembles and annotates the in-frame fusion sequence and correctly detects that the ALK protein kinase domain is completely conserved. Various fusions involving the ALK gene have been reported in literature and the oncogenic effect is generally promoted by ALK signaling. Interestingly, TRAF1 is also known to be involved in both the canonical and non-canonical NFkB pathway. Pegasus correctly annotates that the meprin and TRAF-C homology (MATH) domain is conserved, a domain that ubiquinates the IKK complex, activating NFkB transcription factors. As depicted in Figure 2.9, Pegasus properly identifies both the presence of an oncogenic protein domain (ALK) and an interacting oncogenic domain (TRAF1 to NFKB activation). Experimental work demonstrates and validates this Pegasus prediction showing the oncogenic effect of TRAF1-ALK *in vivo*, with activation of both ALK and NFkB signaling [3].

Beyond ALCL is a histopathologically diverse series of T-cell lymphomas, including angioimmunoblastic lymphoma (AITL), peripheral T-cell lymphoma not otherwise specified

(PTCL-NOS), natural killer / T-cell lymphoma (NKTCL), and cutaneous T-cell lymphoma (CTCL). We recently performed a systematic analysis of genetic alterations using RNA-seq data from a cohort of 154 PTCL samples, including 41 PTCL-NOS, 60 AITL, 17 NKTCL, and 36 ALCL tumors [160, 228, 41]. Pegasus successfully led to the identification and functional characterization of recurrent activating fusions involving the VAV1 gene specifically in AITL and PTCL-NOS subtypes. We identified three different fusion transcripts encoding proteins in CH which the C-terminal SH3 domain of VAV1 is replaced by the calycin-like domain of THAP4 (in two cases), the SH3 domain of MYO1F, or the EF domains of S100A7 (Figure 2.10). Reverse-transcription PCR amplification and DNA sequencing validated the expression of each of these VAV1 chimeric mRNAs in all samples analyzed.

The VAV1 protooncogene encodes a guanine nucleotide exchange factor (GEF) and adaptor protein with crucial signaling roles in protein tyrosine kinase-regulated pathways [27]. Structurally, VAV1 contains a calponin homology domain and an acidic domain in the N terminus followed by a GEF catalytic active core consisting of a central Dbl homology domain, pleckstrin homology domain, and C1 domain [27]. Finally, the C-terminal region of VAV1 contains three Src homology domains in an SH3-SH2-SH3 arrangement [27]. The GEF activity of VAV1 stimulates the transition of RAC1 and RHOA small GTPases from their inactive (GDP-bound) to the active (GTP-bound) configuration [27, 42, 212]. In addition, the adaptor function of VAV1 mediates activation of the nuclear factor of activated T cells (NFAT) in synergy with signals from antigenic receptors in lymphoid cells [27, 212, 226, 119, 237, 182, 173]. In basal conditions, un-phosphorylated VAV1 adopts an inactive closed configuration in which the N-terminal calponin homology and acidic domains and the C-terminal SH3 (C-SH3) domain block access of small GTPases to the catalytic core and limit the noncatalytic activities of the protein [27, 14, 229]. Activation of VAV1 by transmembrane and cytosolic protein kinases reverses these intramolecular inhibitory interactions by promoting an open active configuration associated with phosphorylation in the acidic, C1 finger, and C-SH3 domains [27, 14, 229].

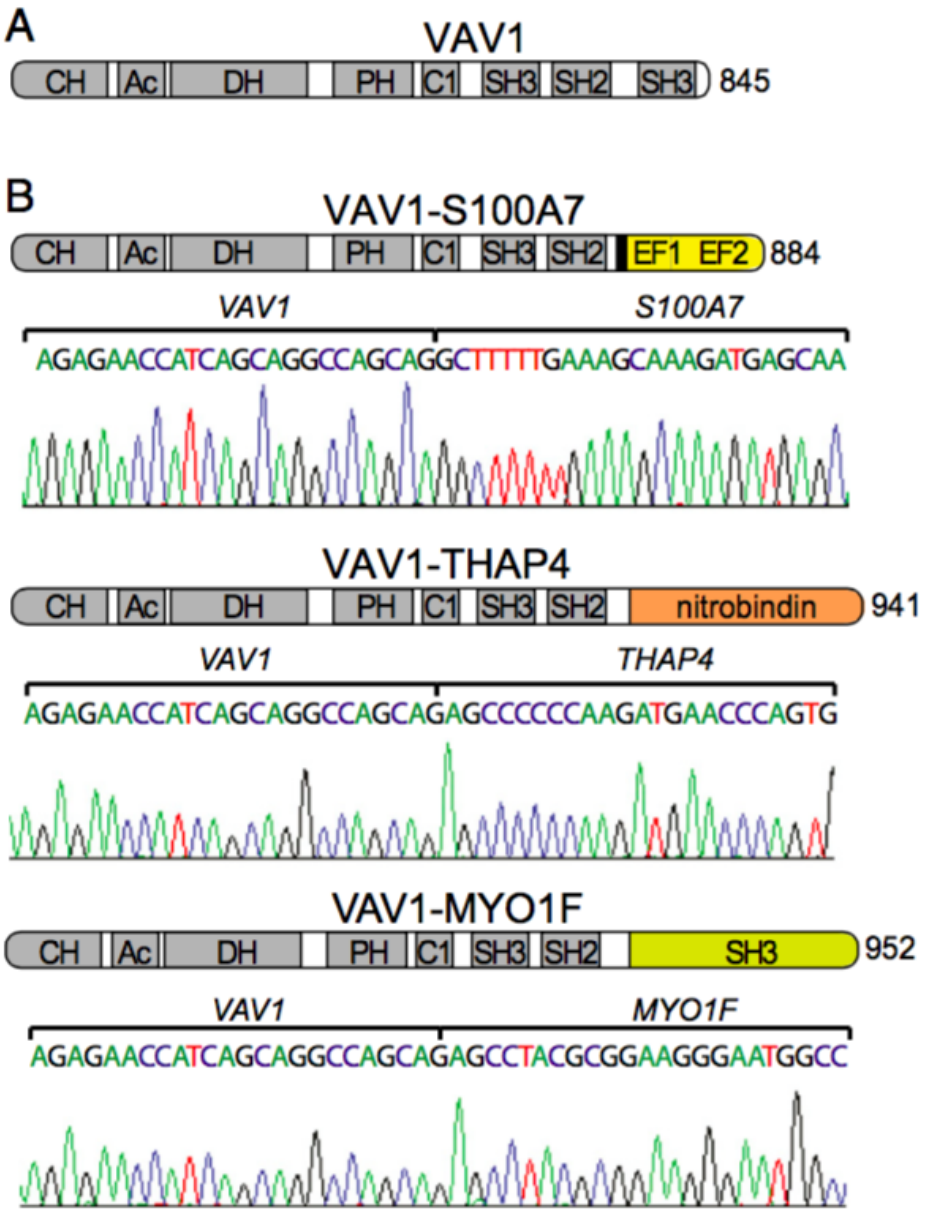


Figure 2.10: **Novel VAV1 fusion genes in PTCL:** (A) Schematic representation of the domain structure of the VAV1 protein. (B) Schematic representation of the domain structures of the VAV1-S100A7, VAV1-THAP4, and VAV1-MYO1F fusion proteins. Ac, acidic domain; C1, C1 domain; recognition motif for diacylglycerol and phorbol esters, atypical; CH, calponin homology domain; DH, DBL homology; EF, pseudo-EF hand domain; nitrobindin, nitrobindin domain; PH, pleckstrin homology domain; SH2, Src homology 2 domain; SH3, Src homology 3 domain.

VAV1 is specifically expressed in hematopoietic tissues, and plays key roles in lymphocyte development and function [212]. VAV1 is essential for T-cell receptor (TCR)-mediated cytoskeletal reorganization, cytokine secretion, proliferation, and survival [212, 182]. Thus, Vav1-deficient mice show a partial block in thymic development at the CD4- CD8- double-negative to CD4+ CD8+ double-positive transition, defective positive selection, and impaired negative selection, which altogether point to a major role for VAV1 in TCR signaling [211, 116]. Biochemically, mouse Vav1 knockout T cells fail to elicit TCR-induced intracellular Ca2+ flux and to activate MAP/ERK pathway and NF-kB signaling [67, 40, 91, 172]. Consistently, the function of mature T-cell populations is also defective in the absence of Vav1, with reduced TCR-induced proliferation and cytokine secretion [212, 201, 235]. Similarly, VAV1-null human JURKAT T cells show impaired TCR-induced calcium flux, IL-2 transcription, and NF-ÎžB activation, as well as decreased TCR-induced JNK and NFAT signaling [29].

Given the prominent role of VAV1 in T-cell activation and to explore the functional consequences of PTCL-associated VAV1 mutations and gene fusions, we analyzed the effect of these genetic alterations on lymphocyte signaling. Analysis of JNK signaling in AP1 reporter assays, a functional readout of VAV1 catalytic-dependent functions downstream of RAC1, showed marked increased JNK activation in JURKAT cells expressing the VAV1-MYO1F, VAV1-S100A7, and VAV1-THAP4 fusions. Notably, this effect was primarily independent of TCR stimulation with anti-CD3 supporting that PTCL-associated VAV1 fusion proteins adopt a constitutively active configuration. The inhibitory role of VAV1 C-terminal SH3 domain involves its folding over to the N-terminal catalytic and pleckstrin homology domains, which occludes the access of VAV effector factors to the catalytic GEF domain [14]. Thus, we postulated that the loss of the C-terminal SH3 domain in the VAV1-MYO1F, VAV1-S100A7, and VAV1-THAP4 fusions would result in VAV1 activation via loss of these inhibitory intramolecular interactions. To test this hypothesis, we analyzed the levels of Tyr174 phosphorylation, a regulatory posttranslational modification indicative of an ac-

tive VAV1 open configuration [42, 229]. Consistent with the loss of the inhibitory role of the VAV1 C-terminal SH3 domain, immunoprecipitation of HA-tagged VAV1 proteins with anti-HA antibody in these cells, followed by immunoblotting with an antibody recognizing phospho-Y174, showed high levels of phosphorylation in VAV1-MYO1F, VAV1-S100A7, and VAV1-THAP4 fusions. These results support that the PTCL-associated fusion proteins VAV1- MYO1F, VAV1-S100A7, and VAV1-THAP4 can adopt an open configuration even in the absence of TCR stimulation and mechanistically implicate the loss or impairment of the inhibitory role of the C-terminal SH3 domain of VAV1 in the pathogenesis of PTCL.

2.6 Conclusions

Gene fusions arising from chromosomal translocation can be highly transforming events, and quite a common finding in cancer genomes. Fusion detection packages typically have many false positives and there is a great need to functional annotation and driver prediction. Pegasus frames the predicting of driver gene fusions as a supervised learning task, taking advantage of large databases of oncogenic fusions as positive training examples. The trained model outperformed a competing package, Oncofuse, and the most informative feature in the classifier was the fusion being in-frame (a feature OncoFuse does not calculate). We successfully recapitulate much of landscape of driver gene fusions in glioma, and predict novel fusions in PTCL, namely TRAF1-ALK and VAV1 with its various partners. Pegasus is successfully deployed at many research labs beyond our own.

Chapter 3

Viruses as Probes of Tumor Genetics and Microenvironment

Similar to the gene fusions of the previous chapter, cancer-causing viruses are a class of potentially high-penetrance alterations in promoting transformation. Oncoviruses are defined as viruses with the potential to transform the cells or tissues they infect into a malignant state. Many viruses have strong epidemiological associations with cancers, and we have a clear picture of the oncogenic mechanism for a subset of these viruses. A virus may inhabit the malignant clonal lineage of a tumor, or infect tumor infiltrating leukocytes or stromal cells. A virus-positive malignant clone may have slightly altered genetics than a virus-negative cancer. Similarly, a tumor in a virus-positive microenvironment of other infected cell types may receive modified immune signals from a tumor growing in a sterile microenvironment. The biology community does not yet have answers to such questions and is in need of robust host-viral characterization methods for high-throughput sequencing data. Sequencing of human tumor tissue has the potential to reveal new associations between

Material presented in this chapter is published, wholly or in part, in: [4] in collaboration with F. Abate, M. Ambrosio, L. Mundo, M.A. Laginestra, F. Fuligni, M. Rossi, S. Gazaneo, G. De Falco, S. Lazzi, C. Bellan, B. Rocca, T. Amato, E. Marasco, M. Etebari, M. Ogwang, V. Calbi, I. Ndede, K. Patel, D. Chumba, P.P. Piccaluga, S. Pileri, L. Leoncini, R. Rabadan; and will be published in the manuscript "Viral landscape and microenvironment of peripheral T-cell lymphomas" (in preparation) in collaboration with F. Abate, R. Rabadan, *et al.*

cancer and viral infection. Regardless of the exact location or cellular tropism of tumor-associated virus, one might view the infection as a biological probe of the local cancer biology. In this chapter we describe a computational strategy for simultaneous characterization of viruses present in human sequencing data and host genomic profiling from RNA-seq data. We then apply the suite of tools, Pandora, to the study of human herpesviruses in lymphomas. The two lymphomas we consider are Burkitt lymphoma and peripheral T-cell lymphoma.

3.1 Abbreviated History of Viruses in Cancer

Viruses were discovered in the late 19th century through the study of a transmissable disease of the tobacco plant. Experiments by Mayer, Ivanovsky, and Beijerinck in succession demonstrated the transmissibility of the tobacco mosaic disease via a filtrate of ground leaves from an afflicted plants that had been passed through a ceramic filter too fine for any cellular pathogen to traverse. The existence of infectious agents too small for microscopic examination had been demonstrated, and it was Beijerink who coined the term virus. In 1911 Peyton Rous published his observations [178] on an aggressive sarcoma of chickens, also relying heavily on filtrates of diseased tissue to demonstrate the transmissibility of the cancer between hosts. Rous's paper never mentions the term virus, though it is widely considered the dawn of the study of viruses as etiologic agents in cancer. In the years since, the biomedical community has come to recognize many cancer-associated viruses. Hepatitis B virus was discovered at the NIH in the mid 1960's although its causative role in hepatocellular carcinoma was not verified until 1981. In 1964, Epstein, Achong, and Barr published a paper [58] reporting virus particles in cultured lymphoblasts from Burkitt lymphoma, a virus that would later bear their names as Epstein-Barr virus (EBV). Between 1979-1981, the first human retrovirus was discovered, human T-lymphotropic virus 1 (HTLV-1), and it also happened to be a potent oncovirus, strongly associated to acute T-cell leukemias. In 1984-1986 there came the discoveries of the human papillomavirus virus strains 16 and 18,

which are responsible for the majority of cervical cancer. In 1989 the cDNA from a non-A non-B hepatitis virus was isolated from diseased liver tissue and heralded the discovery of hepatitis C virus [36]. After the onset of the AIDS epidemic, Chang and Moore isolated DNA sequences from Kaposi sarcoma that were homologous to herpes viruses and that were rarely present in non-Kaposi sarcoma tissue samples from AIDS patients. Thus, in 1994, came the discovery of Kaposi sarcoma herpesvirus (KSHV) [32]. In 2008, the same Chang and Moore uncovered the most recent oncovirus, Merkel cell polyomavirus (MCPV), which is clonally integrated into Merkel cell cancers [64]. The method they pioneered for the MCPV discovery is called digital transcriptome subtraction, and serves as the inspiration for our development of Pandora.

The are many biological differences among the oncoviruses highlighted above, and certainly differences between the different cancers they cause. HBV, EBV, HPV, KSHV, and MCPV are all DNA viruses, while HTLV-1 and HCV are RNA viruses. HBV, MCPV, HPV, and HTLV-1 integrate into the host cell's DNA, while herpesviruses like EBV and KSHV rarely integrate (they remain nuclear episomes). EBV is associated to many cancers including Burkitt lymphoma, Hodgkin lymphoma, gastric carcinoma, nasopharyngeal carcinoma, and peripheral T-cell lymphoma (PTCL). The other viruses are associated with only a single malignancy each. Perhaps the most important distinction to draw, however, is whether a virus is directly or indirectly implicated in certain cancer type. A direct oncogenic mechanism would require the viral infection be clonal in the malignant cell lineage. In the case of a virus like HPV, not only is it integrated into the genome of its host cell, but we have elucidated the exact mechanisms through which the viral gene products (E6, E7) degrade host tumor suppressor proteins (p53, pRb) and therefore promote transformation. An indirect mechanism could be similar to the case of chronic HCV infection, where chronic inflammation of the liver ultimately drives transformation. Another indirect mechanism may be involved in EBV's association to cancers in which it's not clonal, such as gastric cancer and PTCL. If a virus is clonal in the malignant cell lineage, it may help probe the *genetics* of

the cancer, while if a virus infects the non-malignant cell-types of the tumor, it may help probe the *microenvironment* of the cancer.

3.2 Pandora: Viral Detection From the Transcriptome

In the spirit of Chang and Moore's discovery of MCPV in 2009 [64], we wish to implement a computational strategy based on digital transcriptome subtraction. We recognize the bulk RNA-seq of virally infected tumor tissue will capture a set of transcripts quite diverse in origin. The goal of Pandora is first to cleanly separate the viral contribution to the sequencing data from the human host contribution. The viral reads can be used to profile the virus's gene expression while the human host reads can help define the cellular composition of the tumor. A pictorial representation of the Pandora strategy can be seen in Figure 3.1.

We segment the Pandora routines into three Modules, all taking paired-end RNA-seq data as input. The first Module is the detection of variants from RNA, leading to somatic mutation and RNA editing calls. The second Module is the viral detection and expression profiling. The third Module is the cell-type deconvolution using host gene expression. Thematically, we notice that each of these three Modules is intended to provide an important clue into the tumor's host-viral interactions: the tumor genetics, the viral expression program, and the tumor microenvironment composition.

The first Module is implemented using the GATK recommendation for variant calling from RNAseq [77, 48] and uses the SAVI algorithm [209] for statistical annotation of predicted variants. Predicted variants are separated into likely somatic mutations vs. likely RNA editing events using the RADAR database annotations for high confidence A to I editing [171]. Gene fusions are called first via chimerascan [97], and then annotated and scored by Pegasus [1] for their oncogenic potential. The set of point mutations, indels, gene fusions, and A to I RNA editing constitutes the extent to which the host tumor genetics is profiled

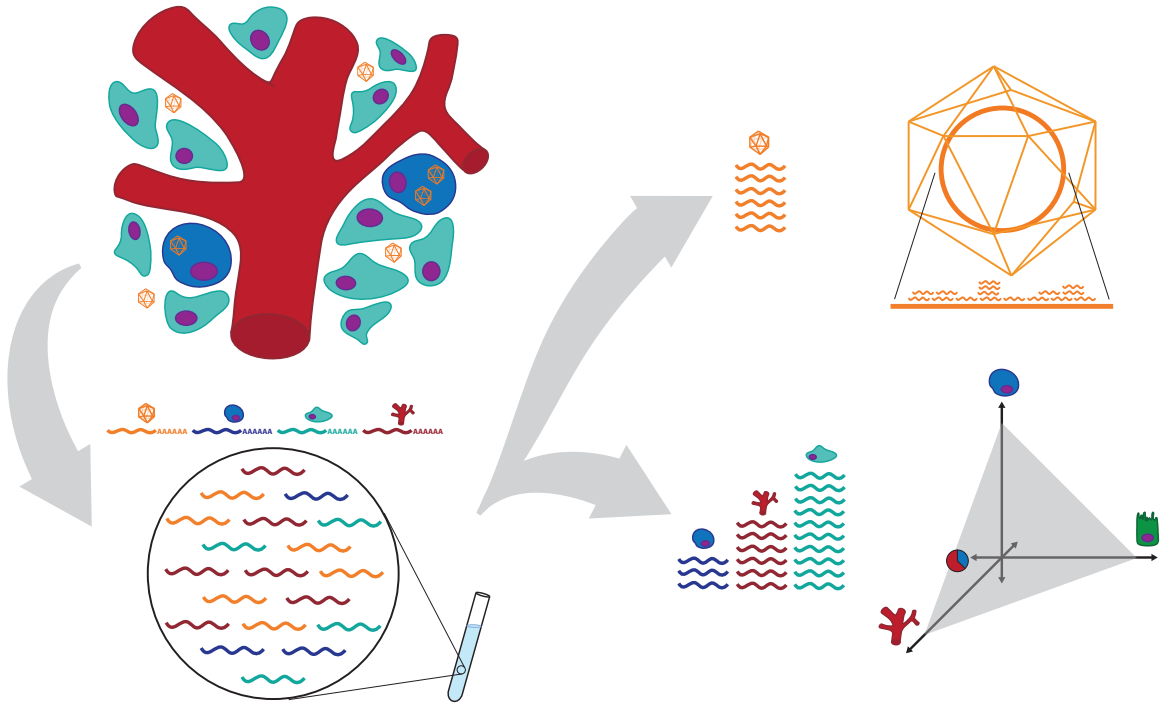


Figure 3.1: **Detection and profiling of viruses in their cellular context:** Whole transcriptome sequencing provides an unbiased view of both viral and host expression prpgrams. Tumors contain a diversity of cell types including the malignant lineage, benign stroma, and infiltrating immune cells. We represent hypothetical virions with the orange icosahedra and further illustrate their infecting the dark blue B-cell. Pandora efficiently separates human from viral transcripts and can be used for both oncovirus discovery and characterization of virus-host genetic interaction.

in Module 1.

Module 2 is concerned with the sorting of reads into viral and human bins, and the subsequent use of the viral reads to call infections. Unlike other virus detection packages, Pandora has a 3-step process of (i) subtracting the human reads, (ii) *de novo* assembling the non-human reads, and (iii) BLAST long contigs against nt database. We feel that the intermediate step of assembling contigs from short reads can decrease false positive alignments and reduce the number of BLAST subcalls to be made (BLAST can be very time-intensive). To subtract the human reads, a two-pass splice-aware alignment is performed, the first pass using STAR aligner [52] and the second pass using Tophat [206]. The hg19 human reference is used for both alignment runs. Any reads mapping to the human genome

reference are separated into a "human_expression" file (to be used in Module 3), while the unmapped reads are written back to disk in preparation for assembly. Trinity [83] is used for *de novo* assembly of the unmapped reads. This produces a set of nucleotide contigs that could range in length from from 200 to 100,000 if there is a high load of a virus with a large genome. The blastn [8] utility within the BLAST package is used to align each contig longer than a user-defined 'contig_threshold' against the entire nt database of NCBI. Reference genome and annotations are downloaded from GenBank for recurrently detected viral taxa. The unmapped read set is then directly aligned to the viral reference with Tophat [206] in order to produce the viral gene expression profile.

We believe there are significant advantages in throughput and accuracy to our decision to perform contig assembly prior to BLAST. Interestingly, there are reads that neither map to the human genome nor assemble into contigs. There are also contigs that have no BLAST hits across all of nt, even at low alignment stringency. These orphan reads and contigs constitute a difficult to characterize set of sequence data, but could be an attractive target for open reading frame (ORF) finding algorithms. Pandora does not currently contain functionality to predict novel viruses or microbes; this may be a future direction.

Module 3 begins with the "human_expression" file generated as a consequence of host subtraction in Module 2. The featurecounts [123] utility within the subread package is used to quantify the read counts for each gene specified in the RefSeq coordinates. Gene expression values are normalized using transcripts per million (TPM). Our next question concerns the relative contributions of infiltrating leukocytes into the tumor, and whether this can be inferred from the host expression profile. We take advantage of recent flow cytometry based sorting of different leukocyte populations followed by RNAseq. The experiments performed in [147] establish canonical expression profiles for 22 purified immune cell populations (LM22), and these pure profiles can be used to infer the relative abundances of those 22 immune cell types in our bulk tumor tissue. A probabilistic deconvolution method [167] is applied to the host gene expression data in the presence of the LM22 purified profiles, from

which we recover our estimates of the microenvironment cellular composition.

In order to validate the detection performance of Pandora, we prepare synthetic sequencing files at varying concentrations of viral reads. Many viral and microbial taxa are included in the simulated reads, across a range of genome sizes. To simulate Illumina reads we use ART [94] in paired-end mode with 100bp reads. The detection task will be evaluated in precision-recall space, since it is impossible to gauge the number of true negatives (TN) and measures like specificity rely on the TN rate. In Figure 3.2 we plot performance of Pandora as a function of the user-defined contig threshold. This parameter affects the sensitivity (recall) because it is the length of contig, below which, we do not search the nt database for alignments. Surprisingly, we do not see a significant loss of precision as we increase recall (corresponding to lowering the contig_threshold so that even short contigs are BLASTed). We also plot the recall vs. the contig threshold and naturally there is a loss of sensitivity first for the larger genomes (cellular pathogens) with fewer reads.

We next compare the detection performance of Pandora against a widely used tool, VirusFinder [216]. Since VirusFinder only claims to detect viral pathogens we omit the synthetic data using cellular microbes. VirusFinder relies on performing BLAST directly on short reads after host removal, unlike Pandora's assembly of long contigs prior to BLAST. We split our synthetic mixture of many different viral reads into species-specific samples, so that we can compare the head-to-head performance by taxon. Both Pandora and VirusFinder had identical recall except for two recently discovered viruses (MVC and PVS) which are not contained within VirusFinder's database. Pandora and VirusFinder have more divergent results when calculating precision, and the panel of comparisons can be seen in Figure 3.3. Within the family of herpesviridae, where there is sequence homology among HHV1-HHV8, VirusFinder consistently calls false positive taxa. Since Pandora is attempting to align much longer sequences against the more up-to-date nt database we suffer no decrease in precision within a family of closely related viruses.

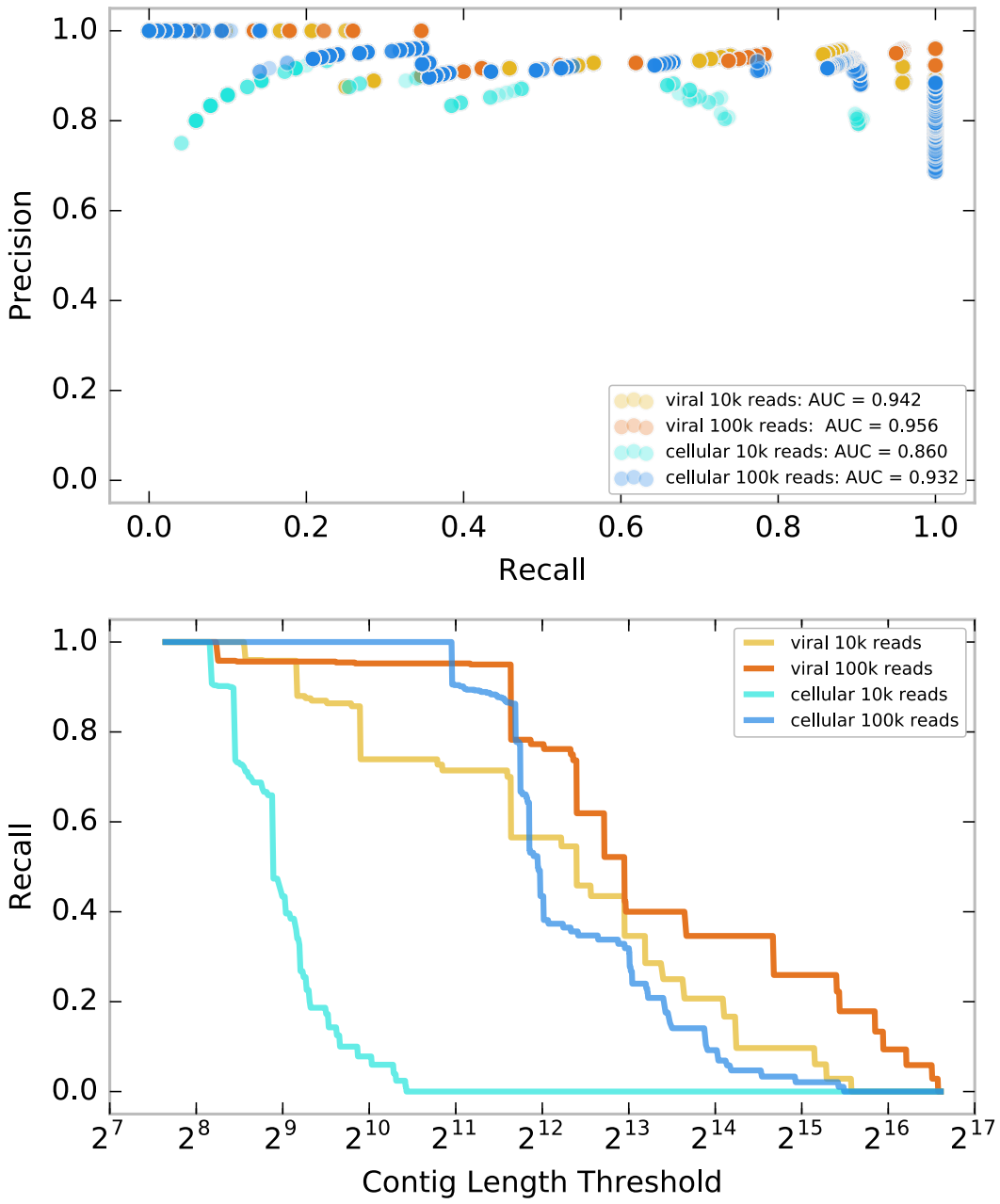


Figure 3.2: **Recovery of virues and microbes in synthetic data:** The length of assembled contig required to initiate local alignment searching is a key user-defined parameter of Pandora. This tunable threshold control the speed with which the pipeline can run, and necessarily impacts detection stringency. (Top) Pandora performance is plotted in PR space for a mixed population of viral pathogens at two concentrations and similarly for a mixed population of cellular pathogens. (Bottom) The recall (sensitivity) depends on a combination of parameters including the contig length threshold, the genomic size of the pathogen, and the concentration of the pathogen.

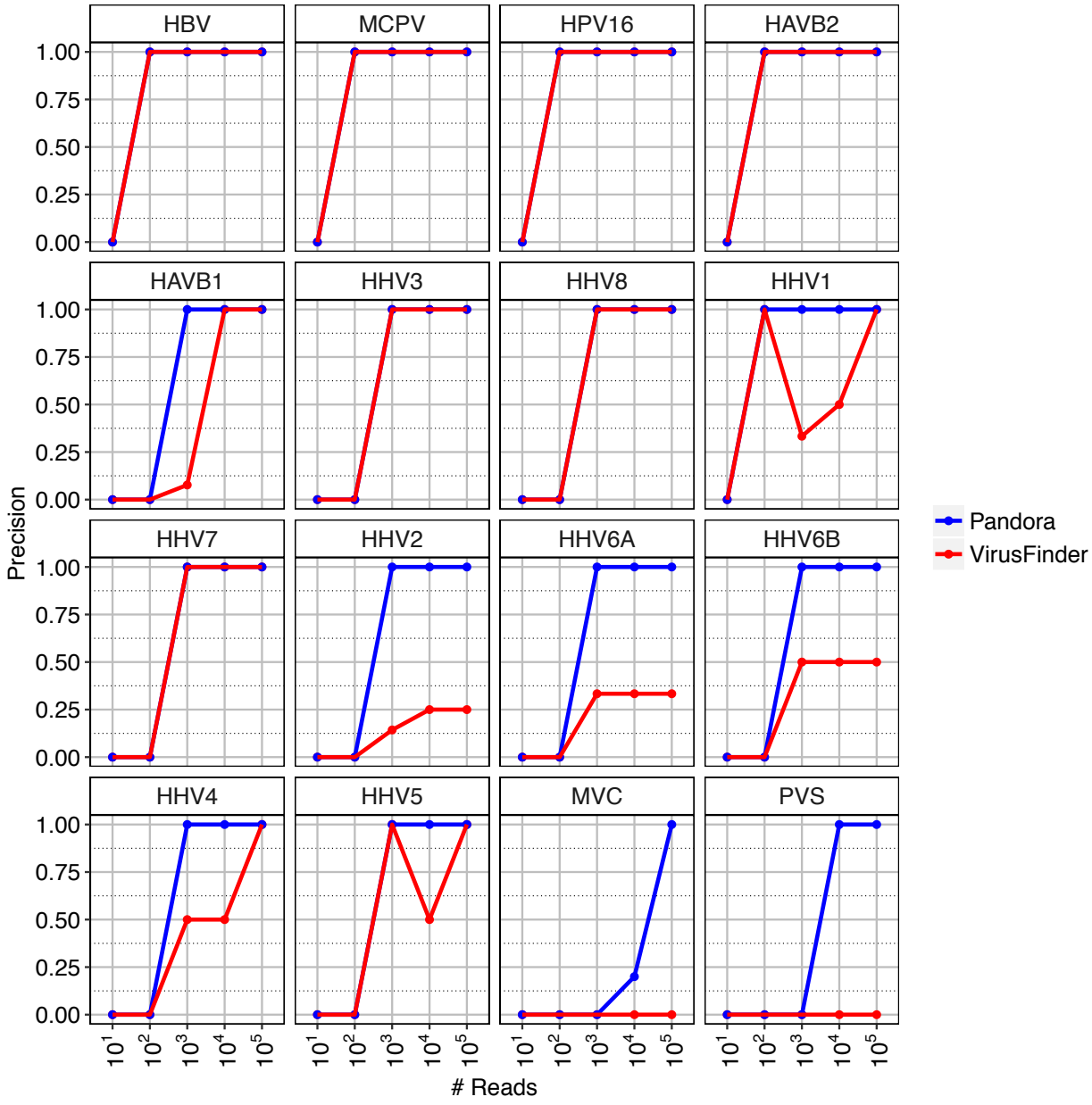


Figure 3.3: **Superior precision in viral detection:** Comparison between Pandora and VirusFinder [216] on synthetic data highlights the importance of our choice to assemble transcripts prior to BLAST. False positives in VirusFinder lead to decreased precision, particular in pairs of highly similar viruses such as the herpes simplex viruses (HHV1 and HHV2) or the HHV6 species. Many existing tools such as VirusFinder also do not maintain up-to-date databases and we expose this with fact with two contrived cases of recently discovered viruses – megavirus chilensis (MVC) and pandoravirus salinus (PVS).

3.3 Applications in Burkitt Lymphoma

Burkitt lymphoma (BL) is the first human cancer to be associated with the Epstein-Barr virus (EBV), the first tumor to exhibit a chromosomal translocation activating an oncogene (MYC), and the first lymphoma to be associated with human immunodeficiency virus (HIV) infection. The World Health Organization classification describes three clinical variants of BL: endemic, sporadic, and immunodeficiency-related. These variants are similar in morphology, immunophenotype, and genetics. While the sporadic variant (sBL) occurs outside of Africa and is rarely associated with EBV infection, the endemic variant (eBL) arises mainly in Africa and is associated with malaria endemicity and EBV infection in almost all cases. Epidemiological studies have shown that malaria and EBV combined do not fully explain the distribution of eBL in high risk regions [155]. Malaria and EBV are in fact ubiquitous within the lymphoma belt of Africa, suggesting that other etiologic agents may be involved [213]. However, it is unclear what other epidemiological factors could play a role in the genesis of eBLs.

Three types of EBV latency have been described in EBV-related lymphomas according to the pattern of EBV nuclear antigen (EBNA) and the latent membrane protein (LMP) expression, namely latency I, II, and III [202]. Specifically, latency I is usually associated with eBL and it denotes a transcriptional program in which an EBV infection does not produce virions and expresses a single protein, EBNA-1. While the latency I program has been extensively characterized *in vitro*, a different form of latency has been recently reported in 15% of eBL that uses a different set of promoters. Termed Wp-restricted latency [109], this program shows a homogeneous host expression signature [108] characterized by down-regulation of BCL-6 and up-regulation of IRF-4 and BLIMP-1. Other reports have described latency program heterogeneity at single cell level [107] and low expression of LMP genes in a fraction of cases [16, 148]. Heterogeneous EBV transcription profiles with LMP expression have been recently reported in some cases of AIDS-related and sporadic BL [10], but extensive data on endemic cases are not available yet. These studies indicate that the transcriptional EBV

programs of primary eBL could be more complex than expected across cases and within individuals. Therefore, the exact role of EBV has remained elusive and further investigation is required.

The genetic hallmark of all three clinical variants of BL is the t(8;14) translocation involving the juxtaposition of the immunoglobulin heavy chain locus (IGH) with the MYC oncogene [44]. However, although transgenic mice expressing MYC under the control of the intronic IGH enhancer (EÎij) develop B cell lymphomas [6], successive molecular characterization demonstrated that this model does not fully recapitulate the human disease. The comparison between the gene expression profile (GEP) of BL and diffuse large B-cell lymphoma (DLBCL) highlighted a distinct signature of BL characterized by the expression of both MYC targets and germinal-center B-cell genes [45]. Furthermore, hypermutation and different breakpoint patterns of IGH/MYC translocation [146, 30] suggests that the origin of human BL derives from aberrant class switching in the germinal center (GC), while transgenic IGH/MYC mice typically arise from precursor/naive B-cells. The more accurate PI3K/MYC transgenic mouse model by Sander *et al.* [181] better recapitulates the human phenotype of BL and highlights the importance of the PI3K pathway in the disease. Moreover, GEP analysis has demonstrated that the transcriptional profile of eBL is different from that observed in sBL [165]. Recent studies have unveiled the genetic landscape of sBL characterized by mutations affecting the B-cell receptor (BCR) pathway and in particular the transcription factor TCF3, its negative regulator ID3, the cell-cycle G1/S regulator CCND3 [186, 185], and the chromatin-remodeling gene ARID1A [82]. On the contrary, very little is known about the spectrum of alterations in eBL, how it might differ from that of sBL, the correlations between host mutation and viral infection, and the specific viral/host transcriptional programs.

We characterize the presence of other potential agents, to define the EBV transcriptional profile and to link these profiles to the mutational status of new and previously reported genes. We leverage Pandora to characterize the mutational and viral landscape of eBL

using 20 cases from Uganda. RNA-Seq, in contrast with earlier microarray-based expression studies, provides the opportunity to identify and associate microbial and tumor mutational and expression profiles.

As we see in Figure 3.4, 20/20 cases of eBL are positive for EBV, 5/20 cases contain CMV, 4/20 KSHV, and 1/20 HTLV-1. HIV is not detected in any case, confirming that pediatric eBL is rarely associated with the immunodeficiency syndrome [144]. Nested PCR and immunohistochemical (IHC) analysis performed on all 20 original samples confirmed the presence of all the viruses in this discovery cohort. To assess whether RNA-Seq findings generalize for EBV, CMV, KSHV, and HTLV-1, we assayed for the presence of these four viruses in 20 additional cases from western Kenya by IHC. In this Kenyan cohort, EBV was detected in 20/20 samples, CMV in 8/20 samples, KSHV in 7/20 samples, and HTLV-1 in 0/20 samples. Therefore, over the 40 cases, we report the overall viral infection frequencies of 40/40 (100%) for EBV, 13/40 (32.5%) for CMV, 11/40 (27.5%) for KSHV, and 1/40 (2.5%) for HTLV-1. IHC analysis demonstrated the presence of CMV in the stromal cells and macrophages localized within the tumors and in the adjacent reactive lymphoid tissue. KSHV was identified not only in normal B-lymphocytes and endothelial cells from the adjacent reactive lymphoid tissue, but also in one case in about 5âĂŞ10% of neoplastic cells. HTLV-1 was detected in reactive T-lymphocytes in the only positive case of the discovery cohort. We also compare the viral landscape of endemic and sporadic cases by analyzing 27 RNA-Seq sBL samples from Schmitz *et al.* [186] with Pandora. The analysis shows the presence of EBV and HIV respectively in 4/27 (15%) and in 1/27 (4%) cases (Figure 3.4), consistent with several literature sources [87].

Beyond identification of EBV presence, RNA-Seq enabled us to quantitatively analyze the viral transcriptional program. In addition to EBER-1 and EBER-2 transcripts, expression analysis of the viral genes showed the expression of EBNA-1, a gene associated to latency I type, in 18/20 cases (Fig 2A). We also detected either LMP-1 or LMP-2A, characterizing the latency II type, in 13/20 samples (65%), and also EBNA-2 in 1/20 cases (5%).

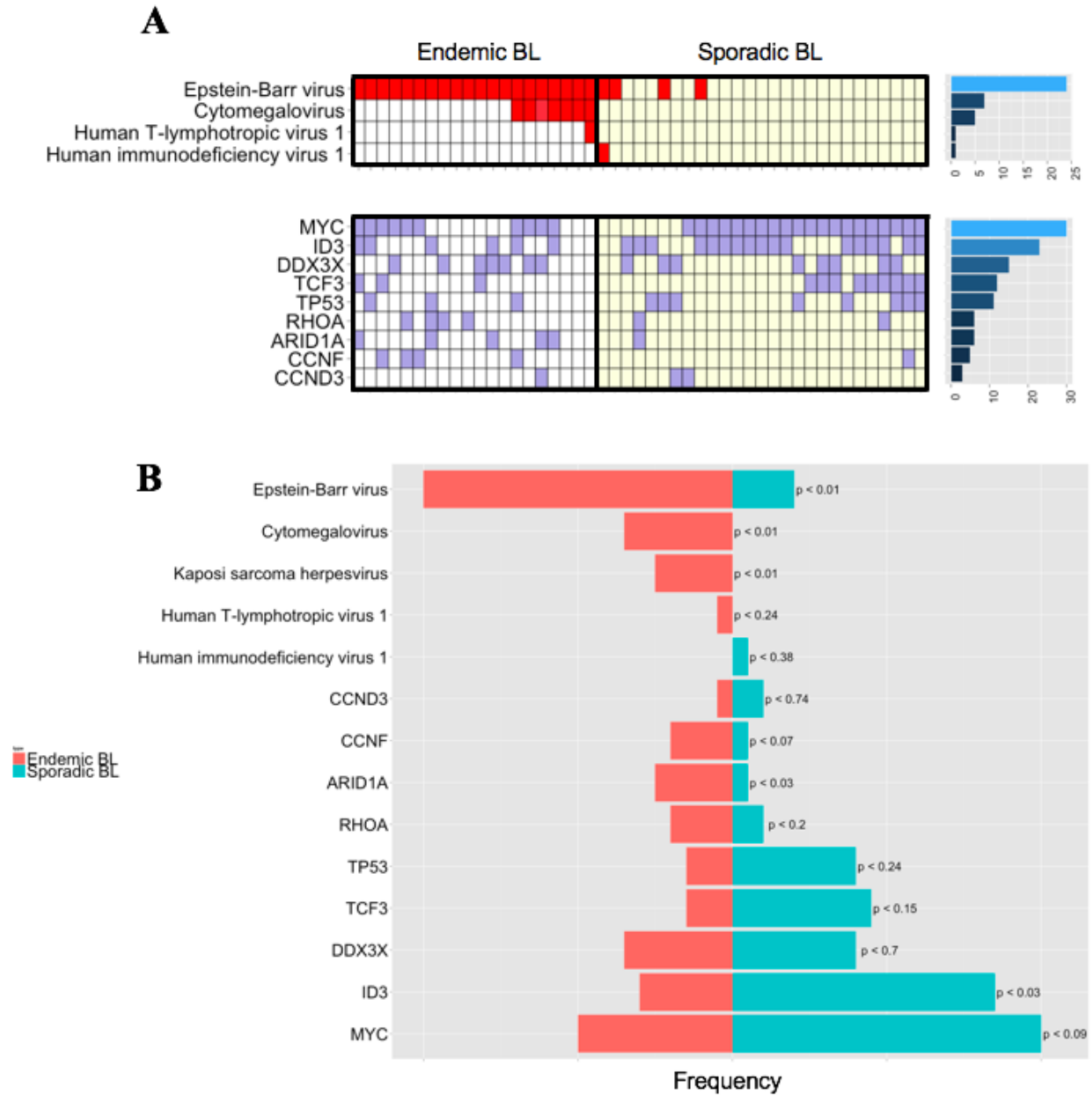


Figure 3.4: **Associations between viral and mutational presence in BL:** (A) There is a dramatic asymmetry between viral burden in endemic vs. sporadic BL. One eBL case had simultaenous expression of 3 viruses, namely EBV, CMV, and HTLV-1. (B) Frequency comparison of virus presence and driver mutations between endemic and sporadic BL. For each comparison we report the p-value associated with rejecting the null hypothesis of equal eBL and sBL prevalences.

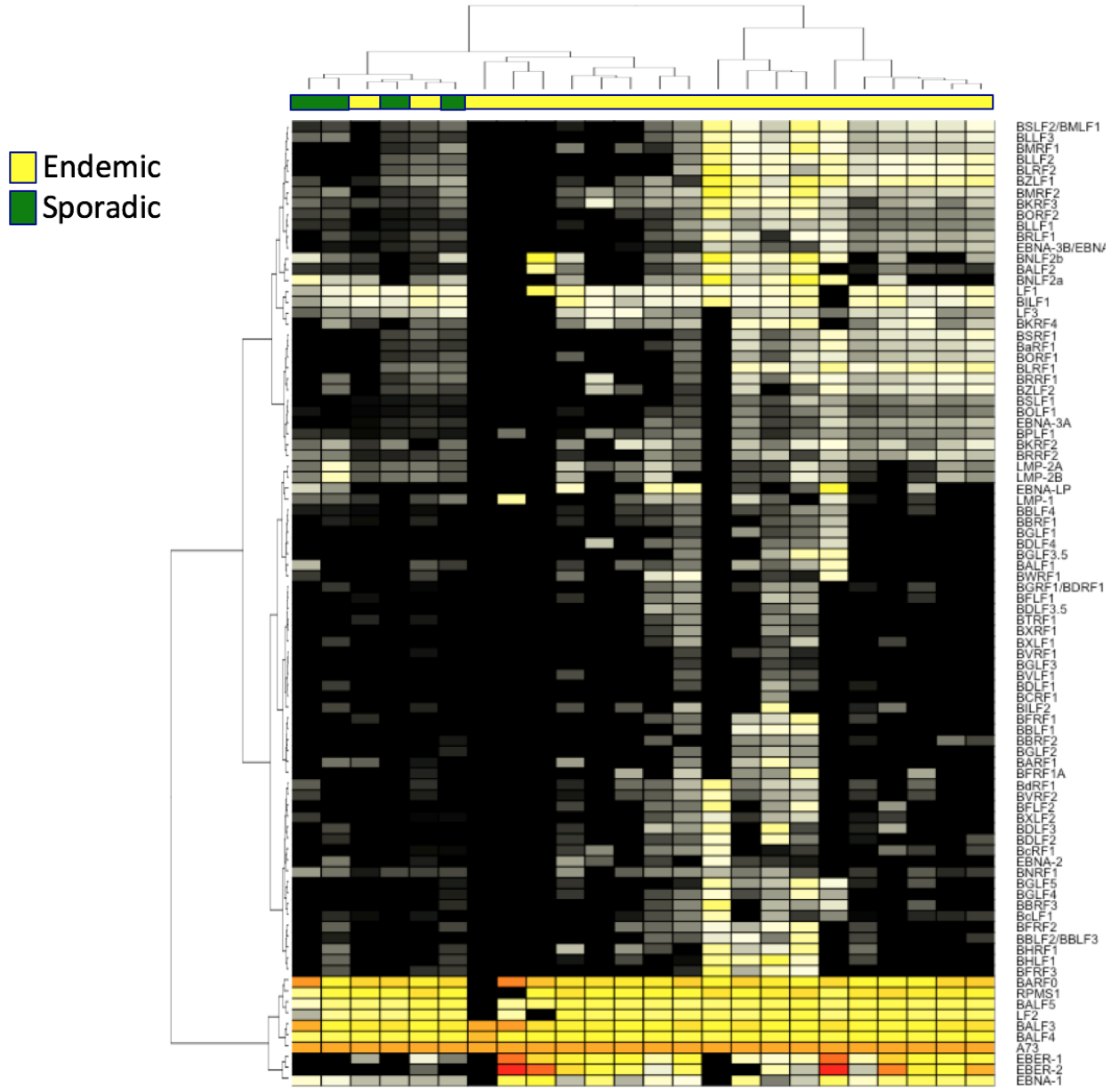


Figure 3.5: **Non-canonical EBV latency and lytic re-activation:** Unsupervised hierarchical clustering of expressed EBV genes demonstrates a diversity of non-canonical latency-associated gene expression programs with a subset of viral episome initiating lytic reactivation as indicated by expression of genes corresponding to the lytic program.

Interestingly, 2/20 cases (10%) were characterized by the expression of EBNA-3A/B/C/LP, together with the lytic gene BHRF-1, suggesting a Wp-restricted program [106]. However, the specific analysis of EBV isoforms showed the presence of H2-HF splicing event, which is hallmark of lytic BHRF-1 expression [203, 157, 122]. Unsupervised hierarchical clustering of expressed EBV genes demonstrates two main clusters (Figure 3.5) distinguished largely by gene products involved in EBV replication (BALF-2, BCRF-1, BHRF-1, BILF-1, BMRF-1, BNLF-2a, BZLF-1). The expression of these genes suggests a non-canonical latency program of the virus with a subset of viral episomes initiating lytic reactivation [106].

Due to the heterogeneity of the viral transcriptional programs, we aimed to validate the latency type by performing RT-qPCR for the EBNA-1, LMP-1, LMP-2A, EBNA-2, EBNA-3C, and BHRF-1 transcripts across an additional series of 26 cases from an extended cohort of samples from Kenya. EBNA-1 was detected in 26/26 (100%), LMP-1 and LMP-2 in respectively 5/26 (20%) and 20/26 (75%) cases, EBNA-2 in 0/26 (0%), and the combination of EBNA-3C and BHRF-1 in 4/26 (15%). These results are largely consistent with the RNA-Seq data with the exception of LMP-1 that has been detected at higher frequency in RNA-Seq. Next, we evaluated the lytic cycle activation and found BILF-1, BALF-4, and LF-2 in all 26 cases, whereas we observed the expression of BALF-2 in 23/26 (90%), BHRF-1 in 20/26 (80%), BZLF-1 and BMRF-1 in 15/26 (60%), BNLF-2a in 13/26 (50%), and BCRF-1 in 11/26 (45%) of the cases. RT-qPCR validation data is fully tabulated in Table 3.1.

Lastly, we wish to consider the host mutational spectrum alongside the presence/absence of EBV infection. The distribution of somatic mutations and viral presence across both eBL and sBL samples exhibit two interesting features (Figure 3.4). First, in eBL samples we observed lower mutational frequencies in the genes MYC, ID3, TCF3, DDX3X, CCND3 and TP53, as compared to their reported recurrence in sBL, and higher mutational frequencies in ARID1A, RHOA, and CCNF [186, 185]. Second, in sBL cases an almost mutual exclusivity can be seen between EBV presence and mutations in TCF3/ID3 both known to be driver

Table 3.1: EBV latency type validation by RT-qPCR

Gene	RNAseq	RT-qPCR	Primers
EBNA-1	100%	100%	5'-TACAGGACCTGGAAATGGCC-3'
			5'-TCTTTGAGGTCCACTGCCG-3'
EBNA-2	5%	0%	5'-TAACCACCCAGCGCCAATC-3'
			5'-GTAGGCATGATGGCGGCAG-3'
EBNA-3C	25%	30%	5'-CTGGCAAAACTTGCTCCA-3'
			5'-GTGCTTCTGCCTTATCAGA-3'
LMP-1	70%	20%	5'-CAGTCAGGCAAGCCTATGA-3'
			5'-CTGGTTCCGGTGGAGATGA-3'
LMP-2A	100%	75%	5'-AGCTGTAACTGTGGTTTCCATGAC-3'
			5'-GCCCCCTGGCGAAGAG-3'
BZLF-1	80%	60%	5'-AAATTTAAGAGATCCTCGTGTAAAACATC-3'
			5'-CGCCTCCTGTTGAAGCAGAT-3'
BHRF-1	50%	80%	5'-AGAAACACCTCTCCGCCTTT-3'
			5'-ATCCACATGTTCGGTGTGTG-3'
BMRF-1	50%	60%	5'-CAACACCGCACTGGAGAG-3'
			5'-GCCTGCTTCACTTTCTTGG-3'
BALF-2	80%	90%	5'-TGCACCTGCTAGAGAACTCG-3'
			5'-CACAGAGTACGCGACTGAGG-3'
BALF-4	100%	100%	5'-AACCTTTGACTCGACCATCG-3'
			5'-ACCTGCTCTTCGATGCACTT-3'
BILF-1	100%	100%	5'-GTCACCTTCACCGGACTCAT-3'
			5'-GTAGTAGCGGGCAACGAGAG-3'
LF-2	85%	100%	5'-CTGACCAGGACATCGTGCTA-3'
			5'-GGGGTTCTTGACCAATCTGA-3'
BCRF-1	60%	45%	
BNLF-2A	65%	50%	

genes in sBL (p-value < 0.02, Fisher exact test). To explore this hypothesis, we performed a hierarchical clustering of both endemic and sporadic cases on TCF3 target genes (previously reported in [186]) and we demonstrate that the first bifurcation of the dendrogram classifies the samples into EBV-positive and EBV-negative BL independently on the specific subtype with an accuracy of 96% (45/47). This result (Figure 3.6) shows that the TCF3 pathway is more activated in EBV-negative cases, as indicated by the significant negative enrichment of TCF3 target genes in EBV-positive samples. Furthermore, we observe that when considering the overall panel of both endemic and sporadic BL samples, the mutually exclusivity between

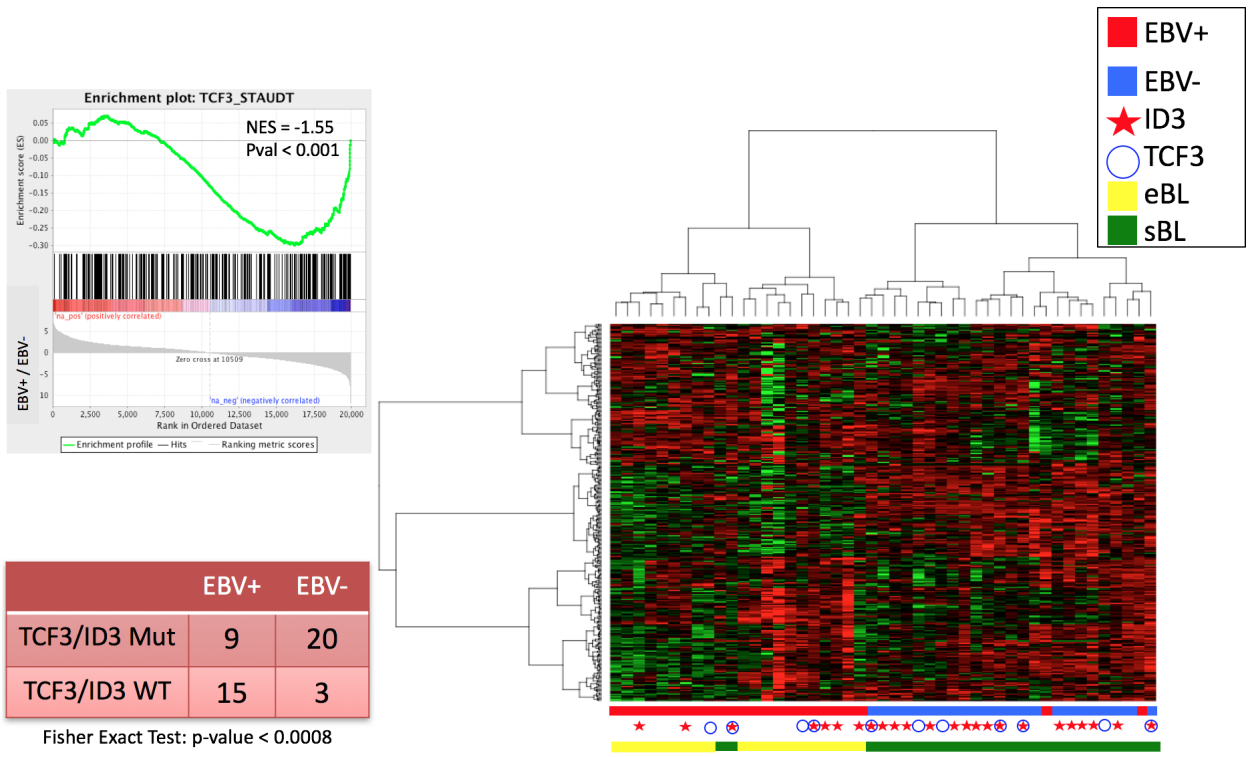


Figure 3.6: **Inverse correlation between EBV and TCF3/ID3 mutations:** The dendrogram classifies the samples into EBV-positive and EBV-negative BL independently of the specific subtype with an accuracy of 96% (45/47). The gene set used in this clustering comes from [186] and corresponds to genes upregulated in the presence of driver TCF3 mutations.

TCF3/ID3 mutations and EBV infection yields a more significant effect (p-value < 0.0008, Fisher exact test).

3.4 Applications in Peripheral T-cell Lymphoma

Peripheral T-cell lymphoma (PTCL) is a heterogeneous set of aggressive non-Hodgkin lymphomas. In order of decreasing prevalence, the four most common histologic subtypes are: peripheral T-cell lymphoma not-otherwise-specified (NOS), anaplastic large cell lymphoma (ALCL), angioimmunoblastic T-cell lymphoma (AITL), and extranodal NK / T cell lymphoma (NK). We collect 191 RNA-seq samples from recent studies in PTCL cases spanning these four histologic subtypes. Our case set includes 72 ALCL [41, 5] (37 unpublished), 60 AITL [160, 228, 5], 42 NOS [160, 5], and 17 NKTCL [118]. We run the three Pandora

Modules in series to characterize the tumor genetics, the viral landscape, and the tumor microenvironment.

We begin with Module 1, in which we quantify the landscape of genomic alterations in the 191 cases. We are primarily interested in establishing the set of highly recurrent point mutations, small indels, and gene fusions across the 4 histologic substypes of the disease. The full genomic landscape is portrayed in Figure 3.7. In this panel we observe many of the genomic hallmarks of PTCL: the mutation cluster associated to AITL (RHOA, TET2. IDH2, CD28, FYN) [160, 5], the ALK+ fusion cluster in ALCL [41] and the ALK- mutation cluster (PRDM1, TP53, STAT3, JAK1) [41]. The VAV1 fusions discussed at the end of the last chapter can be seen in NOS and AITL subtypes. The NK subtype shows no significant enrichment of a driver mutation. The NOS subtype appears to have some subpopulation that looks similar in genetics to the AITL. The RNA editing events are not included in Figure 3.7, however we point out a well-known A to I editing event in Figure 3.8. NEIL1 is a gene involved in DNA repair and the A to I editing introduces a protein changing mutation in the protein that changes its lesion specificty [227]. The appears to be a mild statistical association between the levels of NEIL1 editing and outcomes in our cohort.

Running Module 2 of Pandora provides the identities and expression profiles of any viruses infecting the tumor samples. Surprisingly we detect 4 members of the human herpresviridae, and in Figure 3.9 we observe the levels of viral expression observed for the HHV4 (EBV), HHV5 (CMV), HHV6b, and HHV8 (KSHV). EBV is known to associate with the AITL and NK subtypes of PTCL, which is confirmed by Pandora's results. Viral expression is quantified in viral reads per million human reads (VRPMHR). We proceed to plot the gene expression profiles for each human herpesvirus indepedently, to facilitate the comparison of latency or lytic activation between samples. HHV4 (EBV) is plotted in Figure 3.10. The EBV expression profile shows heterogeneous expression patterns across the samples, though there is a discernible cluster for the NK cases and an overall enrichment of AITL and NK cases in the region with higher EBV expression. HHV5 (CMV) is plotted in Figure 3.11.

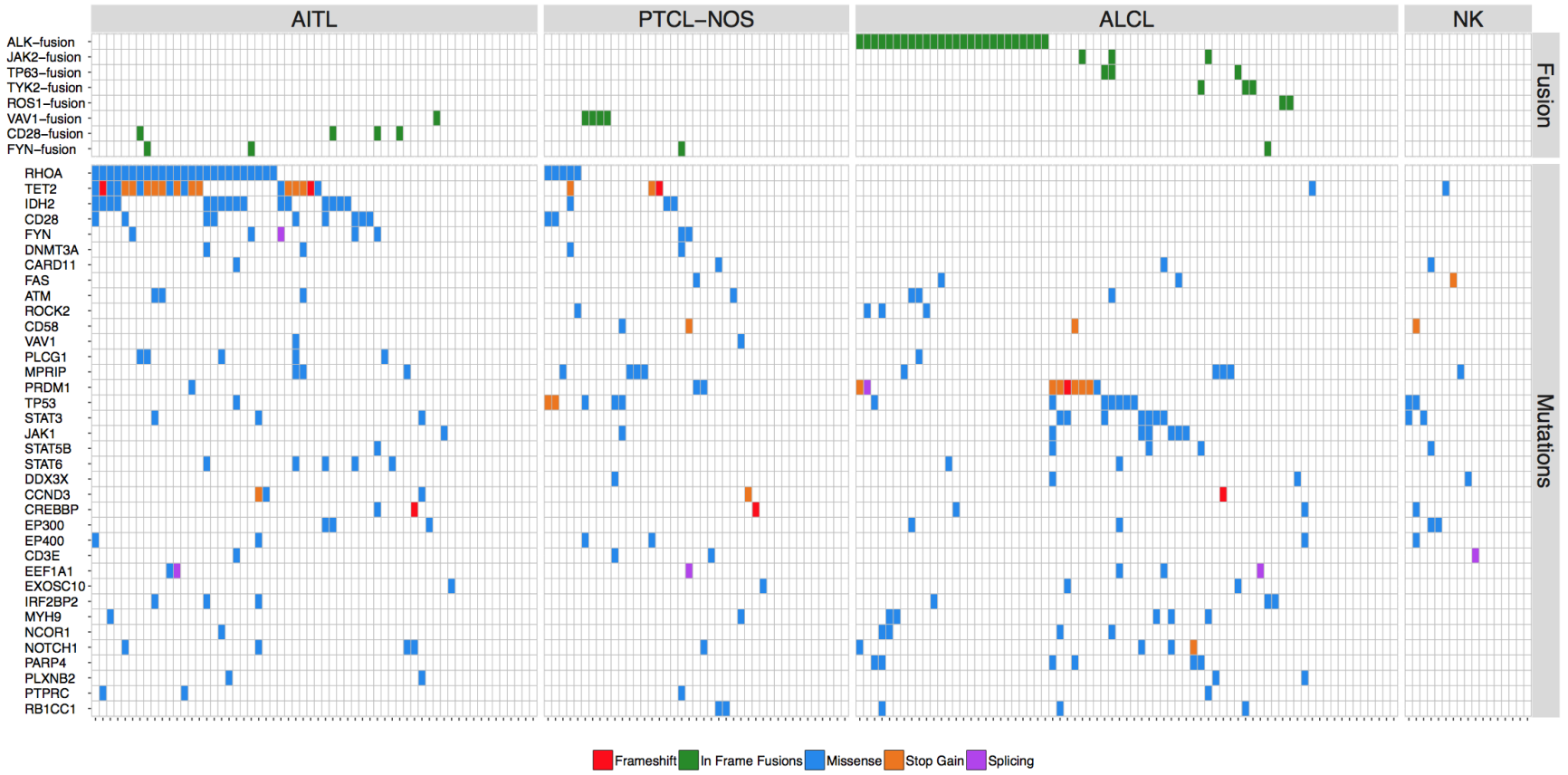


Figure 3.7: **Landscape of genomic alterations detected in PTCL:** The histologic subtypes of PTCL are segmented in this panel. Driver gene fusions identified with Pegasus [1] are displayed in the upper section, while point mutations and small indels are displayed in the lower section. We recapitulate known mutational hallmarks of the diease such as the associations between RHOA and AITL, or the association between ALK fusion and ALCL.

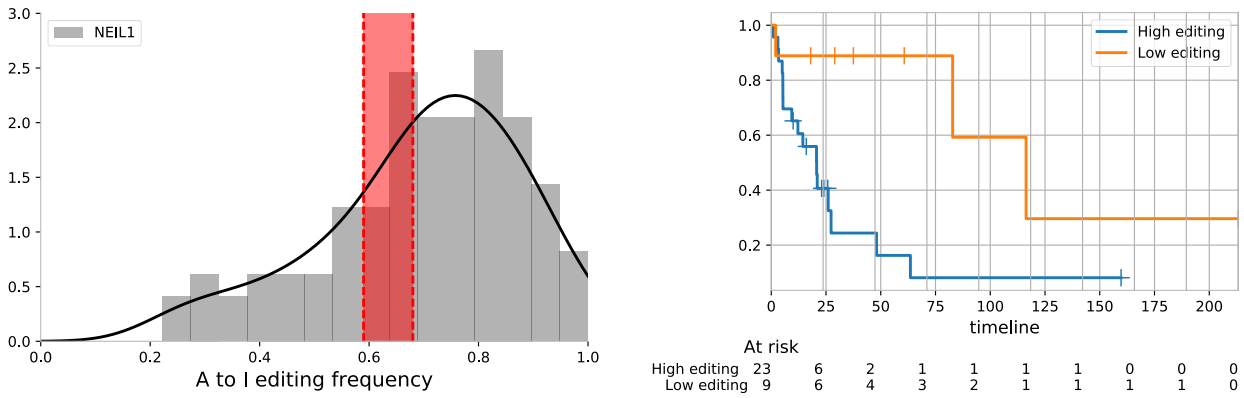


Figure 3.8: **A to I editing of host RNA:** Neil1 is a DNA repair protein that that sustains an amino acid substitution as a result of A to I editing. (Left) We highlight in red the range of editing frequences in which a high/low division could be made to preserve the statistical separation of survival curves at a p-value of 0.05. (Right) Kaplan-Meier curve describing the stratification of the PTCL cases on high/low editing frequency of NEIL1 transcripts.

HHV6b is plotted in Figure 3.12. HHV48 (KSHV) is plotted in Figure 3.13.

Running Module 3 of Pandora provides the immune cell-type abundances comprising the tumor microenvironment. The majority of the 22 cell-types we can measure (from LM22 [147]) do not have any presence in the PTCL tumors, however we highlight the 8 cell-types for which at least one PTCL subtype showed nonzero abundance (Figure 3.14). The AITL and NOS subtypes tend to track together in every cell-type. There is also a strong signature of macrophages (M1 polarization > M2), and it is generally believed that the M1 polarization is involved in modulating innate immunity [128] and responding to pathogens. The M2 polarization is associated with tissue repair and tumor progression [128]. In general there is data [153] to suggest that higher levels of infiltrating macrophages imply worse prognosis. We see certain positive controls as well. First, the NK subtype appears enriched for NK cells and gamma/delta T-cells, which is correct since those are the cells of origin. Similarly, AITL is believed to originate from Th follicular cells [47] and indeed they are enriched in the AITL subtype. In Figure 3.15 we correlate the viral load of EBV with three different host cell-type abundances. It is clear that in the case of the NK subtype, the EBV infection is in the cell of origin. The positive correlation between EBV presence and NK

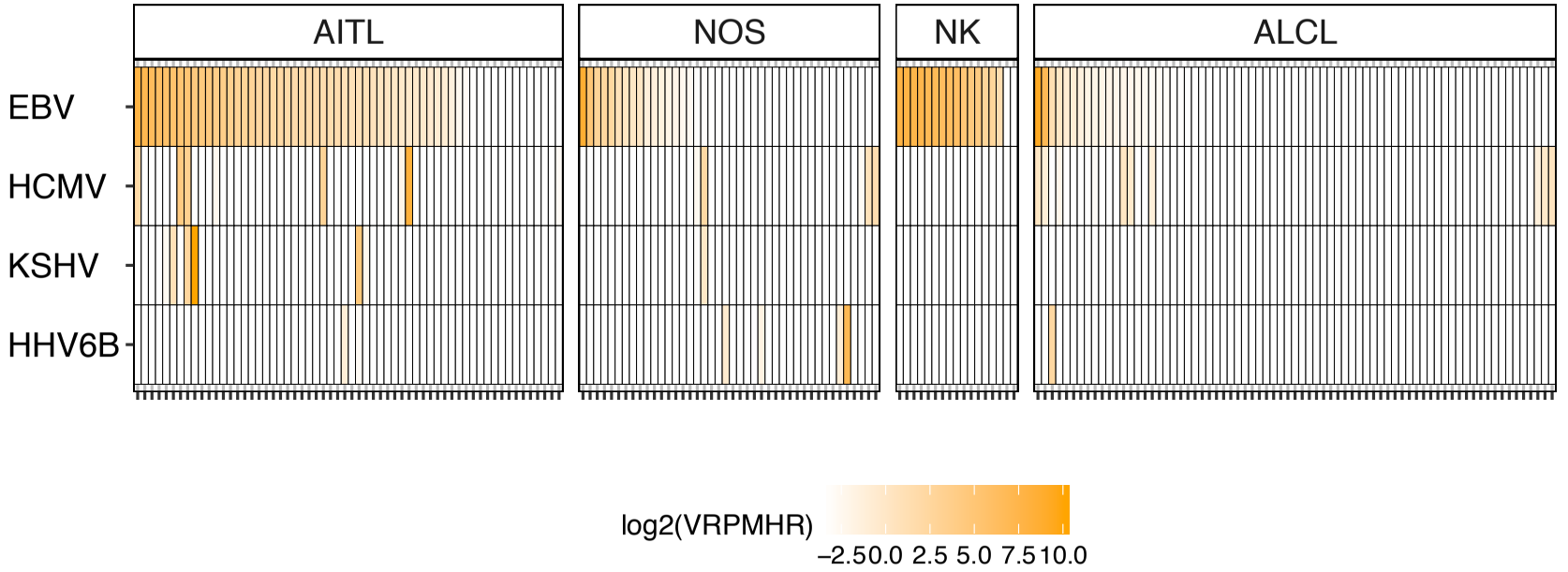


Figure 3.9: **Landscape of human herpesviruses detected in PTCL:** Viral concentrations are measured in viral reads per million human reads (VRPMHR). These four members of the herpesviridae family are detectable in multiple samples, and a marked enrichment of EBV is observed in the AITL and NK subtypes of the disease.

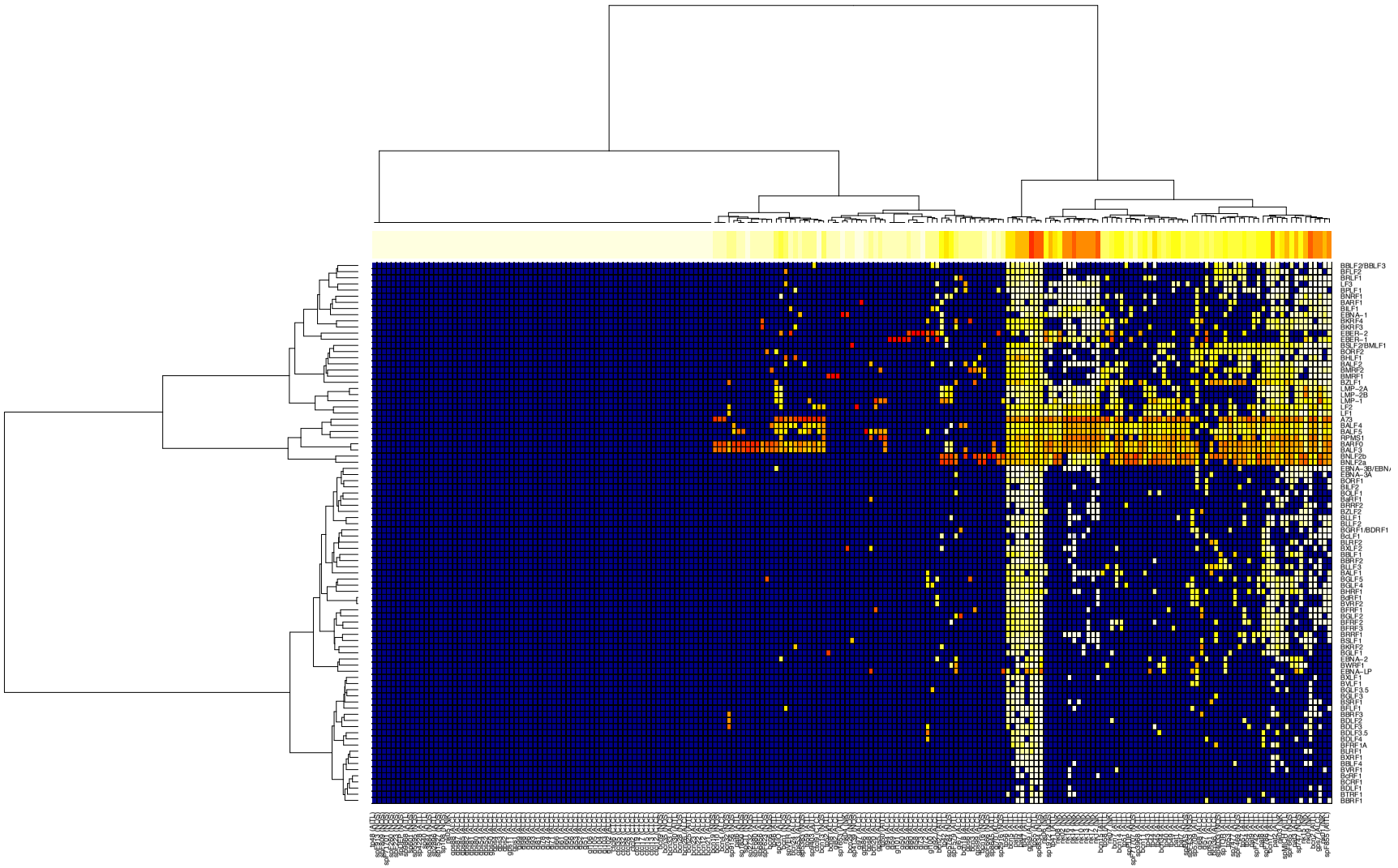


Figure 3.10: **Expression profiling of EBV:** Gene expression profiling of EBV using RefSeq annotation NC007605.

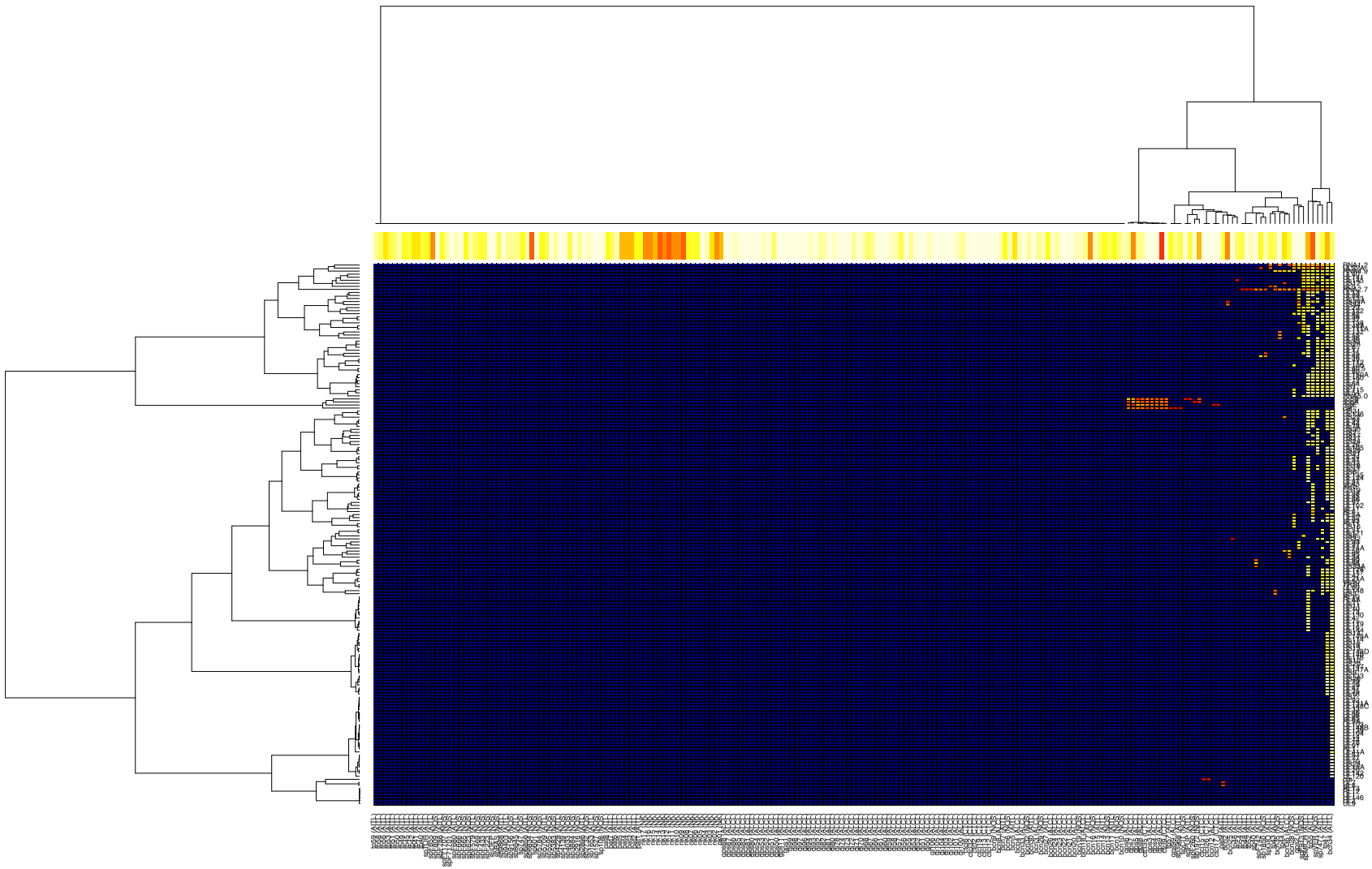


Figure 3.11: **Expression profiling of CMV:** Gene expression profiling of CMV using GenBank annotation GU179001.

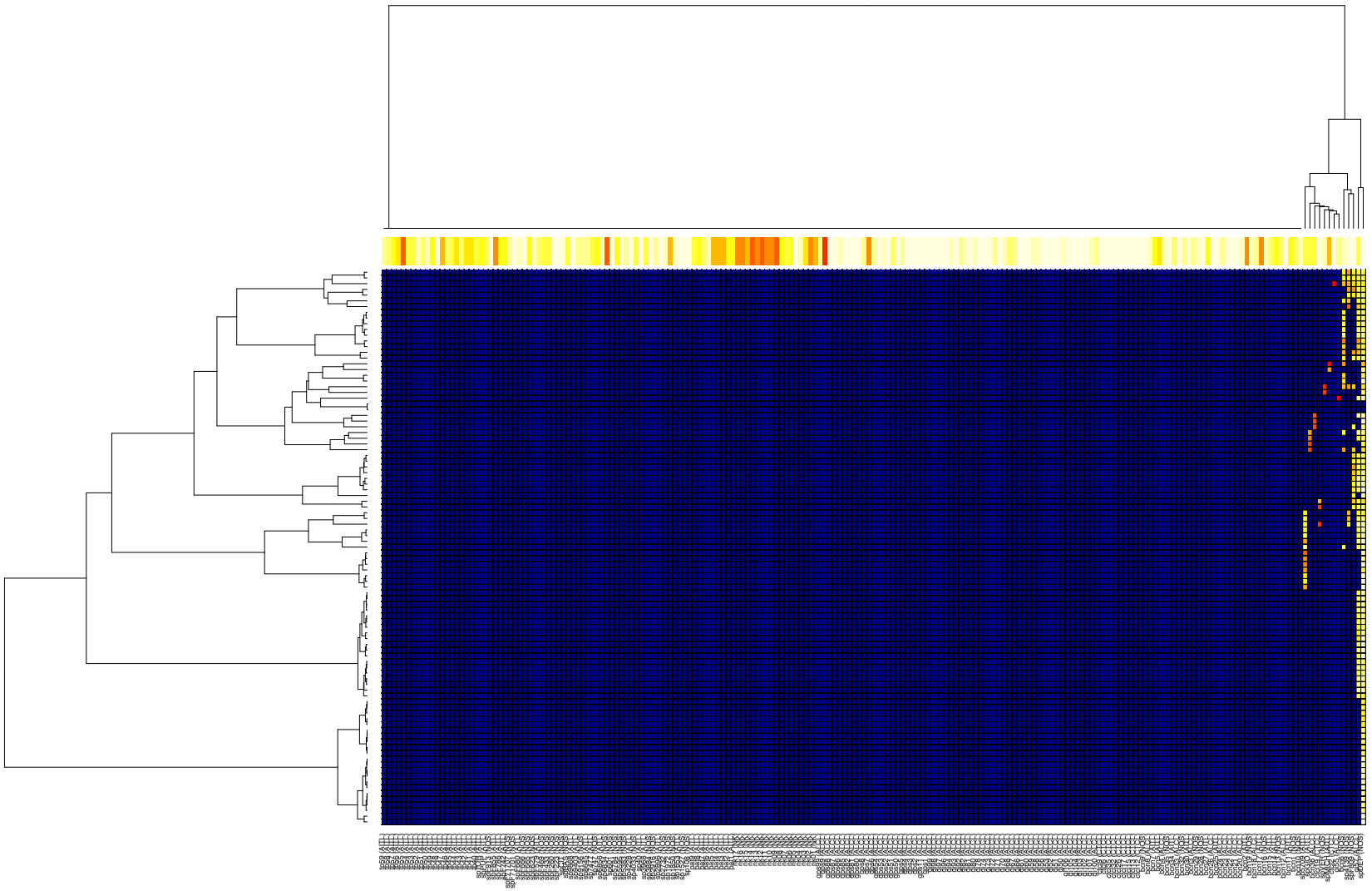


Figure 3.12: **Expression profiling of HHV6b:** Gene expression profiling of HHV6b using RefSeq annotation NC000898.

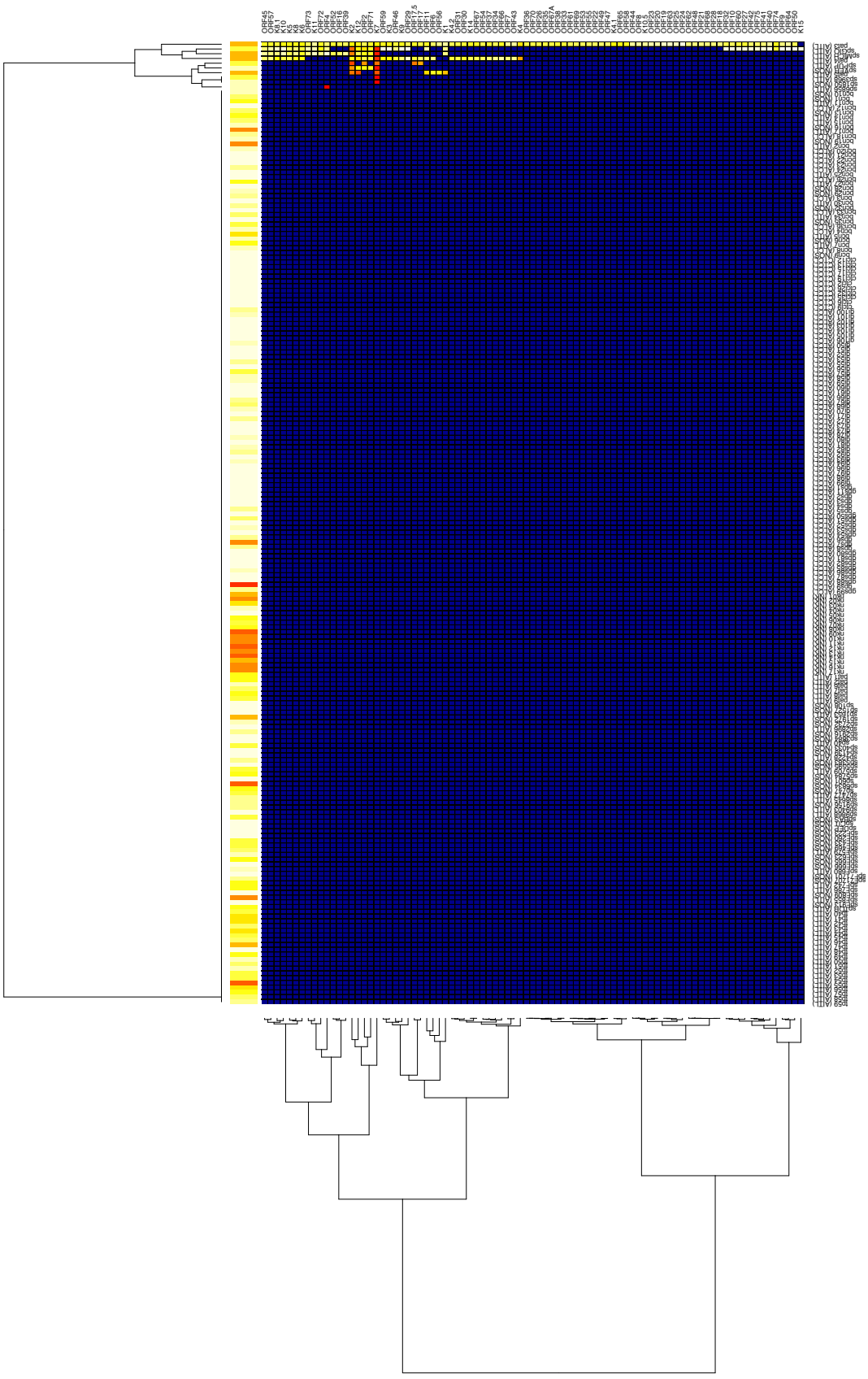


Figure 3.13: **Expression profiling of KSHV:** Gene expression profiling of KSHV using GenBank annotation GQ994935.

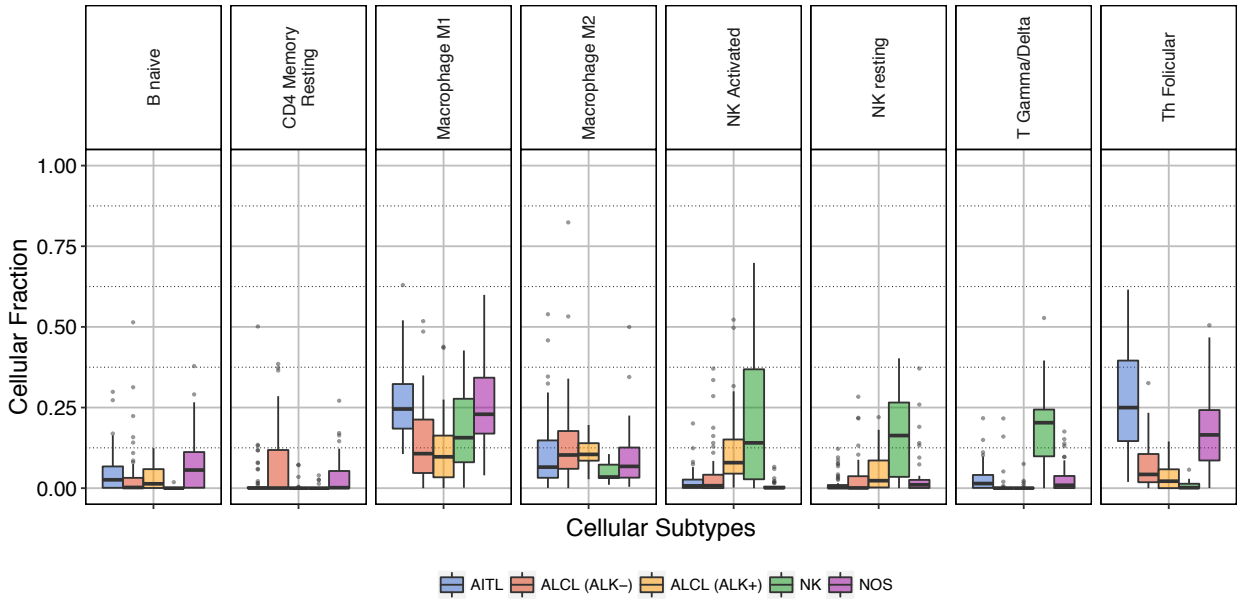


Figure 3.14: **Immune cell fractions in the microenvironment:** Of the 22 immune cell types characterized in the LM22 matrix [147], these 8 have some nonzero representation across the PTCL cohort. We observe good agreement between AITL and NOS subtypes, and we note the strong presence of macrophages (M1 polarization > M2).

cell abundance likely points to the virus playing a direct role in this malignancy as a clonal alteration. The lack of correlation between EBV and Th follicular cells indicates that EBV is likely not clonal in this subtype of PTCL, and is therefore an indirect factor. Further confirmation of EBV's microenvironmental role, as opposed to clonal role, in AITL comes from the positive correlation between viral load and infiltrating B cells (Figure 3.15).

3.5 Conclusions

Tumor associated viruses have a long history in illuminating oncogene and tumor suppressor biology. Viruses can have many roles with respect to a tumor's growth They can be a clonal infection that drives the disease or they can be a microenvironmental factor residing in non-malignant cells. We presented Pandora, a suite of modules for the simultaneous characterization of viral expression and host tumor genetics and cell composition. EBV was studied first in Burkitt lymphoma where the infection is clonal and the virus has an

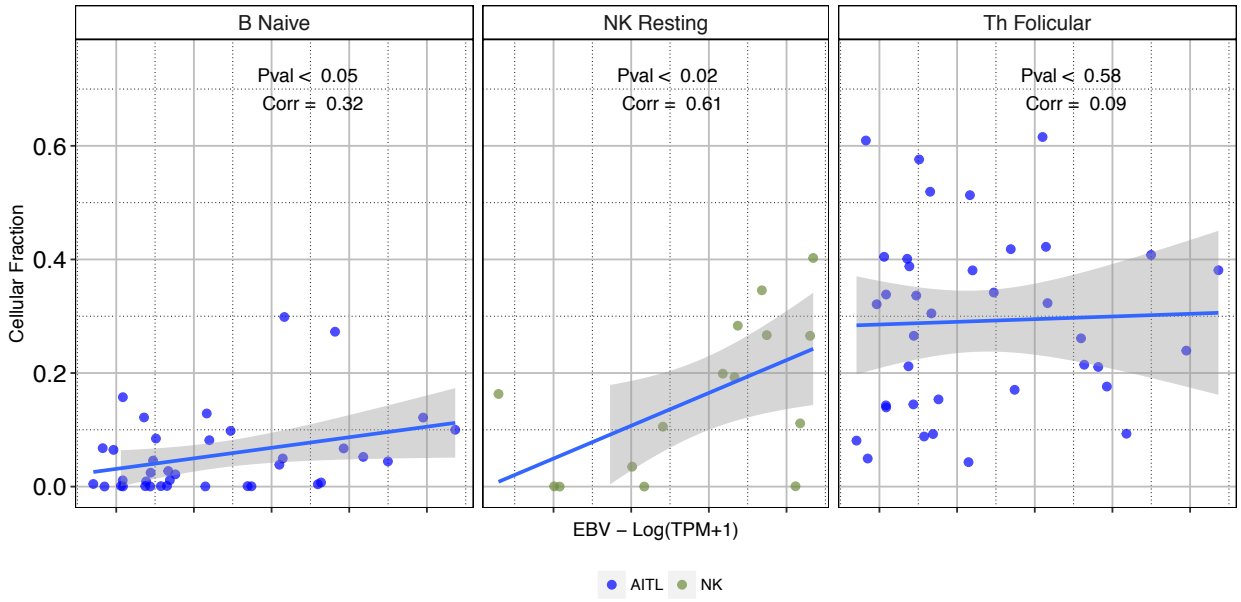


Figure 3.15: **Cellular host of EBV across subtypes:** Specific examination of the AITL and NK subtypes, which are most associated with EBV infection. The associations between viral load and cell-type fraction demonstrate a fundamental difference between the role of EBV in these two subtypes. We claim that EBV is infecting B-cells in AITL, and therefore is a micorenvironmetal factor, while in NK the virus is clearly infecting the malignant clone itself.

inverse correlation with TCF3/ID3 mutations. Then EBV was studied in peripheral T-cell lymphoma, where the infection can be either clonal (NK) or microenvironmental (AITL).

Part II

Longitudinal View of Clonal Evolution in Tree Spaces

Chapter 4

Tree-Like Evolution and Representations of Phylogeny

The previous two chapters focused a great deal on characterizing the heterogeneous landscapes of genomic alterations in cancer. We used cross-sectional sequencing data to recover specific gene fusions, point mutations, and viral infections in tumors and saw a high degree of recurrence for many of the alterations. A natural next question is how to track a genomic landscape through time, given sequencing data at multiple time points along the clonal process. To answer this question we will rely heavily on objects called phylogenetic trees. This chapter will begin with an overview of basic evolutionary theory and discuss common approaches to phylogenetic inference. After establishing the phylogenetic tree as a computable and natural representation of a clonal process, the question arises of how to represent large collections of trees. We pay considerable attention to a construction by Billera, Holmes, and Vogtmann (BHV) [19] wherein trees are mapped to points and an efficient algorithm for calculating shortest paths exists. Our own contribution begins with a definition of projective BHV space, an idea that is well-suited to the analysis of heterogeneous

Material presented in this chapter is published, wholly or in part, in: [233, 234] in collaboration with H. Khiabanian, A.J. Blumberg, R. Rabadan; and will be published in the manuscript "Phylogenetic dimensionality reduction" (in preparation) in collaboration with H. Khiabanian, A.J. Blumberg, R. Rabadan

genomic data sets in practice. By understanding the projective tree space as a join of simpler spaces we show an explicit visualization scheme for point clouds of trees on three, four, or five leaves. We next survey approaches for statistical summarization of point clouds in tree space. Finally, we end with a discussion of a structured tree dimesnionality reduction approach for mapping large phylogenies to distributions of smaller trees more amenable to visualization and statistical analysis.

4.1 Evolutionary Theory and Phylogenetic Inference

4.1.1 Tree-like evolution

Evolutionary theory was placed on a mathematical footing by the 1930's with the work of Ronald Fisher and Sewall Wright. Two classical approaches to simulating an evolving population forward in time are the Moran process [137] and the Wright-Fisher [224, 68] process. Both are instances of birth-death processes and both make the simplifying assumptions of a well mixed population, a constant population size, and no genetic recombination. The absence of recombination is perfectly appropriate for our current studies in clonal evolution, while the biological validity of the first two assumptions may be questioned in the settings of tumor growth or viral infection. The Moran process is a forward-time Markov process in which, at each step, a single genome in the population is chosen to reproduce (selected randomly in proportion to its fitness) and another genome is chosen to be eliminated (selected at random uniformly). The Wright-Fisher process is similar except that at each time step the entire population is replaced by randomly drawing from the previous generation in proportion to fitness of genotypes. The Wright-Fisher process is equivalent to N generations of the Moran process, where N is the fixed size of the population [15]. When there is no fitness advantage to either wildtype or mutant genotype, both of these processes simulate neutral evolution [15, 112]. There is value in forward-time simulations if one is interested in longitudinal sampling of evolution, since the entire population of genomes is tracked in memory and

simulations can be stopped, restarted, forked, etc. The tradeoff, however, in comparison to reverse-time simulations such as Kingman's coalescent [113], is that forward-time simulation requires more computational resources. A recent and quite promising simulation package for forward-time modeling of diallelic loci under many user-controlled evolutionary parameters (mutation rate, selection strength, number of driver genomic sites, number of passenger genomic sites, epistatic interactions, etc.) is OncoSimulR [50]. This package simulates both fixed-size and exponential growth populations, and can capture the positive fitness effects of driver mutations as well as the deleterious fitness effects of passenger mutations [133].

Though it may be obvious to the reader, genotypes sampled from these sorts of clonal simulations must fit a phylogenetic tree exactly. The transition rules for the simulations are specified in terms of parent–child relationships, meaning that any mutational profiles observed at termination of the simulation must obey a tree-like distance metric. On this point, we briefly review the conditions for additivity of a metric and also ultametricity. A dissimilarity measure $d(\cdot, \cdot)$ can be called a distance metric if it satisfies the properties that (i) $d(x, x) = 0$, (ii) $d(x, y) \geq 0$, (iii) $d(x, y) = d(y, x)$, and (iv) $d(x, z) \leq d(x, y) + d(y, z)$. A distance constitutes an additive metric if it satisfies Buneman's four point condition [25], which is illustrated in Figure 4.1. A distance matrix fits a tree if, for every quartent of 4 observations, the pairwise distances satisfy $d_{ik} + d_{kl} \leq d_{ik} + d_{jl} = d_{il} + d_{jk}$ under one of the three possible labelings $\{i, j, k, l\}$. A distance matrix fits an ultrametric tree if the condition is made more stringent to $d_{ik} + d_{kl} \leq \max(d_{ik} + d_{jl}, d_{il} + d_{jk})$.

4.1.2 Inferring phylogeny

Phylogeny is the study and inference of ancestral relationships between observed organisms. The field predates the molecular biology revolution and early phylogenetic trees were constructed on the basis of phenotypic traits rather than genetic data. Today phylogenetic trees are inferred from aligned sequences. The simplest distance metric is the Hamming distance, which counts the number (or percentage) of discordant positions between a pair of sequences.

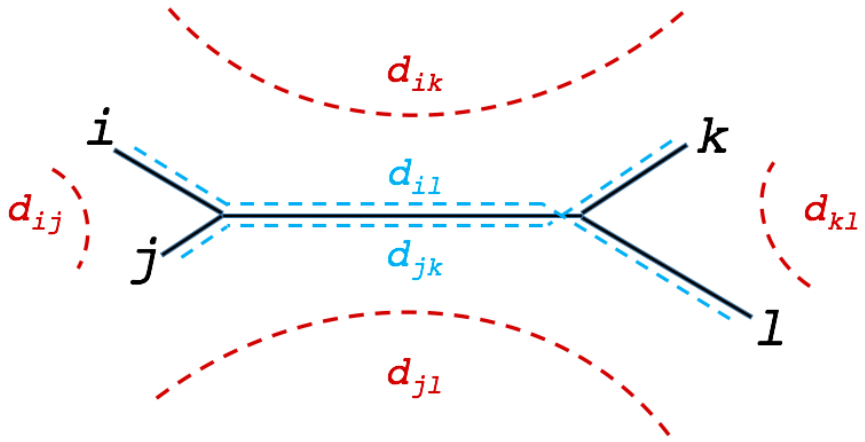


Figure 4.1: **Four point condition for additive metrics:** There are three distinct ways to partition 4 vertices into sets of two unordered pairs. Let the unordered pairs stand for the graph distance between their vertices, and suppose a set of two unordered pairs denotes the addition of the pairwise vertex distances. Buneman first noted [25] that a graph is tree $\iff$ two of the three pairwise distance sums are equal and they are at least as large as the third pairwise distance sum.

Evolutionary distances, rather than just edit distances on a string, come from considering DNA substitution models which parametrize the rates of transition between nucleotides. The earliest substitution model came from Jukes and Cantor [103], using a single mutation rate parameter μ and equal transition rates for all letters. In 1980 Kimura published a two parameter model [111] that discriminated between the rate of transition and transversion, and more highly parametrized models have been developed since. Regardless of the choice of evolutionary distance metric, sequence data rarely adhere to strict additivity. Deviations from a tree-like distance matrix can arise from back mutations at the same position, from horizontal (non-clonal) evolutionary events, or something as mundane as technical error in measurement. We do not consider biological systems undergoing horizontal evolution in this work, but measurement error and repeated mutation at the same position (especially for shorter RNA genomes) are relevant sources of noise.

A schematic of a clonally evolving population is given in Figure 4.2, along with a depiction of tree-like amplification of genomes. In the Figure we illustrate the construction of a phylogenetic tree on longitudinal sampling of the evolving population. A phylogenetic tree

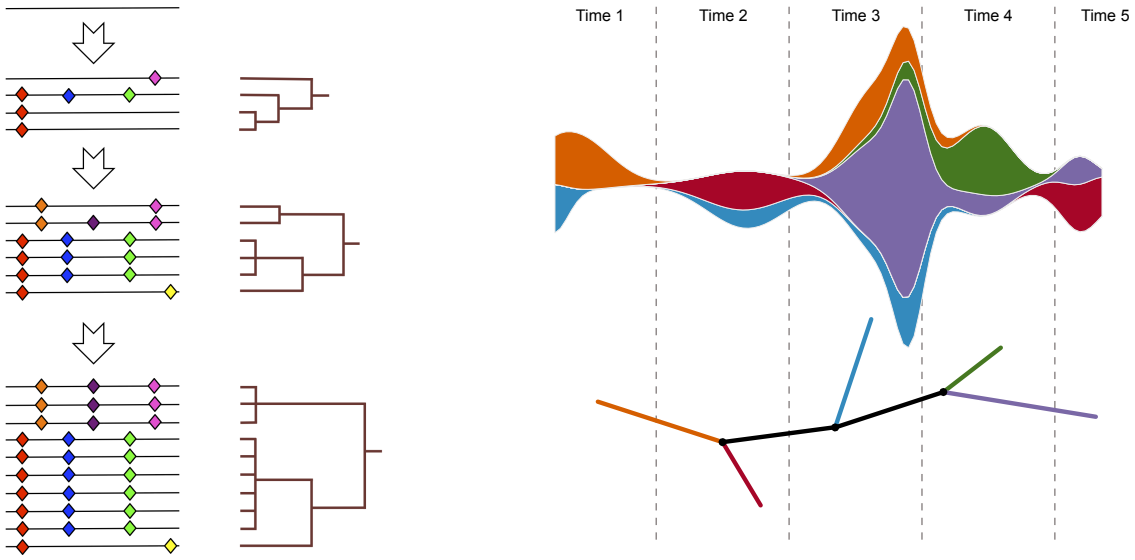


Figure 4.2: **Clonal evolution of an asexually reproducing genome:** (Left) Through acquisition of mutations, the primordial clone gives rise to a large heterogeneous population, whose evolutionary history can be accurately described by a tree. (Right) Longitudinal sampling of a clonal population permits the construction of phylogenetic trees that approximate the underlying history. Subpopulations are represented in different colors; random sampling of a particular genotype at each time point is illustrated in the color of the external branch in the tree. This tree is one of many that could be observed when sampling this population.

with m leaves is a weighted, connected graph with no cycles, having m distinguished vertices (referred to as *leaves*) of degree 1 and labeled $\{1, \ldots, m\}$. All the other vertices are of degree ≥ 3. We refer to edges that terminate in leaves as *external* edges and the remaining edges are *internal*. Phylogenetic trees must be inferred from aligned sequence data, and there are many competing approached to tree building. The character-based methods that explicitly optimize a criterion of fit include maximum parsimony and maximum likelihood. The distance-based methods are clustering algorithms that do not explicitly pose the search for the best fit tree as an optimization of a particular quantity. The two most prominent distance methods are Unweighted Pair Group with Arithmetic Mean (UPGMA) and the neighbor joining method of Saitou and Nei [180]. UPGMA makes the strong assumption that the data obey a molecular clock, or equivalently that the distance matrix is ultrametric. The phylogenetic trees inferred in this thesis use the neighbor joining method, as it is guaranteed to recover the correct tree if one exists and is computationally quite efficient.

4.2 Geometric Representations of Phylogenetic Trees

Comparison between trees requires a definition of distance, and many competing tree distances can be found in the literature. The Robinson-Foulds metric [177] computes the symmetric difference between the partitions defined be each tree's internal edges. Nearest neighbor interchange considers the swapping of vertices in 4-leaf subtrees and is the simplest instance of a rearrangement procedure for quantifying dissimilarity between trees [63]. More complex tree arrangement procedures exist such as subtree prune and regrafting, or tree bisection and reconnection, but will not be described here. A general theme shared by many of these tree distance is that they prioritize tree topology and do not account for edge lengths. In the setting of phylogenetic trees the edge lengths have an important interpretation, as the degree of evolutionary divergence between samples. We therefore seek a distance metric and an associated space of trees that captures differences in both the connectivity as well as edge lengths between phylogenies.

The foundation of our framework for analysis is the *metric geometry* of the space of phylogenetic trees [19]. The purpose of this section is to review in detail the spaces of phylogenetic trees that we work with and their geometric structure. We begin with a rapid review of the geometry of geodesic metric spaces and the theory of cubical complexes. We then review the definition and properties of the Billera-Holmes-Vogtmann metric space of phylogenetic trees and its metric geometry, following the the excellent original treatment. Finally, we discuss the properties of projective versions of tree spaces which are relevant for some of the biological applications.

4.2.1 A rapid review of metric geometry

In this subsection we quickly explain the foundations of metric geometry. See [23, 26] for comprehensive textbooks on the subject. A metric space (X, d) is a set X equipped with a distance function $d\colon X \times X \to \mathbb{R}^{\geq 0}$ having the properties that $d(x, y) = d(y, x)$, $d(x, y) = 0$

if and only if $x = y$, and $d(x, z) \leq d(x, y) + d(y, z)$ for all $x, y, z \in X$ (the triangle inequality). Although metric spaces often arise in contexts in which there is not an evident notion of geometric structure, it turns out that under very mild hypotheses a metric space (X, d) can be endowed with structures analogous to those arising on Riemannian manifolds. A metric space is a length space if the distance $d(x, y)$ is realized as the infimum of the lengths of paths joining x and y. A length space (X, d) is a geodesic metric space if any two points x and y can be joined by a path with length precisely $d(x, y)$. A key insight of Alexandrov is that *curvature* makes sense in any geodesic metric space [7].

The idea is that the curvature of a space can be detected by considering the behavior of the area of triangles, and triangles can be defined in any geodesic metric spaces. Specifically, given points p, q, r, we have the triangle $T = [p, q, r]$ with edges the paths that realize the distances $d(p, q)$, $d(p, r)$, and $d(q, r)$. The connection between curvature and area of triangles comes from the observation that given side lengths $(\ell_1, \ell_2, \ell_3) \subset \mathbb{R}^3$, a triangle with these side lengths on the surface of the Earth is "fatter" than the corresponding triangle on a Euclidean plane. To be precise, we consider the distance from a vertex of the triangle to a point p on the opposite side — in a fat triangle, this distance will be larger than in the the corresponding Euclidean triangle (and smaller in a thin triangle).

Given a triangle $T = [p, q, r]$ in (X, d), we can find a corresponding triangle $\tilde{T}$ in Euclidean space with the same edge lengths. Given a point z on the edge $[p, q]$, a comparison point in $\tilde{T}$ is a point $\tilde{z}$ on the corresponding edge $[\tilde{p}, \tilde{q}]$ such that $d_E(\tilde{z}, \tilde{p}) = d(z, p)$ and $d_E(\tilde{z}, \tilde{q}) = d(z, q)$. (Where here d_E denotes the Euclidean metric.) We say that a triangle T in M satisfies the CAT(0) inequality if for every such pair $(z, \tilde{z})$, we have $d(r, z) \leq d_E(\tilde{r}, \tilde{z})$. If every triangle in M satisfies the CAT(0) inequality then we say that M is a CAT(0) space.

More generally, let M_κ denote the unique two-dimensional Riemannian manifold with curvature κ. The diameter of M_κ will be denoted D_κ. Then we say that a geodesic metric space M is CAT(κ) if every triangle in M with perimeter $\leq 2D_\kappa$ satisfies the inequality above for the corresponding comparison triangle in M_κ. If $\kappa' \leq \kappa$, any $CAT(\kappa')$ space is also

CAT(κ). A n-dimensional Riemannian manifold M that is sufficiently smooth has sectional curvature $\leq \kappa$ if and only if M (regarded as a metric space) is CAT(κ). For example, Euclidean spaces are CAT(0), spheres are CAT(1), and hyperbolic spaces are CAT(-1).

As described, CAT(κ) is a global condition; we will say that a metric space (X, d) is locally CAT(κ) if for every x there exists a radius r_x such that $B_{r_x}(x) \subseteq X$ is CAT(κ). For example, the flat torus (obtained by identifying opposite edges in a rectangle) is locally CAT(0) but not globally CAT(0). The Cartan-Hadamard theorem implies that a simply-connected metric space that is locally CAT(0) is also globally CAT(0).

A remarkably productive observation of Gromov is that many geometric properties of Riemannian manifolds are shared by CAT(κ) spaces. In particular, CAT(κ) spaces with $\kappa \leq 0$ (referred to as *non-positively curved metric spaces*) admit unique geodesics joining each pair of points x and y, balls $B_\epsilon(x)$ are convex and contractible for all x and $\epsilon \geq 0$, and midpoints of geodesics are well-behaved. As a consequence, there exist well-defined notions of mean and variance of a set of points, and more generally one can develop some of the foundations of classical statistics, as we review below in Section 4.4.

4.2.2 Cubical complexes and their links

In this subsection, we review the theory of cubical complexes, which provide a rich source of examples of CAT(0) metric spaces (again, see [26] or [23] for textbook treatments). It is in general very difficult to determine for an arbitrary metric space whether it is CAT(κ) for any given κ. Even for finite polyhedra where the metric is induced from the Euclidean metric on each face, this problem does not have a general solution. The important of cubical complexes in this context comes from an effective criterion for determining if they are non-positively curved (i.e., CAT(0)).

Let $I^n \subseteq \mathbb{R}^n$ denote the n-dimensional unit cube $[0, 1] \times \ldots \times [0, 1]$, regarded as inheriting a metric structure from the standard metric on $\mathbb{R}^n$. A codimension k face of the cube I^n is determined by fixing k coordinates to be in the set $\{0, 1\}$. A cubical complex is a metric

space obtained by gluing together cubes via the data of isometries of faces, subject to the condition that two cubes are connected by at most a single face identification and no cube is glued to itself. The metric structure is the length metric induced from the Euclidean metric on the cubes, i.e., the distance between x and y is the infimum of the lengths over all paths from x to y that can be expressed as the union of finitely many segments each contained within a cube. When the cubical complex C is finite or locally finite, results of Bridson [22] and Moussong [140] imply that C is a complete geodesic metric space.

Gromov gave a criterion for a cubical complex to be CAT(0) that is often possible to check in practice. In order to explain this criterion, we need to review the notion of the link of a vertex in a cubical complex.

Fix a vertex v in a cubical complex C and a cube $C_i \cong I^m \subseteq C$ such that v is a vertex of C_i. For fixed $\epsilon > 0$, the all-right spherical simplex associated to (C_i, v) is the subset

$$S(C_i, v) = \{z \in C_i \,|\, d(z, v) = \epsilon\}.$$

The set $S(C_i, v)$ has a metric induced by the Euclidean angle metric. The faces of $S(C_i, v)$ are defined as the intersections of $S(C_i, v)$ with faces of C_i; equivalently, these are the all-right spherical simplexes associated to faces of C_i. The collection of all-right spherical simplices for all pairs (C_i, v) forms a polyhedral complex with metric given by the length metric induced from the angle metrics; this is referred to as a spherical complex. Forgetting the metric structure, the all-right spherical simplices also form an abstract simplicial complex. (Recall that an abstract simplicial complex is simply a set of subsets of a set V that is closed under passage to subsets.)

The link of a vertex v in a cubical complex C is the spherical complex obtained as the subset

$$L(v) = \{z \in C \,|\, d(z, v) = \epsilon\},$$

for fixed $0 < \epsilon < 1$. Gromov's criterion now states that the cubical complex C is locally CAT(0) if and only if the link is CAT(1) or the abstract simplicial complex underlying the

link is flag. (Recall that a flag complex is a simplicial complex in which a k-simplex is in the complex if and only its 1-dimensional faces are in the complex.)

As an easy application of Gromov's criterion, we conclude the section by showing the standard result that the Cartesian product of locally CAT(0) cubical complexes is itself a locally CAT(0) cubical complex. Let X and Y be cubical complexes that are CAT(0). Since $I^n \times I^m \cong I^{n+m}$, it is clear that $X \times Y$ has the structure of a cubical complex. The set of vertices of $X \times Y$ is given by the product of the sets of vertices of X and Y respectively. The link of a vertex (v, v') in $X \times Y$, regarded as an abstract simplicial complex, is the join of $L(v)$ and $L(v')$, which we denote $L(v) * L(v')$ (the join of complexes S_1 and S_2 is obtained by considering all pairwise unions of elements of S_1 and S_2). Finally, since the join of flag complexes is easily seen to be a flag complex, Gromov's criterion now implies that $X \times Y$ is locally CAT(0).

4.2.3 The BHV spaces of phylogenetic trees

The space BHV_m of isometry classes of rooted phylogenetic trees with m-labelled leaves where the nonzero weights are on the internal branches was introduced and studied by Billera, Holmes, and Vogtmann [19]. The space BHV_m is constructed by gluing together $(2m-3)!!$ positive orthants $\mathbb{R}^m_{\geq 0}$; each orthant corresponds to a particular tree topology, with the coordinates specifying the lengths of the edges. A point in the interior of an orthant represents a binary tree; if any of the coordinates are 0, the tree is obtained from a binary tree by collapsing some of the edges. We glue orthants together such that a (non-binary) tree is on the boundary between two orthants when it can be obtained by collapsing edges from either tree geometry. Put another way, two tree topologies are adjacent when they are connected by a *rotation*, i.e., one topology can be generated from the other by collapsing an edge to length 0 and then expanding out another edge from the incident vertex.

The metric on BHV_m is induced from the standard Euclidean distance on each of the orthants. For two trees t_1 and t_2 which are both in a given orthant, the distance $d_{\mathrm{BHV}_m}(t_1, t_2)$

is defined to be the Euclidean distance between the points specified by the weights on the edges. For two trees which are in different quadrants, there exist (many) paths connecting them which consist of a finite number of straight lines in each quadrant. The length of such a path is the sum of the lengths of these lines, and the distance $d_{\mathrm{BHV}_m}(t_1, t_2)$ is then the minimum length over all such paths. For many points, the shortest path goes through the "cone point", the star tree in which all internal edges are zero.

Allowing potentially nonzero weights for the m external leaves corresponds to taking the cartesian product with an m-dimensional orthant. We will focus on the space

$$\Sigma_m = \mathrm{BHV}_{m-1} \times \mathbb{R}^m_{\geq 0},$$

which we refer as the evolutionary moduli space (the $m-1$ index arises from the fact that we consider unrooted trees.) There is a metric on Σ_m induced from the metric on BHV_{m-1}. Specifically, for a tree t, let $t(i)$ denote the length of the external edge associated to the vertex i. Then

$$d_{\Sigma_m}(t_1, t_2) = \sqrt{\left(d_{\mathrm{BHV}_{m-1}}(\bar{t}_1, \bar{t}_2)\right)^2 + \sum_{i=1}^{m} (t_1(i) - t_2(i))^2},$$

where $\bar{t}_i$ denotes the tree in BHV_{m-1} obtained by forgetting the lengths of the external edges (e.g., see [158]). As explained in [19, §4.2], efficiently computing the metric on Σ_m is a nontrivial problem, although there exists a polynomial-time algorithm [158].

The main result of Billera, Holmes, and Vogtmann is that the length metric on BHV_n endows this space with a (global) CAT(0) structure (Figure 4.3. By subdividing each orthant into cubes in the evident fashion, Σ_m is naturally a cubical complex where the metric we have described is the one induced from the Euclidean metric on the cubes; a straightforward combinatorial analysis of the link of Σ_m implies the result via Gromov's criterion. In addition, Σ_m is clearly a complete and separable metric space; any tree can be approximated by a sequence of trees in the same orthant that have rational edge lengths.

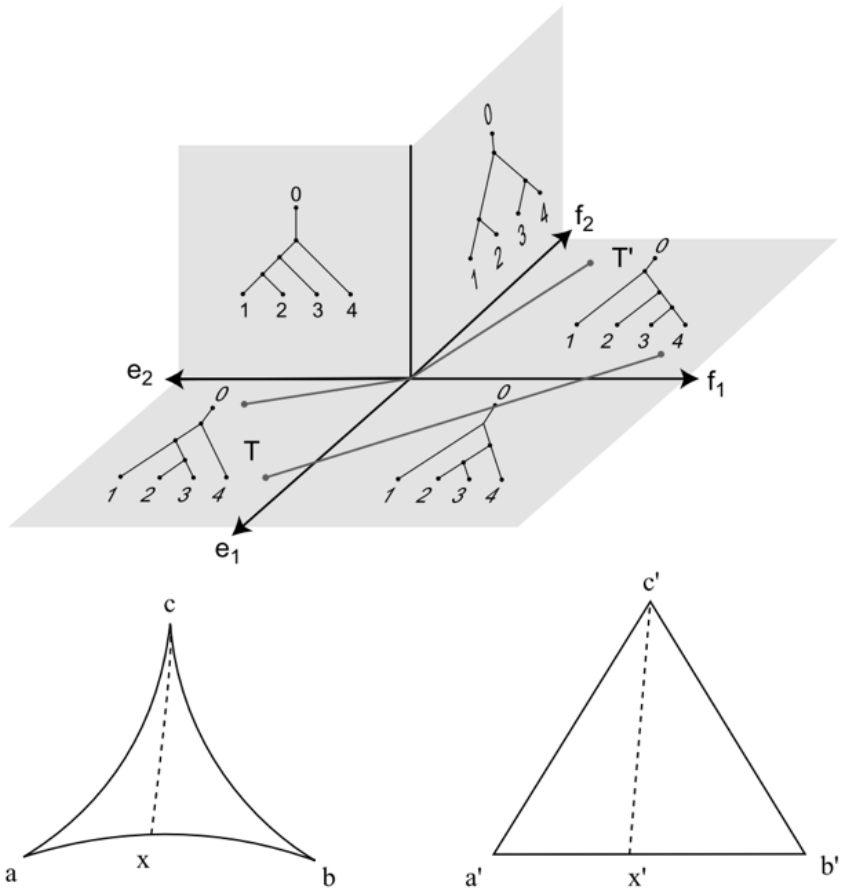


Figure 4.3: **BHV orthants and the CAT(0) condition:** (Top) BHV space is constructed by appropriate gluing of Euclidean orthants along codimension 1 boundaries. The gluing respects the rotational distances between pairwise tree topologies. All orthants meet at a single point, the origin. A shortest path between trees in different orthants that goes through the origin is termed a "cone path," and we see illustrated that geodesics may or not be cone. (Bottom) A useful aspect of BHV space is the CAT(0) property which globally guarantees that any triangle formed will be at least as thin as the analogous Euclidean triangle. Images reproduced with permission from [19].

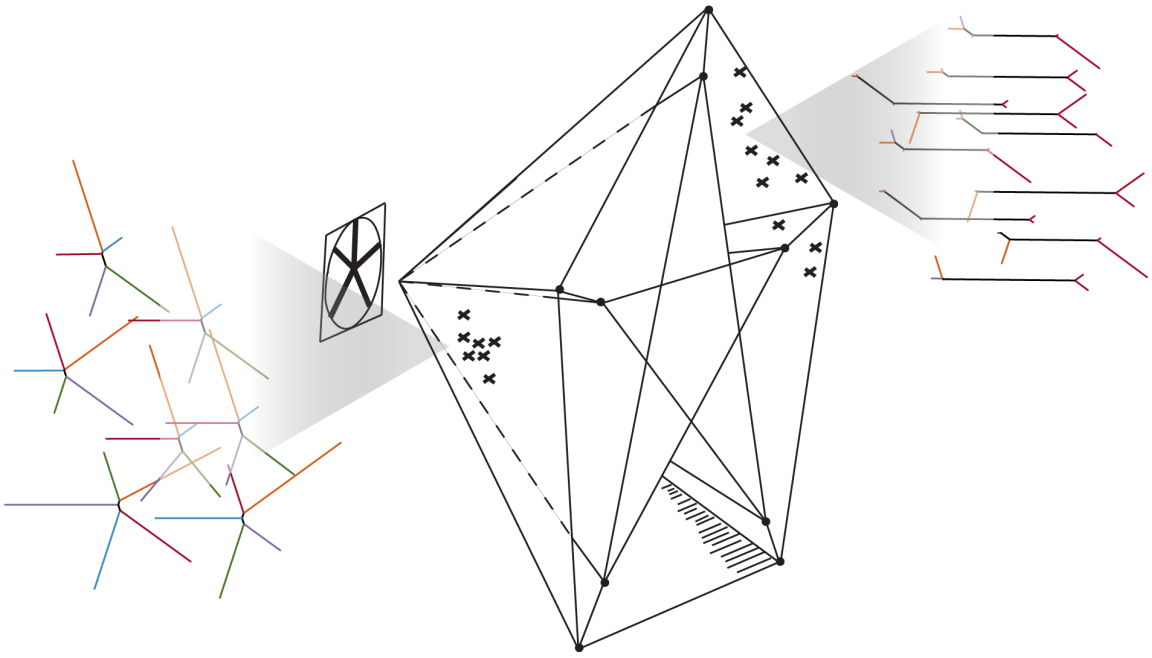


Figure 4.4: **Moduli space of phylogenetic trees describing clonal evolution:** Collections of trees are points in a metric space, forming a point cloud. Trees with the same topology live in the same orthant, and crossing into an adjacent orthant corresponds to a tree rotation. Points closer to the vertex of the cone have relatively little internal branch length, while points near the base of the cone have little weight in the external branches.

4.3 Projective BHV Space

In evolutionary applications, we are often interested in classifying and comparing distinct behaviors by understanding the relative lengths of edges: rescaling edge lengths should not change the relationship between the branches [233]. Motivated by this consideration, we define $\mathbb{P}\Sigma_m$ to be the subspace of Σ_m consisting of the points $\{t_i\}$ in each orthant for which the constraint $\sum_i t_i = 1$ holds.

We denote the space of trees with internal edges of fixed length by τ_{m-1}. The space of m external branches whose lengths sum to 1 is the standard $m-1$ dimensional simplex Δ_{m-1} in $\mathbb{R}^m$. The constraint that the length of internal branches plus the external branches sum to 1 implies that

$$\mathbb{P}\Sigma_m = \tau_{m-1} \star \Delta_{m-1},$$

where here $\star$ denotes the join of two spaces, using the Milnor model of the join. We can also

describe $\mathbb{P}\Sigma_m$ as the link on the origin in Σ_m.

There are various possible natural metrics to consider on $\mathbb{P}\Sigma_m$. The simplest way to endow $\mathbb{P}\Sigma_m$ with a metric is to use the induced intrinsic metric specified by paths in Σ_m constrained to lie entirely within $\mathbb{P}\Sigma_m$. From the perspective of metric geometry, the characterization of $\mathbb{P}\Sigma_m$ as the link of the origin endows it with a "spherical" metric, and Gromov's criteria imply that with this metric, $\mathbb{P}\Sigma_m$ is a CAT(1) space. (Alternatively, $\mathbb{P}\Sigma_m$ is the spherical join of τ_{m-1} and the spherical realization of the Δ_{m-1}; since τ_{m-1} and Δ_{m-1} are CAT(1), so is their spherical join [23, p. II.3.15].) The theory of polyhedral complexes implies that in either case $\mathbb{P}\Sigma_m$ is a complete geodesic metric space [23, p. I.7.19], and it is evidently separable.

Moreover, with the induced intrinsic metric $\mathbb{P}\Sigma_m$ is in fact a CAT(0) space; although τ_{m-1} has points which are not connected by unique geodesics (see Section 4.3.1 below for a more detailed discussion), the join with Δ_m introduces a new "cone direction" that changes the geometry.

Theorem 4.3.1. *The projective moduli space $\mathbb{P}\Sigma_m$ endowed with the intrinsic metric is a CAT(0) space.*

Proof. First, recall that, for any $k \geq 0$, $\tau_{m-1} \star \Delta_k$ is isomorphic to

$$\tau_{m-1} \star \underbrace{\Delta_0 \star \Delta_0 \ldots \star \Delta_0}_{k+1}.$$

To see this, observe that a point in the join of $\tau_{m-1} \star \Delta_{k-1}$ with Δ_0 can be described as a tuple

$$((wt_0, \ldots, wt_n), (wx_0, \ldots, wx_{k-1}), 1-w),$$

where $\sum_{i=0}^{n} t_i + \sum_{i=0}^{k-1} x_i = 1$ and $w \in [0,1]$. This data is clearly equivalent to a tuple

$$((t_0, \ldots, t_n), (x_0, \ldots, x_{k-1}, x_k))$$

where $\sum_{i=0}^{n} t_i + \sum_{i=0}^{k} x_i = 1$.

The fact that BHV_{m-1} is CAT(0) implies that the cone $\tau_{m-1} \star \Delta_0$ is CAT(0), and from this it follows by induction that $\tau_{m-1} \star \Delta_{m-1}$ is also CAT(0). □

To compute the intrinsic metric on $\mathbb{P}\Sigma_m$, we use ϵ-nets and a local-to-global construction. Recall that a set of points S in a metric space (X, ∂) is an ϵ-net if for every $z \in X$, there exists $q \in S$ such that $\partial(z, x) < \epsilon$. For a compact metric space equipped with a probability measure such that all non-empty balls in the metric space have nonzero measure, we can produce an ϵ-net by sampling. More precisely, it is straightforward to show that given a finite collection of measurable sets $\{A_1, A_2, \ldots, A_k\}$ and a probability measure μ on $\cup_i A_i$ such that $\mu(A_i) \geq \alpha > 0$, then given at least

$$\frac{1}{\alpha}\left(\log k + \log(\frac{1}{\delta})\right)$$

samples, with probability $1 - \delta$ there is at least one sample in every A_i [149, p. 5.1].

Next, suppose that we have a metric space (X, ∂) where there exists a constant κ such that if $\partial(x, y) < \kappa$, it is easy to compute $\partial(x, y)$. An algorithm for approximating ∂ on all of X is then to take a dense sample $S \subset X$, form the graph G with vertices the points of S and edges between x and y when $\partial(x, y) < \kappa$, and define the distance between x and y in X to be the graph metric on G between the nearest points to x and y in S. This distance can be efficiently computed using Dijkstra's algorithm [51].

When S is an ϵ-net for sufficiently small ϵ relative to κ, we can describe the quality of the resulting approximation to ∂ [18, Thm. 2]. In particular, if $\epsilon < \frac{\kappa}{4}$, then

$$\partial(x, y) \leq \partial_G(x, y) \leq (1 + 4\frac{\delta}{\epsilon})\partial(x, y).$$

Putting this all together, to approximate the metric on $\mathbb{P}\Sigma_m$ we take the union of ϵ-nets on all of the simplices (including the faces) and form a κ-approximation ∂_G as above. In practice, the required density of samples is determined by looking at when the approximation converges (i.e., when the change in distances drops below a specified precision bound). We sample densely on each simplex representing a tree topology on m leaves, and explicitly

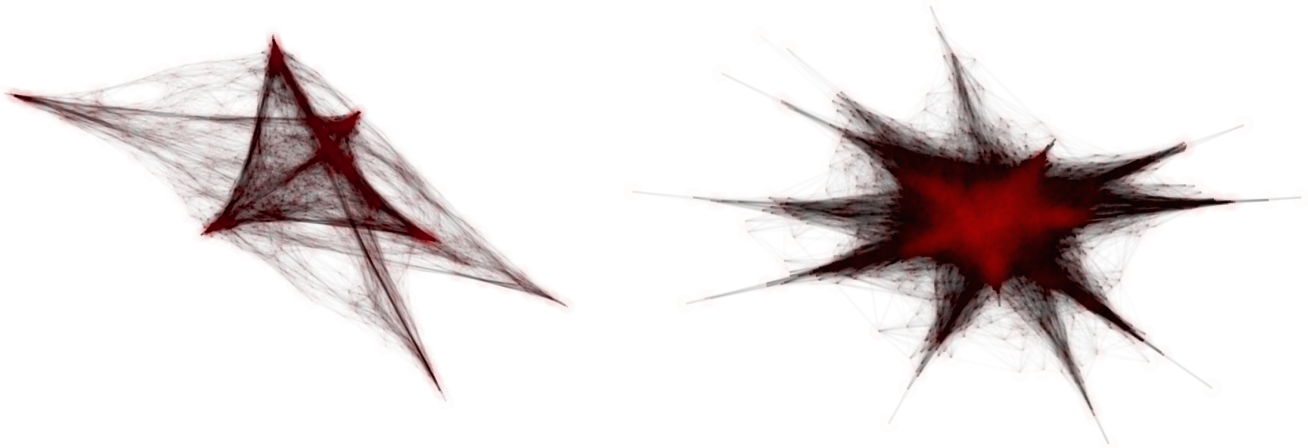


Figure 4.5: **Discrete approximation of $\mathbb{P}\Sigma_m$, projective BHV space:** Unit norm phylogenetic trees on m leaves are densely sampled and used as vertices in a graph weighted by pairwise BHV distances. All edges above a threshold distance are removed, with the threshold determined as the smallest value that maintains a single connected component. We call this graph the ϵ-net. The projective distance is then defined as the graph distance along the ϵ-net, and we visualize a force directed layout of the graph. On the left we represent $\mathbb{P}\Sigma_4$, and on the right $\mathbb{P}\Sigma_5$. The case of four leaves is more easily appreciated as a join space, as described earlier.

include certain key singular points of the projective space. Figure 4.5 contains two visualizations of the projective space, using force-directed layouts of the ϵ-nets constructed on 4–leaved and 5–leaved trees respectively.

Our definition of $\mathbb{P}\Sigma_m = \tau_{m-1} \star \Delta_{m-1}$ as a join space implies an intuitive visualization scheme for point clouds of rescaled trees. The topological join of two spaces X, Y will produce a space of dimension $\dim X + \dim Y + 1$, since it is defined as $X \times Y \times I/\sim$, the direct product of the two input spaces and the unit interval I quotienting by an equivalence relation. Specifically, the equivalence is $(x, y_1, 0) \sim (x, y_2, 0) \quad \forall x \in X \quad \forall y_1, y_2 \in Y$ and $(x_1, y, 1) \sim (x_2, y, 1) \quad \forall x_1, x_2 \in X \quad \forall y \in Y$. We can therefore visualize $\mathbb{P}\Sigma_m$ as a pair of plots in which each phylogenetic tree is represented as a point in the internal-branch specifying plot and a point in the external-branch specifying plot. The external branches of a rescaled 4-tree can be plotted as a point in tetrahedron, while a rescaled 5-tree requires two tetrahedral shadows of pentachoron. The internal branches of a rescaled tree can be

plotted on a Boardman space augmented with the unit-length 'radial' coordinate of the join. For 4-trees this corresponds to three unit-length line segmenting meeting at a shared origin, while for 5-trees this corresponds to fifteen planes, arranged on a Petersen graph, meeting at a share origin. The height of trees off the Petersen graph, in the radial coordinate, can be easily encoded by the size of points plotted. We add more schematic diagrams of this visualization strategy in the following chapter before applying it to genomic data sets.

4.3.1 The size of $\mathbb{P}\Sigma_m$

We now describe the size of $\mathbb{P}\Sigma_m$ (see also [19, p. 3.3] for a related discussion). For simplicity, we will focus on the link in BHV_n, which we will denote by $\mathcal{P}_n$, and temporarily ignore the join with the simplex coming from the external edge lengths. Observe that adjacent top-dimensional simplices in $\mathcal{P}_n$ differ by a *rotation* of tree topologies, where a rotation collapses an internal edge and then expands out from the resulting node. Next, recall that the homotopy type of $\mathcal{P}_n$ is a wedge of $(n-1)!$ spheres of dimension $(n-3)$ [176] (and see also [49, Thm. 6]). Moreover, we can explicitly describe these spheres, as follows.

As discussed in [49, Prop. 1] and [19, §3.1], the boundary of the dual polytope to the standard associahedron on n letters (parametrizing parenthesizations of n terms) embeds in many different ways into $\mathcal{P}_n$. Following [49], let us denote this boundary by $\mathcal{K}_n$. Explicitly $\mathcal{K}_n$ is a simplicial sphere of dimension $(n-3)$ where a k-simplex corresponds to a planar rooted tree with n leaves and $k+1$ internal edges. Then the homotopy type of $\mathcal{P}_n$ can be described in terms of various embedded copies of $\mathcal{K}_n$. As a consequence, to understand the size of $\mathcal{P}_n$, we need to compute the diameter of $\mathcal{K}_n$.

For convenience, we describe this diameter in terms of counts of simplices; the actual value can then be obtained by multiplying by the diameter of a simplex. In this guise, the problem is an old one which can be described in many different forms, perhaps most relevantly as the computation of maximal rotation distances between binary trees.

The main result here is that, for unrooted trees on n leaves, the diameter of $\mathcal{K}_n$ is

$2n - 8$ for $n > 11$ [166]; this bound was established asymptotically (for sufficiently large but indeterminate n) in [192]. For smaller values, we have the following table (taken from [192, §2.3]) of explicit values:

Table 4.1: Diameters for $\mathcal{K}_n$ for small values of n

4	5	6	7	8	9	10	11
2	4	5	7	9	11	12	15

Warning 4.3.2. *The results given in [166] and [192] differ slightly from the formula above and from each other due to divergent choices of indexing convention.*

More generally, the maximum rotation distance between labelled trees on n leaves is $O(n \log n)$ [193]. Of course, the cone point associated to the join with standard simplex means that the size is considerably smaller.

4.4 Statistics in Tree Space

Our motivation for using the metric geometry of Σ_m and $\mathbb{P}\Sigma_m$ comes from the problems of describing and comparing collections of trees generated from experimental data. Regarding such collections as samples from distributions on the evolutionary moduli spaces, we are interested in basic statistical inference — estimating parameters describing these distributions and determining if two samples came from the same or different distributions. More generally, we would like to understand the kinds of distributions that can arise in evolutionary moduli spaces. We are also interested in clustering and classification (i.e., unsupervised and supervised learning) problems in this context. Given a set of unlabeled samples, we want to infer clusters of points that have similar clinical outcomes. Given a set of labeled samples, we want to produce classifiers that can assign labels to new points in order to predict clinical outcomes. In this section, we will review available tools for these kinds of problems.

4.4.1 Statistics for distributions in Σ_m and $\mathbb{P}\Sigma_m$

In order to study probability distributions in evolutionary moduli spaces, it is necessary to have reasonable notions of moments of the distribution, expectation of random variables, and analogues of the law of large numbers. Since Σ_m and $\mathbb{P}\Sigma_m$ are CAT(0) spaces, points are connected by unique geodesics and there is a sensible notion of a centroid of a collection of points. Discussion of statistical inference in Σ_m was initiated in [19], and subsequently Holmes has written extensively on this topic [89, 89, 90] (and see also [65]). More generally, Sturm explains how to study probability measures on general CAT(0) spaces [199]. He shows that there are reasonable notions of moments of distribution, expectation of random variables, and analogues of the law of large numbers on CAT(0) spaces.

Definition 4.4.1. *Given a fixed set of n trees $\{T_0, \dots T_{n-1}\} \subseteq \Sigma_m$, the Fréchet mean T is the unique tree that minimizes the quantity*

$$E = \sum_{i=0}^{n-1} d_{\Sigma_m}(T_i, T)^2.$$

The variance of T is the ratio $\frac{E}{n}$.

Sturm provides an iterative procedure for computing the mean and variance of a set of points in Σ_m, and by exploiting the local geometric structure of Σ_m, Miller, Owen, and Provan produce somewhat more efficient algorithms for computing the mean [135]. Furthermore, Sturm proves versions of Jensen's inequality and the law of large numbers in this context. The situation for the central limit theorem is less satisfactory. Barden, Le, and Owen study central limit theorems for Fréchet means in Σ_m [13]; as they explain, the situation exhibits non-classical behavior and the limiting distributions depend on the codimension of the simplex in which the mean lies. Finally, there has been some work on principal components analysis (PCA) in Σ_m [154].

However, in contrast to classical statistics on $\mathbb{R}^n$, we do not know many sensible analytically-defined distributions on the evolutionary moduli spaces. Billera-Vogtmann-Holmes briefly

introduce a family of Mallows distribution on Σ_m with density function

$$x(t) = \kappa e^{\alpha d_{\Sigma_m}(t_1,t)}$$

for fixed $t_1 \in \Sigma_m$, and an analogous family can be defined on $\mathbb{P}\Sigma_m$. Sampling from these distributions is not easy; in general, the behavior of distributions on Σ_m and sampling algorithms is somewhat perverse due to the pathological behavior near the origin due to the exponential growth in the mass of an ϵ ball. A much more tractable source of distributions on Σ_m and $\mathbb{P}\Sigma_m$ arise from resampling from a given set of empirical data points.

4.4.2 Distributions in Σ_m and $\mathbb{P}\Sigma_m$ via distributions in $\mathbb{R}^n$

One way to grapple with the difficulties in dealing with distributions on Σ_m and $\mathbb{P}\Sigma_m$ is to instead study associated projections into distributions on Euclidean space. The advantage of this approach is evident; we are now in a setting where the theory of moments, the central limit theorem, and asymptotic consistency for resampling procedures are all very familiar. Of course, it is important to keep in mind that the moments derived in this setting will reflect the geometry of the evolutionary moduli space is complicated ways, and inverses to the projections will not usually exist. Nonetheless, for purposes of many kinds of statistical tests (e.g., hypothesis testing about distributions generating observed samples), this approach can be very effective. There are a number of natural ways to map metric measure spaces into Euclidean space; in this section, we discuss several strategies derived from the use of the metric.

An intrinsic map comes from looking at the distance distribution on $\mathbb{R}$ induced by ∂_M. Specifically, given a Borel distribution Ψ on (M, ∂_M), the product distribution $\Psi \otimes \Psi$ on $M \times M$ induces a distibution on $\mathbb{R}$ via ∂_M. Applying this construction to the empirical measure on a finite sample yields the empirical distance distribution. More generally, for any fixed n, we can consider the distribution on $\mathbb{R}^{n^2}$ induced by taking the product distribution $\Psi^{\otimes n}$ on $M^{\times n}$ and applying ∂_M to produce the $n \times n$ matrix of distances. Gromov's "mm-reconstruction

theorem" showed that in the limit as $n \to \infty$, the distance matrix distributions completely characterize the distribution Ψ on (M, ∂_M) [84]. Once again, given a sufficiently large finite sample, we can construct the empirical distance matrix distributions for any fixed n.

Another approach involves choosing a fixed set of n landmarks and considering the vector of distances from a fixed point to the landmarks. Given a set $L = \{t_1, t_2, \ldots, t_n\} \subset M$, there is a continuous map

$$d_L \colon M \to \mathbb{R}^n$$

specified by

$$x \mapsto (\partial_M(x, t_1), \partial_M(x, t_2), \ldots, \partial_M(x, t_n)).$$

Pushforward along d_L again induces a distribution on $\mathbb{R}^n$ from one on M. One expects that as k increases (provided the landmarks are "generic"), the induced distributions in $\mathbb{R}^n$ will characterize the distribution on M.

In both cases, choice of the parameter k depends on some sense of the intrinsic dimension of the support of the distribution as well as the number of points available (in the case of finite samples). Unfortunately, the required k may well be quite large.

Finally, there is a substantial body of work on low-distortion embeddings of finite metric spaces into ℓ^p spaces, in particular Euclidean spaces. Recall that the distortion of a non-contractive (distance expanding) embedding of metric spaces $f \colon (X, \partial_X) \to (Y, \partial_Y)$ is given by $\sup_{x_1 \neq x_2} \frac{\partial_Y(f(x_1), f(x_2))}{\partial_X(x_1, x_2)}$. Notably, Abraham, Bartal, and Neiman show that one can construct a probabilistic embedding of an n-point finite metric space into an $O(\log n)$ dimensional space with distortion $O(\log n)$. Pushing forward distributions on Σ_m and $\mathbb{P}\Sigma_m$ provide another way of reducing statistical questions to Euclidean space.

4.4.3 Distinguishing samples from different underlying distributions

Given a set of samples $X \subset \Sigma_m$ and a partition $X = \mathcal{C}_1 \cup \mathcal{C}_2 \ldots \cup \mathcal{C}_n$ (where $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$), it is often useful to be able to determine whether or not the different $\mathcal{C}_i$ were generated from the same or different underlying distributions. For instance, $\mathcal{C}_1$ might represents samples from patients who received treatment and $\mathcal{C}_2$ is untreated patients, or the different groups $\mathcal{C}_i$ represent different observed genetic markers. Based on the discussion of the previous two subsections, we can study this problem directly in Σ_m or in via projections to $\mathbb{R}^n$.

The Fréchet mean and variance provides a summary of each collection of samples $\mathcal{C}_i$. A standard comparison between groups is then given by the distance

$$\theta_{ij} = d_{\Sigma_m}(T(\mathcal{C}_i), T(\mathcal{C}_j))$$

between the means. In order to understand the variability due to sampling, we can use bootstrap resampling (or more general k out of n resampling without replacement) to generate confidence intervals for the value of θ_i. Asymptotic consistency for the bootstrap follows from the fact that the VC dimension of the collections of balls in Σ_m and $\mathbb{P}\Sigma_m$ is bounded, via the usual criteria [79, 80, 81].

However, it is often simpler to consider tests induced by the projections into $\mathbb{R}^n$ discussed above. Here, we can compare collections $\mathcal{C}_i$ and $\mathcal{C}_j$ by using any of the many standard non-parametric comparison techniques for real distributions, for example χ^2 tests or two-sample Kolmogorov-Smirnov tests.

One pervasive problem in clinical applications is that often the number of samples is quite small, and so we are often far from the asymptotic regime for statistical tests. Standard small-sample corrections can be applied. However, for this reason a machine learning approach to analyzing the data is often more useful.

4.4.4 Clustering in Σ_m and $\mathbb{P}\Sigma_m$

Given the difficulties with statistical inference in Σ_m and $\mathbb{P}\Sigma_m$, it is useful to complement these approaches with techniques from machine learning. The most basic family of techniques we might consider is clustering, a kind of unsupervised learning. Here, given a finite set X in Σ_m or $\mathbb{P}\Sigma_m$, we search for a partitionings of the points into clusters which optimize some criterion for the "goodness" of the clustering.

Regarding Σ_m and $\mathbb{P}\Sigma_m$ simply as metric spaces, we can apply standard clustering algorithms that operate on arbitrary metric spaces. For example, we can apply standard k-means clustering, using the centroids as defined above. A related alternative is the k-medoids algorithm. Like k-means, k-medoids seeks partitions which are optimal in the sense of minimizing the sum of squared distances; the cost function for a cluster $C = \{x_i, x_j\}$ is given by $\sum_{i<j} \partial(x_i, x_j)^2$. But instead of using cluster centroids as in k-means, cluster assignments are determined by medoids, which are points $z \in C$ that minimize $\sum_i d(z, i)$. The advantage of k-medoids over k-means is that the problem of finding a centroid in Σ_m or $\mathbb{P}\Sigma_m$ is avoided.

Another natural family of clustering algorithms comes from spectral clustering techniques. Recall that spectral clustering can be applied to finite subsets of any metric space; one constructs an embedding into Euclidean space using the graph Laplacian associated to a graph encoding the local metric structure of the set of points and then performs k-means clustering. As such, spectral clustering can be applied both to Σ_m and $\mathbb{P}\Sigma_m$. However, as illustrated in Figure 4.6, there is substantial distortion under such embeddings for low dimensions.

4.4.5 Supervised Learning

Although clustering algorithms are very useful for exploratory data analysis, for clinical applications we expect that classification problems are more salient. Specifically, a temporal sequence of tumor samples will be linked with a categorical or numeric label denoting the

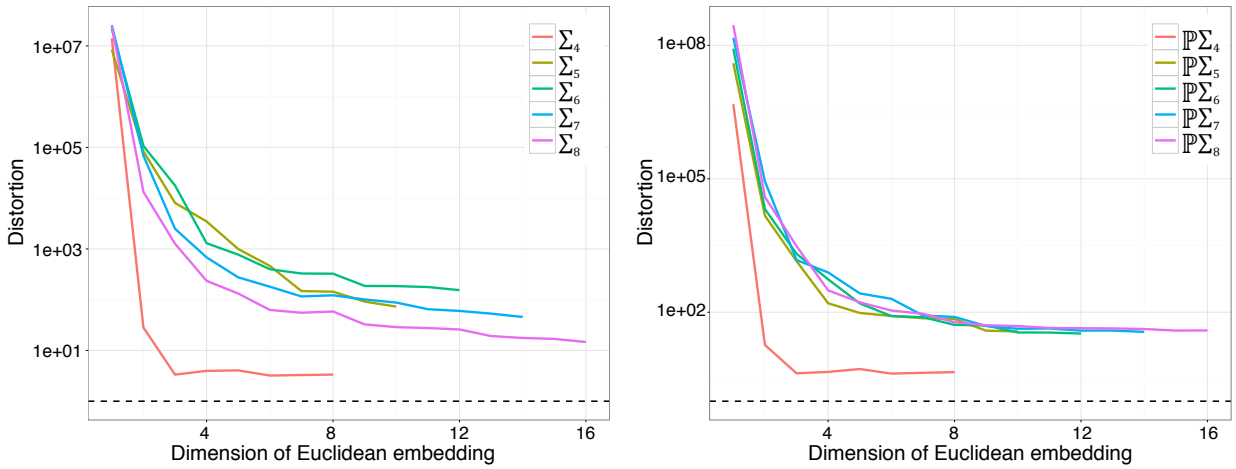


Figure 4.6: **Euclidean embedding of the affine and projective tree spaces, Σ_m and $\mathbb{P}\Sigma_m$:** A spectral embedding approach is taken to finding Euclidean approximations of our ϵ-nets. We are interested in the relation between m and the smallest d with acceptable distortion of embedding into $\mathbb{R}^d$.

clinical management of the patient. We would then like to predict patient outcomes or expect response to treatment using a discriminative supervised learning algorithm operating in Σ_m or $\mathbb{P}\Sigma_m$. Analogous to our use of k–medoids clustering for unsupervised grouping, the most basic algorithm for supervised learning is a k–nearest neighbor (k–NN) predictor. In this algorithm the predicted label for a given point is generated by taking a majority or weighted vote over the labels of the k nearest trees. The optimal value of k then specifies an order–k Voronoi tesselation of the space that provides a description of the sizes of the predictive neighborhoods surrounding each element of the data set. We use this classification algorithm to study clinical correlates of trees determined by tumor samples from glioma patients in the following chapter.

4.5 Tree Dimensionality Reduction

When analyzing a large number of genomes, phylogenetic trees are often too complex to visualize and analyze as they can contain thousands of branches. In this section, we will explain a technique for dimensionality reduction that projects a single tree in Σ_m or $\mathbb{P}\Sigma_m$

to a "forest" of trees in Σ_n, for $n < m$. The main idea is that by subsampling leaves of a large tree we can have a distribution of smaller trees that can capture properties of the more complex structure. This procedure makes it easy to visualize and analyze high-dimensional data, and avoids scalability issues with algorithms for working with the spaces of phylogenetic trees. We believe that the analysis and visualization of the resulting clouds of trees is an effective way to study high-dimensional evolutionary moduli spaces. To provide theoretical justification for this claim, we prove that this procedure is stable, in the sense that it preserves distances up to a constant factor.

4.5.1 Structured dimensionality reduction

Let $\mathcal{E}_m$ denote either Σ_m or $\mathbb{P}\Sigma_m$.

Definition 4.5.1. *For $S \subseteq \{1, \ldots, m\}$, define the tree projection function*

$$\Psi_S \colon \mathcal{E}_m \to \Sigma_{|S|}$$

by specifying $\Psi_S(T)$ to be the unique tree obtained by taking the full subgraph of t on the leaves that have labels in S and then deleting vertices of degree 2. An edge e created by vertex deletion is assigned weight $w_1 + w_2$, where the w_i are the weights of the incident edges for the deleted vertex. (It is easy to check that the order of vertex deletion does not change the resulting tree.)

A representative example of Ψ_S is shown in Figure 4.7. Using Ψ, we can describe a number of dimensionality reduction procedures. The most basic example is simply to exhaustively subsample the labels. Let $\mathcal{D}(\Sigma_m)$ denote the set of distributions on Σ_m.

Definition 4.5.2 (Tree dimensionality reduction)**.** *For $1 \leq k < m$, define the map*

$$\Psi_k \colon \mathcal{E}_m \to \mathcal{D}(\Sigma_k)$$

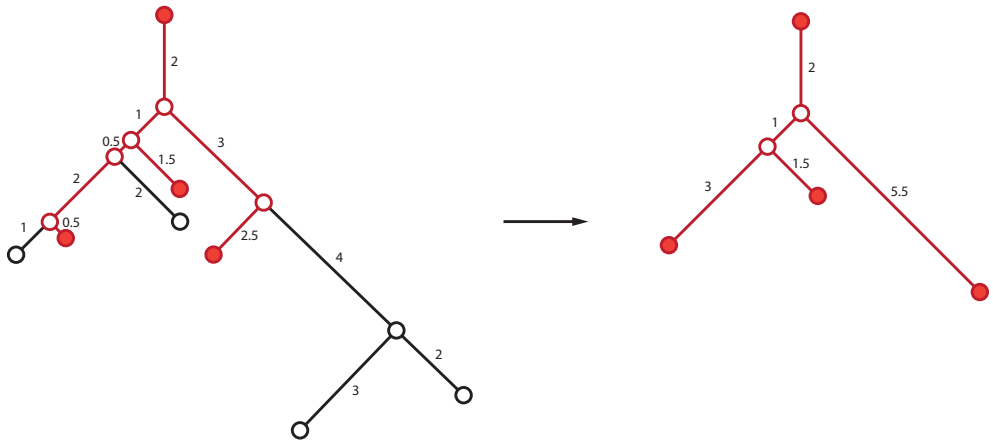


Figure 4.7: **Tree dimensionality reduction:** The leaves that are not highlighted in the starting phylogeny are pruned and their external edges removed; internal vertices of degree 2 are collapsed and edge weights on either side are summed.

as the assignment that takes $T \in \mathcal{E}_m$ *to the empirical distribution induced by* Ψ_S *as* S *varies over all subsets of* $\{1, \ldots, m\}$ *of size* k*. Define the map*

$$\Psi'_k \colon \mathcal{E}_m \to \prod_{S \subseteq \{1,\ldots,m\}, |S|=k} \Sigma_k$$

as the map that takes $T \in \mathcal{E}_m$ *to the product of* $\Psi_S(T)$ *as* S *varies over all subsets of* $\{1, \ldots, m\}$ *of size* k*.*

(In practice, we approximate Ψ_k using Monte Carlo approximations.)

Often there is additional structure in the labels that can be exploited. For instance, in many natural examples, the genomic data has a natural chronological ordering. When this holds, sliding windows over the labels induces an ordering on subtrees generated by Ψ_S. Rather than just regarding such a sequence as a distribution, the ordering makes it sensible to consider the associated trees as forming a piecewise-linear curve in Σ_k. (Note that given a set of points in Σ_k it is always reasonable to form the associated piecewise-linear curve because each pair of points is connected by a unique geodesic.) Let $\mathcal{C}_k(\Sigma_m)$ denote the set of piecewise linear curves in Σ_m; equivalently, $\mathcal{C}_k(\Sigma_m)$ can be thought as the set of ordered sequences in Σ_m of cardinality k. A schematic example of this sequential operation is given

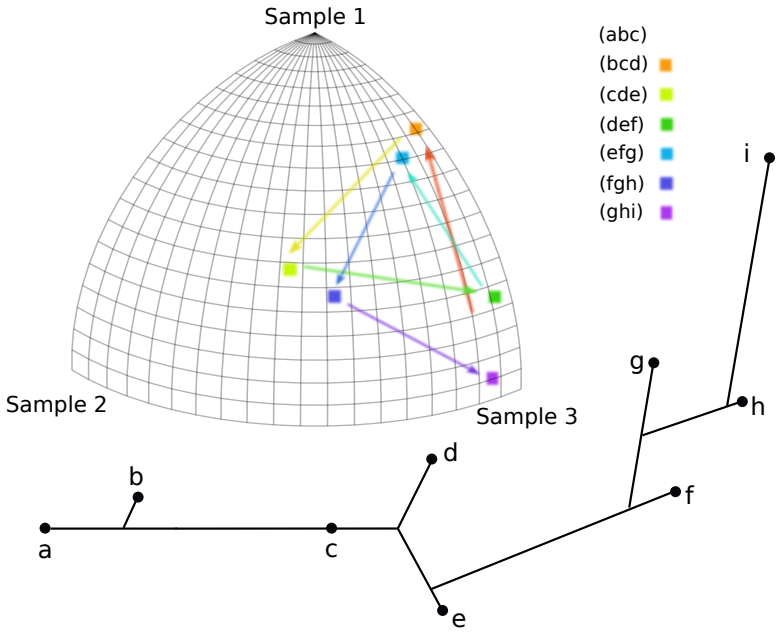


Figure 4.8: **Sequential tree decomposition on a set of ordered samples:** The subsets of trees generated from the initial phylogeny respect the ordering on the leaves, as would be the case in a temporally ordered set of samples. Each subtree can be visualized on a common set of axes, to chart motion through time. Periodicity in the sequence of branch lengths, for example, might give rise to cycles in the evolutionary moduli space.

in Figure 4.8.

Definition 4.5.3 (Sequential tree dimensionality reduction)**.** *For* $1 \leq k < m$*, define the map*

$$\Psi_C \colon \mathcal{E}_m \to \mathcal{C}_{m-k}(\Sigma_k)$$

as the assignment that takes $T \in \mathcal{E}_m$ *to the curve induced by* Ψ_S *as* S *varies over the subsets* $\{1, \ldots, k\}$*,* $\{2, \ldots, k+1\}$*, etc.*

Equivalently, we can regard this as producing a map

$$\Psi_C \colon \mathcal{E}_m \to \mathcal{D}(\Sigma_k).$$

There are many variants of Ψ_C depending on the precise strategy for windowing that is employed.

4.5.2 Tree dimensionality reduction and neighbor-joining

We have described tree dimensionality reduction in terms of the map Ψ, which is an operation on tree spaces. In practice, this technique would be applied by producing a very large phylogenetic tree from the raw data and subsequently applying Ψ. However, there is an alternative form of tree dimensionality reduction that instead subsamples the raw data to produce smaller phylogenetic trees. In this section, we discuss the relationship between these two procedures in the context of neighbor-joining.

Neighbor-joining is an algorithm for producing a tree from metric data (or more broadly, a dissimilarity measure) [63]. That is, the input is a set of points X and a metric $\partial_X \colon X \times X \to \mathbb{R}$. One of the main theorems about consistency of neighbor-joining is that when ∂_X is "close" to a tree metric ∂_T, neighbor-joining recovers T. Recall that given a tree T, the associated metric ∂_T is defined by taking the distance between leaves i and j to be

$$\partial_T(i,j) = \sum_{e \in P_{ij}} \ell(e),$$

where P_{ij} is the unique path in T from i to j and $\ell(e)$ is the length of the edge e.

The specific consistency theorem we use is due to Atteson [11]: if

$$\max_{x_i, x_j \in X} |\partial_X(x_i, x_j) - \partial_T(x_i, x_j)| \leq \frac{1}{2} \min_{e \in T} \ell(e),$$

then neighbor-joining recovers T. In this case we say that (X, ∂_X) is consistent with T.

Proposition 4.5.4. *If (X, ∂_X) is consistent with T, for any subset $S = \{x_1, x_2, \ldots x_k\} \subseteq X$, the associated submetric space is consistent with $\Psi_S(T)$.*

Proof. It is clear from the definition of $\Psi_S(T)$ that

$$\min_{e \in \Psi_S(T)} \ell(e) \geq \min_{e \in T} \ell(e).$$

On the other hand,

$$\max_{x_i, x_j \in S} |\partial_X(x_i, x_j) - \partial_T(x_i, x_j)| \leq \max_{x_i, x_j \in X} |\partial_X(x_i, x_j) - \partial_T(x_i, x_j)|$$

The result follows. □

For a metric space (X, ∂_X), let $T(X)$ denote the tree obtained from neighbor-joining applied to (X, ∂_X). As a consequence of Proposition 4.5.4, given a metric space (X, ∂_X) that is consistent with $T(X)$, the distribution $\Psi_k(T(X))$ is identical to the distribution $\{T(X_S)\}$ where S varies over all subsets of X of cardinality k and X_S denotes the metric space structure on S induced by ∂_X.

4.5.3 Stability of tree dimensionality reduction

In order to apply tree dimensionality reduction in the face of potentially noisy data, we would like to know that small random perturbation of the original sample results in a distribution of subsamples that is "close" in some sense (e.g., small shifts in the centroid in Σ_m). Conversely, if two distributions of subsamples are suitably close, we would like to be able to conclude that the sampled trees are also close.

We begin with a lemma describing the interaction of Ψ_S and the boundaries of orthants.

Lemma 4.5.5. *Fix $S \subseteq \{1, \ldots, m\}$. Let T be a point in the interior of an orthant of Σ_m, and let $\gamma\colon [0,1] \to \Sigma_m$ be the geodesic path contained in that orthant from T to T', where T' is obtained from T by collapsing an interior edge to length 0. Then $\Psi_S(T')$ is obtained from $\Psi_S(T)$ by shrinking an interior edge, and $\Psi_S(\gamma)$ is the geodesic path from $\Psi_S(T)$ to $\Psi_S(T')$.*

Proof. It suffices to show that $\Psi_S(T')$ is obtained from $\Psi_S(T)$ by shrinking an edge; given this, the assertion about γ is clear. Let $e = (v_1, v_2)$ denote the edge to be collapsed, with v_1 the vertex closer to the root and v_2 the vertex closer to the leaves. There are three possibilities. If the edge e is not present in $\Psi_S(T)$, then this means that none of the leaves below e are in S; as a consequence, none of the leaves below v_1 in T' are in S, and so $\Psi_S(T')$ will also not contain e and so $\Psi_S(T) = \Psi_S(T')$. If the edge e is present in $\Psi_S(T)$ and does not participate in a vertex collapse, this means that both edges emanating from v_2 are present in $\Psi_S(T)$ and therefore that collapsing e to 0 commutes with applying Ψ_S. Finally, if applying Ψ_S to T causes e to be concatenated with another edge, then there are

two cases to analyze — e could be concatenated via the deletion of v_1 or v_2. Suppose that the concatenation occurs because the other "downward" edge with endpoint v_2, which we will denote e', leads to leaves that are not in S. If we collapse e to 0 before applying Ψ_S, e' will still be deleted when we apply Ψ_S, and so the result will be the same. The case of deletion of v_1 is analogous. □

In light of Lemma 4.5.5, the projection Ψ_S preserves paths. The other thing we need to understand is the potential increase in length caused by applying Ψ_S. Specifically, we need to consider the impact of the addition of edge lengths that occurs when a degree 2 vertex is produced by the reduction process. In the simplest case, we are considering the map $\mathbb{R}^2 \to \mathbb{R}$ specified by $(x_1, x_2) \mapsto x_1 + x_2$, and in general, we are looking at $\mathbb{R}^n \mapsto \mathbb{R}$ specified by $(x_1, x_2, \ldots, x_n) \mapsto \sum_{i=1}^n x_i$. Squaring both sides, it is clear that

$$\partial_{\mathbb{R}^n}((x_i), (y_i))^2 \leq \partial_{\mathbb{R}}(\sum_{i=1}^n x_i, \sum_{i=1}^n y_i)^2.$$

On the other hand, since

$$\partial_{\mathbb{R}}(\sum_{i=1}^n x_i, \sum_{i=1}^n y_i) \leq n(\max_i |x_i - y_i|) \leq n\partial_{\mathbb{R}^n}((x_i), (y_i)),$$

the addition of edge lengths can result in an expansion bounded by the size of the sum.

Remark 4.5.6. *Another way to interpret the previous result is to observe that the addition map is an isometry for the Manhattan distance (when working in the positive orthant) but not for the Euclidean distance.*

For a rooted tree T, let depth(T) denote the length of the longest path from a leaf to the root.

Proposition 4.5.7. *Let $S \subseteq \{1, \ldots, m\}$ such that $|S| > 1$ and let $\gamma\colon [0,1] \to \mathcal{E}_m$ be a path from T to T'. Then $\gamma \circ \Psi_S\colon [0,1] \to \Sigma_{|S|}$ is a path from $\Psi_S(T)$ to $\Psi_S(T')$ and $|\gamma'| \leq \max(\mathit{depth}(T), \mathit{depth}(T'))|\gamma|$.*

Proof. First, observe that if T and T' are in the same orthant of $\mathcal{E}_m$, the result is clear. In this case, for any S, $\Psi_S(T)$ and $\Psi_S(T')$ will be in the same orthant of $\Sigma_{|S|}$. By the discussion above, the length of the projected path in that orthant is bounded by the length of the path in $\mathcal{E}_m$ scaled by the depth of the tree. This argument also shows that result holds for trees joined by the cone path; Ψ_S applied to the cone point produces the cone point.

Now suppose that T and T' are not in the same orthant and neither T nor T' is contained in a positive codimension subspace of $\mathcal{E}_m$ (i.e., they are not on the boundary of any orthant). Further, we assume that γ does not go through the origin and that γ can be expressed in terms of a sequence of contractions and expansions of a single edge. That is, we assume that γ only goes through codimension 1 faces of each orthant. It suffices to consider this case, since for a general γ that does not go through the origin but might traverse faces of codimension larger than 1, observe that for any $\epsilon > 0$, we can perturb γ to produce a path γ' with the same endpoints which satisfies the hypothesis above and has $|\gamma'| = |\gamma| + \epsilon$. Passing to limits then implies that the bound holds for such a path. Similarly, a limit argument implies the result for a path that starts or ends on a positive codimension subspace of $\mathcal{E}_m$. Moroever, given this case, more complicated paths that involve both rotations and also pass through the origin satisfy the bound by an easy induction.

Thus, fix a subset $S \in \{1, \ldots, m\}$. Lemma 4.5.5 now implies that $\Psi_S(\gamma)$ is a path from $\Psi_S(T)$ to $\Psi_S(T')$, and the discussion preceding the proposition implies that the potential expansion in length is $\max(\text{depth}(T), \text{depth}(T'))$. □

Remark 4.5.8. *In fact, the expansion factor in Proposition 4.5.7 depends on the number of edge conactenations that occur when Ψ_S is applied; in situations where an estimate of this is available, tighter bounds can be used.*

Using Proposition 4.5.7, it is straighforward to deduce the next two theorems that provide the theoretical support for the use of tree dimensionality reduction. The following theorem is an immediate consequence of Proposition 4.5.7, choosing the path realizing the distance between T and T'.

Theorem 4.5.9. *For $T, T' \in \mathcal{E}_m$ such that $d_{\mathcal{E}_m}(T, T') \leq \epsilon$, then for any $S \subseteq \{1, \ldots, m\}$ such that $|S| > 1$,*

$$d_{\Sigma_{|S|}}(\Psi_S(T), \Psi_S(T')) \leq \max(\mathit{depth}(T), \mathit{depth}(T'))\epsilon.$$

Moreover, this bound is tight.

Let A and B be subsets of $\mathcal{E}_m$ such that each item of A and B has a label in $L \subset \mathcal{P}(\{1, \ldots, m\})$ (where $\mathcal{P}(-)$ denotes the power set of $\{1, \ldots, m\}$). Then we can define a matching distance as

$$d_{M,L}(A, B) = \max_{S \in L} d_{\mathcal{E}_m}(A(S), B(S)).$$

Without assuming such a labelling, we define the matching distance between A and B to be

$$d_M(A, B) = \min_{\phi} \max_{a \in A} d_{\mathcal{E}_m}(a, \phi(a)),$$

where ϕ varies over all bijections $A \to B$.

The following is now also immediate from Proposition 4.5.7.

Lemma 4.5.10. *For $T, T' \in \mathcal{E}_m$ and $L \subset \mathcal{P}(\{1, \ldots, m\})$, ,*

$$d_{\mathcal{E}_m}(T, T') \geq \left(\frac{1}{\max(\mathit{depth}(T), \mathit{depth}(T'))} \right) d_{M,L}(\Psi_{S \in L}(T), \Psi_{S \in L}(T')).$$

We conclude the discussion by describing some computational results on simulated data that illustrates the divergence between d_M and $d_{\mathcal{E}_m}(T, T')$; see Figure 4.9. To demonstrate the stability of the tree projection operation, and to empirically test its approach of the distance bound, we constructed a panel of pairs of m-dimensional trees, $m > 10$. The distances between the pairs of trees were computed and compared against the distributions of distances induced by the k-dimensional projection operator, Ψ_k with $k \in 3, 4, 5, 6$. The distances between elements of the projections can exceed the original inter-phylogeny distance (ϵ), but rarely approach the upper bound. To further characterize the behavior of the distribution of subtree distances as a function of k, we compared the distributions of projected distances

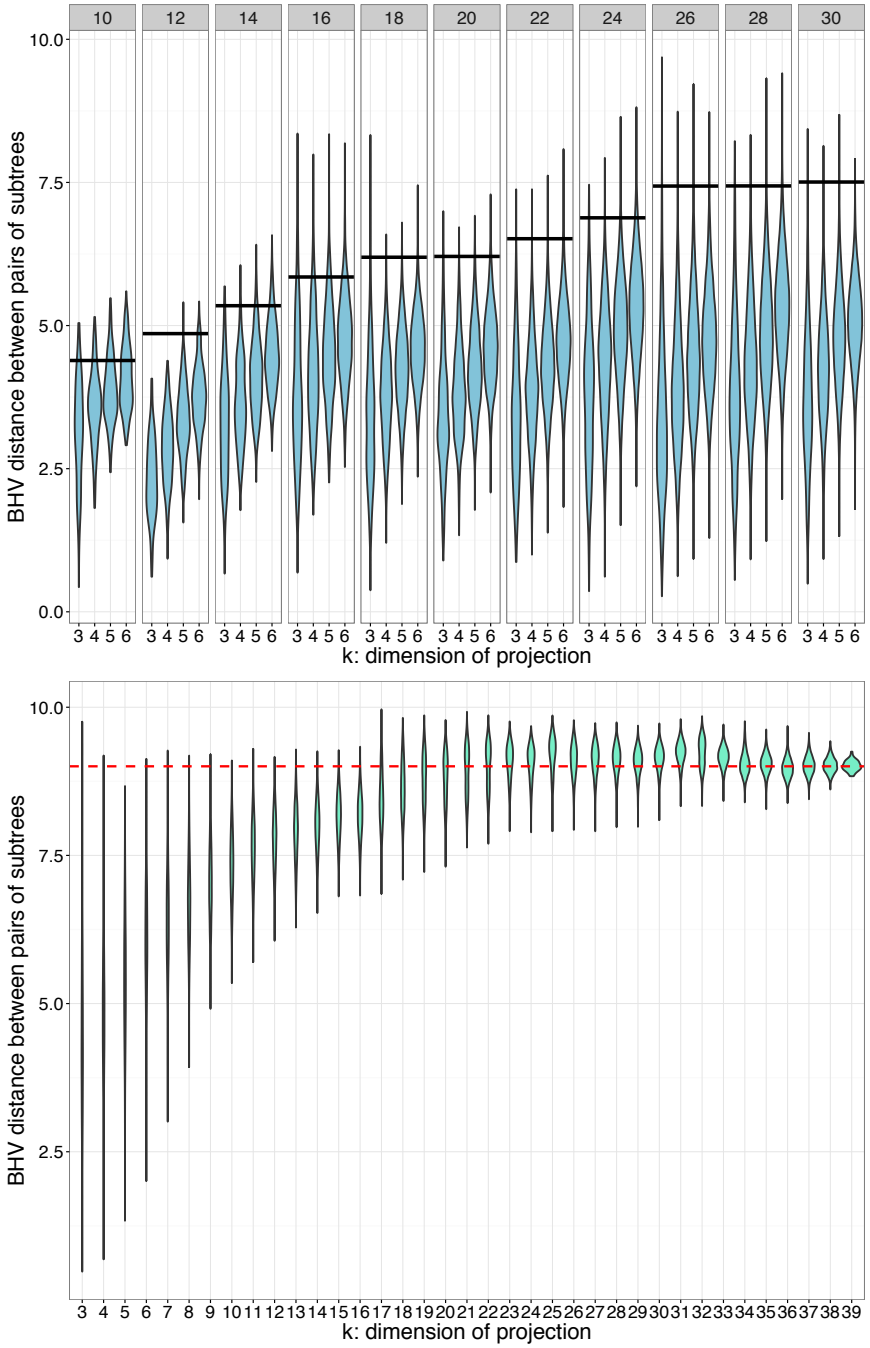


Figure 4.9: **Distributions of subsample distances under the tree projection operation:** (Top) Pairs of m-dimensional phylogenies with known distance (horizontal black lines) are projected into distributions of low dimensional trees. The distances between elements of the projections can exceed the original inter-phylogeny distance (ϵ), but rarely approach the upper bound. (Bottom) As the dimension of the projection operator approaches that of the initial phylogenies, there is a decrease in the variance of the distribution of subtree distances and its median approaches the 40-dimensional $d_{BHV}(T, T')$.

ranging from $k = 3$ to $k = m - 1$ for a fixed pair of m-dimensional phylogenies. We see that the median approaches the true m-dimensional $d_{BHV}(T, T')$ with larger values of k, while the variance appears to decrease monotonically.

4.5.4 Using tree dimensionality reduction for inference and machine learning

Broadly speaking, the various tree dimensionality reduction operators transform questions about comparison of trees or analysis of finite sets of trees to questions about comparisons and analysis of sets of clouds of trees. This has several advantages. First, when working with $\mathbb{P}\Sigma_m$, the resulting clouds live in Σ_k, and as discussed above analysis in the non-projectivized space can be simpler. Second, when projecting to Σ_3 or Σ_4, both visualization and analysis is easier (particularly in Σ_3, since that has a Euclidean metric). For example, in order to perform supervised classification on a labelled set of trees X, we can simultaneously solve classification problems in $\Psi'_k(X)$ and use majority voting in order to assign labels to new trees. Another possibility is to perform hierarchical clustering on the clouds in $\Psi'_k(X)$, resulting in another tree, and use the distance in tree space as a test statistic to discriminate between clouds. In the next chapter we describe an application that involves producing a predictor for influenza vaccine effectiveness using the variance of the distribution produced by Ψ.

Remark 4.5.11. *Another interesting direction of research is to consider the use of topological data analysis summaries (i.e., hierarchical clustering dendrograms or barcodes) for the clouds of projected points. The stability results above easily imply stability results for the associated barcodes of the projections.*

4.6 Conclusions

Phylogenetic trees are a natural and easily computed representation of evolutionary relationships in a clonal process. Billera, Holmes, and Vogtmann (BHV) defined a geometric space of the internal branches of rooted phylogenetic trees. We extended their construction here by considering unrooted trees whose total branch length is rescaled to 1. The set of all such trees forms our projective tree space, $\mathbb{P}\Sigma_m$, which we understand how to visualize and which we prove is CAT(0). We reviewed approaches for calculating statistics in tree space or performing distance-based classification such as k-NN, which will be applicable in the next chapter. In anticipation of analyzing large phylogenetic trees (or densely sampled clonal processes), we ended this chapter with a discussion of tree dimensionality reduction, which maps a large tree to a distribution of subtrees.

Chapter 5

Clonal Evolution in Tumor and RNA Virus Genomes

This chapter leverages the infrastructure developed in the previous chapter to examine instances of clonal evolution under strong selection in the realms of oncology and infectious disease. The specific settings of clonal evolution we consider are (i) serially biopsied cancers along the clinical progression of disease, (ii) xenografted tumors sampled across animal passages, (iii) H3N2 influenza hemagglutinin segments sampled seasonally over a period of over two decades. The phylogenetic tree inference implicit in our preparation of this chapter's data sets is always performed via neighbor-joining with a Hamming metric, leveraging Felsenstein's PHYLIP package [62].

Material presented in this chapter is published, wholly or in part, in: [233, 234] in collaboration with H. Khiabanian, A.J. Blumberg, R. Rabadan; and will be published in the manuscript "Phylogenetic dimensionality reduction" (in preparation) in collaboration with H. Khiabanian, A.J. Blumberg, R. Rabadan

5.1 Observing Cancer Evolution in Projective Tree Space

To summarize one of the main points of the previous chapter, we aim to understand the diverse patterns of clonal evolution in human cancer by first placing the data in the appropriate dimensional projective tree space, $\mathbb{P}\Sigma_m$. Over the clinical course of a cancer's lifetime there are alternating epochs of rapid expansion and profound selection bottlencks. Tumors are generally assumed to have a monoclonal lineage, although sufficient time and a dynamic fitness landscape can generate a large amount of genetic diversity within the population of malignant cells. We illustrate a coarse-grained picture of a patient's treatment course in Figure 5.1, with the understanding that the underlying evolutionary process may be sampled with arbitrary frequency.

In the context of cancer patients, triplet samples are often comprised of 1) **normal** tissue, 2) malignant tissue at **diagnosis**, and 3) malignant tissue at a later clinical time point such as local **relapse**. The moduli space of unrooted phylogenetic 3-trees, Σ_3, is a Euclidean 3-orthant whose basis vectors represent the 3 external edge lengths (l_n, l_d, l_r). We project each tree onto $\mathbb{P}\Sigma_3$, the space formed by the intersection $\mathbb{R}^{3+} \cap S^2$, by rescaling the branch lengths. This space is visualized in Figure 5.2

The general case of three nonzero external branch lengths is called branched evolution and such phylogenetic trees will be found far from the boundary of $\mathbb{P}\Sigma_3$. We would also like to understand the possible singular cases that occur when one or more branches degenerate. If all branch lengths are zero then we have the trivial situation of no evolution among the three samples. The edges of $\mathbb{P}\Sigma_3$ represent trees in which a single branch has collapsed to zero. As $l_n \to 0$ we have the situation where the diagnosis and the relapse are completely distinct tumors whose earliest common ancestor is in fact normal tissue. We call this divergent evolution. As $l_d \to 0$ we have the situation where the diagnosis is a perfect intermediate between the normal and relapse genotypes, the well known case of linear evolution. Lastly,

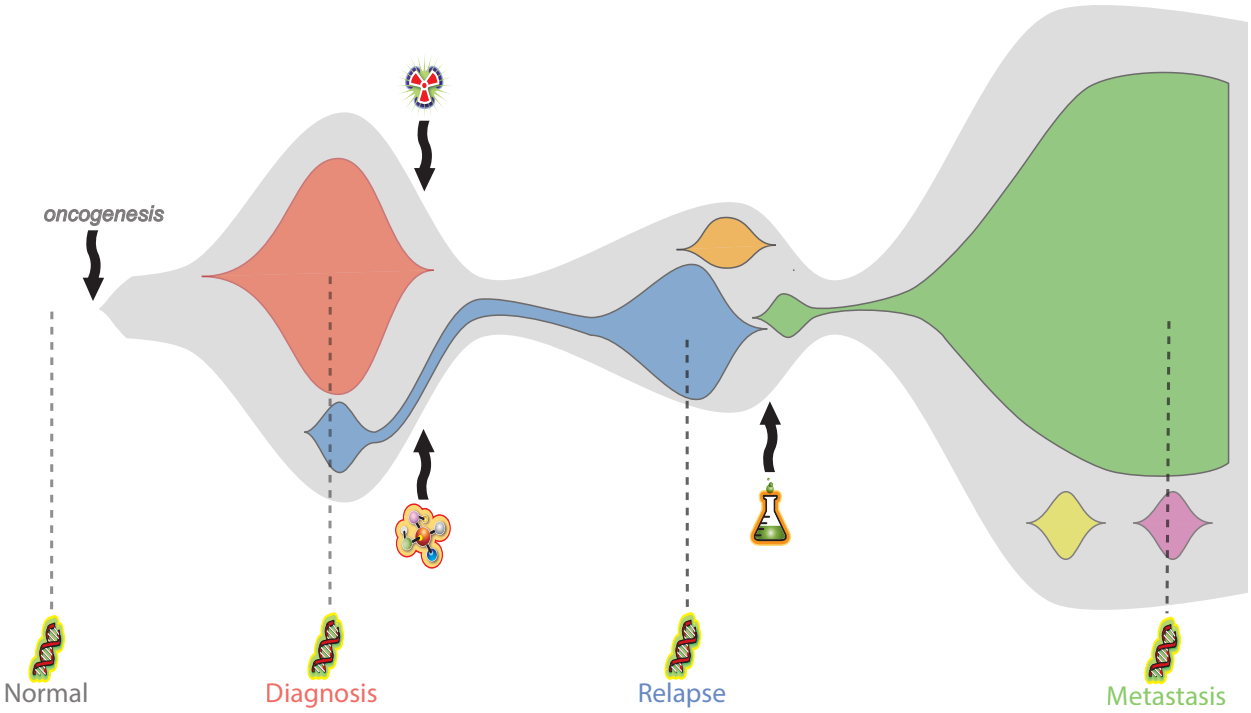


Figure 5.1: **The dynamic nature of clonal evolution in cancer:** We depict the evolution of a tumor through various clinical stages, with time running from left to right. There is an expansion of malignant cells beginning at "oncogenesis" and within the larger gray cell mass are contained different subclonal populations, represented in different colors. The overall size of the malignant cell mass is affected by therapeutic interventions, here depicted via symbols for radiation treatment, targeted molecular agents, and salvage chemotherapy. The schematic shows a progression from oncogenesis through clinically distinct phases: primary tumor, remission following initial therapy, relapse, remission following salvage therapy, and finally uncontrolled metastatic spread. Sequencing may be performed at multiple clinical time points, however the mutational spectrum primarily reflects the dominant clone at that time point.

as $l_r \to 0$ we have the situation where the relapse sample is actually the intermediate between normal and diagnosis genotypes, indicating the emergence of an ancient clone that was not dominant at the time of diagnosis. We call this revertant evolution. The vertices of $\mathbb{P}\Sigma_3$ represent trees in which two branches have collapsed to zero. Near the "shared" vertex is the case where the tumor genomics are almost identical between diagnosis and relapse samples with respect to normal tissue. From a clinical perspective, no further mutations are needed beyond the diagnosis stage for the disease to relapse, and we term this scenario frozen evolution. Near the "diagnosis" vertex is the case where the relapsed tumor is almost

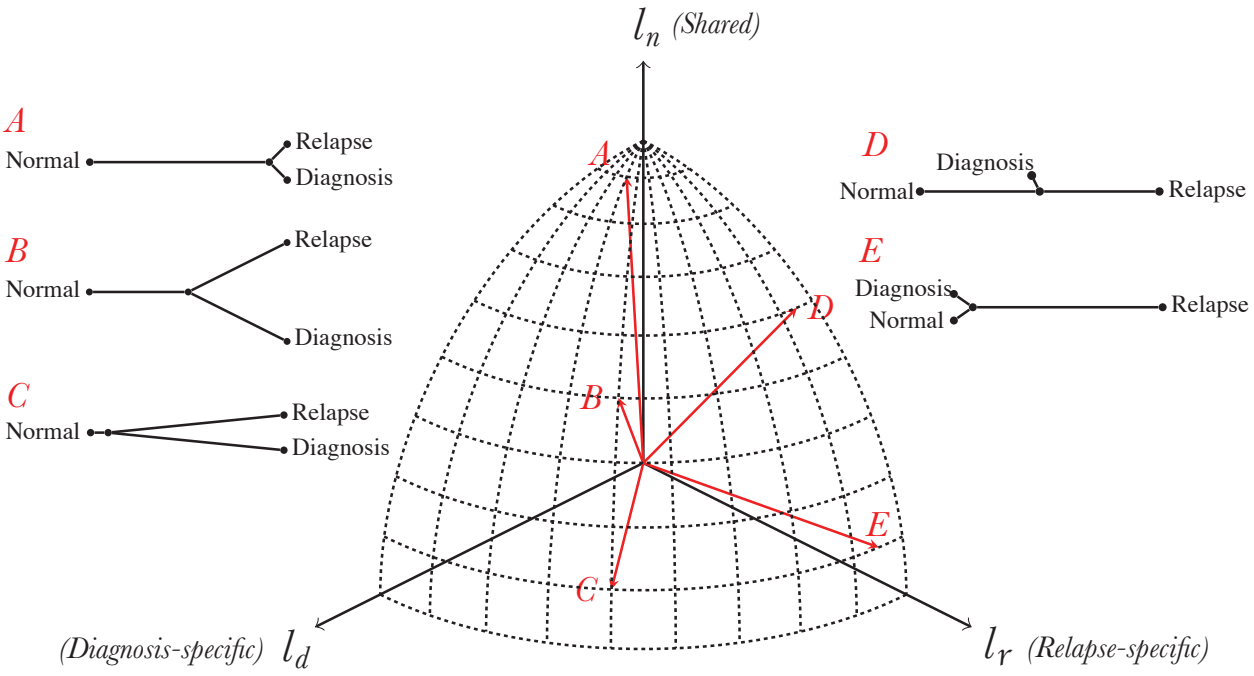


Figure 5.2: **Common evolutionary vocabulary mapped onto $\mathbb{P}\Sigma_3$:** The cancer biology literature has developed certain qualitative descriptors for the genetic relationships between serial time points in a tumor's genetics. This terminology tends to be limited to relationships between triplets of samples, and we map this qualitative vocabulary to different regions of $\mathbb{P}\Sigma_3$. A: frozen evolution, B: branched evolution, C: divergent evolution, D: linear evolution, E: hypermutation.

identical to normal, healthy tissue with respect to the lesion at diagnosis. This would be a highly unusual set of genotypes to observe since advanced cancers require some genomic deviation from normal. Near the "relapse" vertex is the case where the tumor at diagnosis is essentially the same as normal tissue compared to the number of mutations specific to the relapsed disease. Rapid accumulation of mutations can result from a shifting fitness landscape during medical therapy, and this region of the space can indicate a hypermutation phenotype. This scenario does not imply that the lesion at diagnosis has zero difference from normal tissue, but rather that the difference is dwarfed by the number of mutations accumulated in the relapsed sample (a situation often encountered in the context of mismatch-repair gene mutation).

Quadruplet samples can arise from 1) **normal** tissue, 2) malignant tissue at **diagnosis**, and 3) malignant tissue at local **relapse** and 4) malignant tissue from distant **metastasis**.

Unrooted trees constructed from quadruplet data contain a single internal edge, implying 3 possible tree topologies. We decompose the moduli space of unrooted phylogenetic 4-trees, Σ_4, into the product of spaces for its internal and external edges respectively, $BHV_3 \times \mathbb{R}^{4+}$. Upon rescaling of the branch lengths we project each tree onto $\mathbb{P}\Sigma_4$, the space formed by $\tau_3 \star \Delta_3$, which is the join of a set of three points and a tetrahedron. In Figure 5.3, we illustrate the two components of $\mathbb{P}\Sigma_4$. A quadruplet is represented by a point in the star plot (on the left) and a point in the tetrahedron (on the right). The three arms of the star plot represent the three possible tree topologies, and they meet at an origin corresponding to the degenerate case of a length zero internal branch. The vertices of the tetrahedron correspond to trees having only one nonzero external branch, the edges to trees with two nonzero external branches, and the faces to trees with three nonzero external branches.

Quintuplet samples might include 1) **normal** tissue, 2) malignant tissue at **diagnosis**, 3) malignant tissue at local **relapse**, 4) malignant tissue from distant **metastasis**, and 5) malignant tissue collected at **autopsy**. Unrooted trees constructed from quintuplet data contain two internal edges, implying 15 possible topologies. We decompose the moduli space of unrooted phylogenetic 5-trees, Σ_5, into the product of spaces for its internal and external edges respectively, $BHV_4 \times \mathbb{R}^{5+}$. Upon rescaling of the branch lengths we project each tree onto $\mathbb{P}\Sigma_5$, the space formed by $\tau_4 \star \Delta_4$. The internal space of Σ_5 can be thought of as a cone on τ_4, the Petersen graph. This object is a cubic graph with no planar embedding whose shortest circuit is 5. In Figure 5.4 we have arranged the possible tree topologies along the 15 edges of τ_4, and each vertex corresponds to an intermediate point of rotations between three adjacent topologies. We represented the space of external branches for triplet data as a 2–simplex, for quadruplet data as a 3–simplex, and naturally for quintuplet data the space will be a 4–simplex, also called a pentachoron. Explicit schemes for visualization of $\mathbb{P}\Sigma_5$ must encode the radial coordinate along the cone on τ_4 and also provide lower dimensional projections of the data residing in the pentachoron.

From a visualization standpoint, the only impediment to explicit constructions of $\mathbb{P}\Sigma_m$

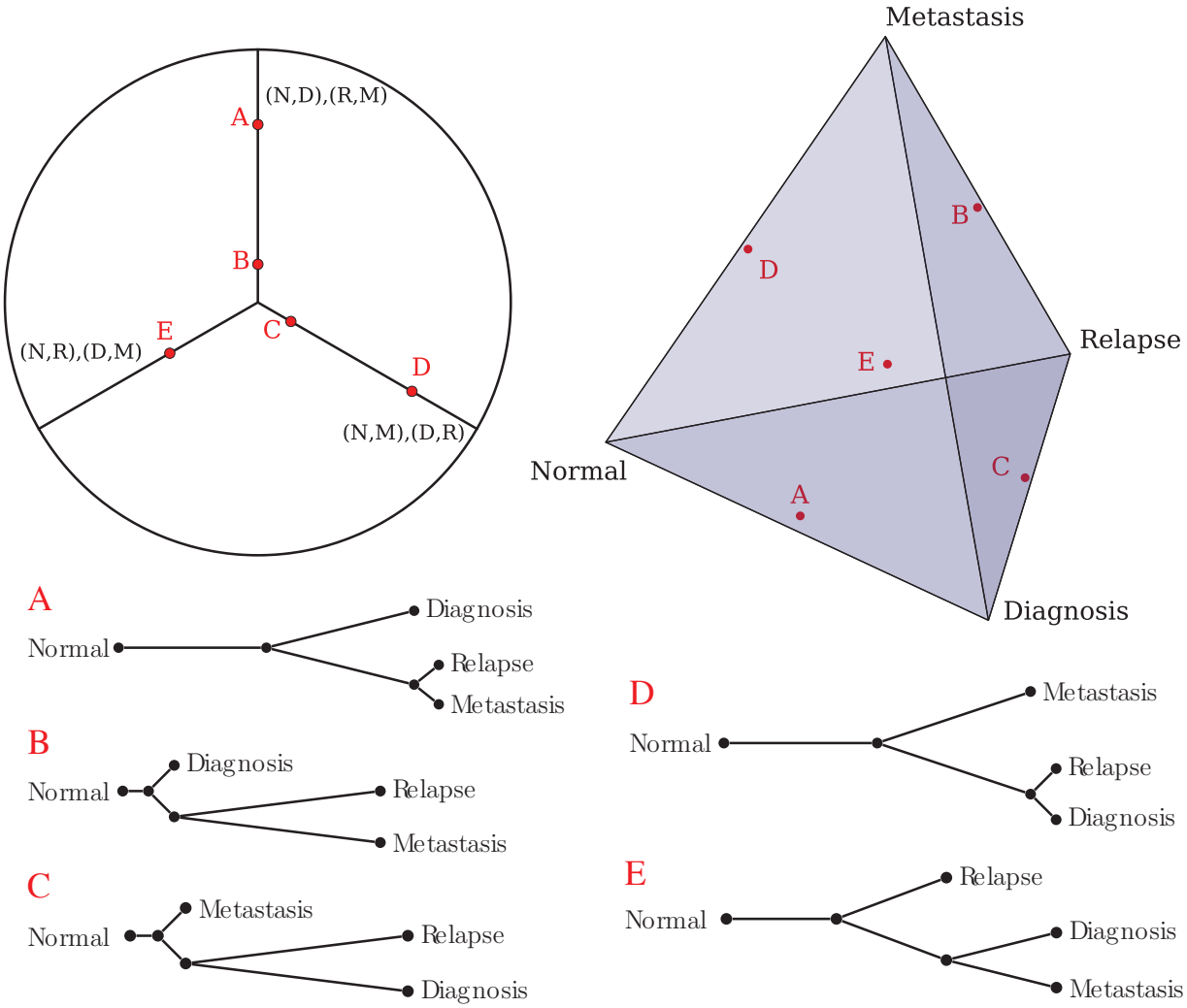


Figure 5.3: **Tree topology emerges in $\mathbb{P}\Sigma_4$:** Each tree, A—E, is represented by a pair of points. The three arms of the star plot specify the internal branch topology, while the tetrahedral plot describes the external branches.

for $m \geq 6$ is the increasing dimensionality of the space. The number of possible topologies of an unrooted phylogenetic tree on m samples grows as $(2m-5)!!$ and the overall dimension of Σ_m is $2m-3$.

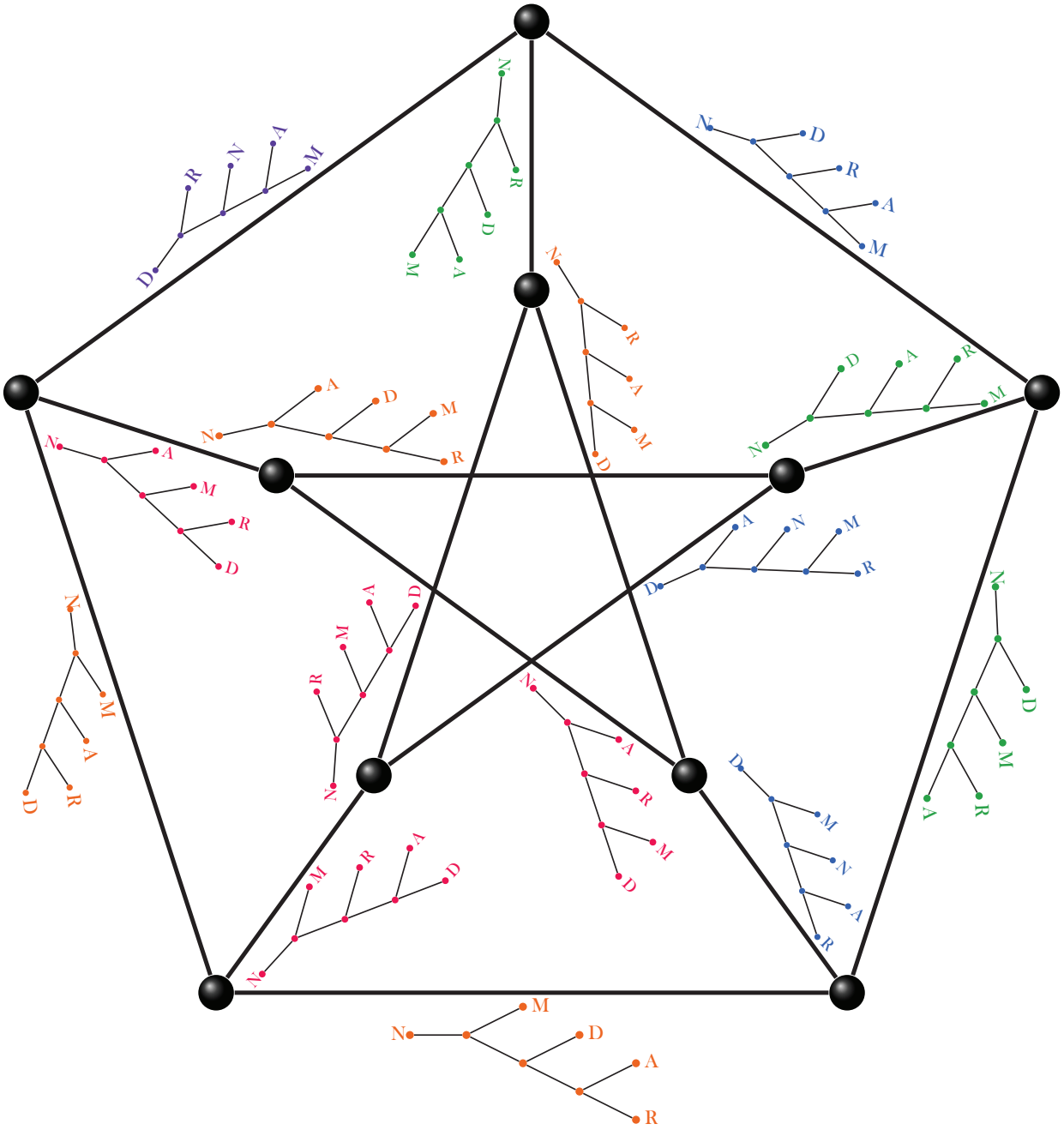


Figure 5.4: **Petersen graph for 5-sample time series:** The 15 possible tree topologies arranged on the edges of τ_4 and color coded to reflect biological plausibility. Warmer colors correspond to evolutionary relationships on the five samples that would be highly surprising, such as normal tissue (N) and the tumor at autopsy (A) being adjacent in the phylogeny.

5.2 Human Cancers Sampled Along the Clinical Progression

Progression of cancer is believed to be intimately related to the accumulation of genomic alterations in tumor cells [152]. Mutations can spur proliferation, either via activation of an oncogene or inactivation of a tumor suppressor. The spatial and temporal heterogeneity of tumors can be addressed by reconstructing the evolutionary history of tumors from different samples. For each location and time point, one can define (partially or totally) the genotype of the dominant clone. As the evolution proceeds in a clonal fashion, the relationships between dominant clones can be structured as a phylogenetic tree. Questions about the nature of the evolutionary process, mechanisms of resistance, stratification of patient tumor histories, or prognosis can be formulated as a comparison between sets of trees.

For example, a first step towards personalizing cancer therapy is to monitor the mutational status of patients along the therapeutic course. Genomic snapshots before and after administration of cytotoxic therapy can reveal the extent of population remodeling. A further goal is to establish *in vivo* mouse models of every patient's tumor, as a means of rapidly exploring drug susceptibility and resistance. Such models can be created by direct implantation of human tumor tissue into immunodeficient mice and are termed patient-derived xenografts (PDX).

5.2.1 Chronic lymphocytic leukemia

Chronic lymphocytic leukemia (CLL) is the most common leukemia in adults, primarily affecting the elderly population (median age at diagnosis is 70) [194]. CLL is a proliferative disorder of B-lymphoctyes characterized by a steady accumulation of clonal, non-functional B-cells. Treatment strategies vary greatly given the heterogeneity in disease course, ranging from watchful waiting, to localized radiation, to systemic chemotherapy. The fact that CLL is a relatively indolent malignancy makes it an excellent model for studying clonal evolution

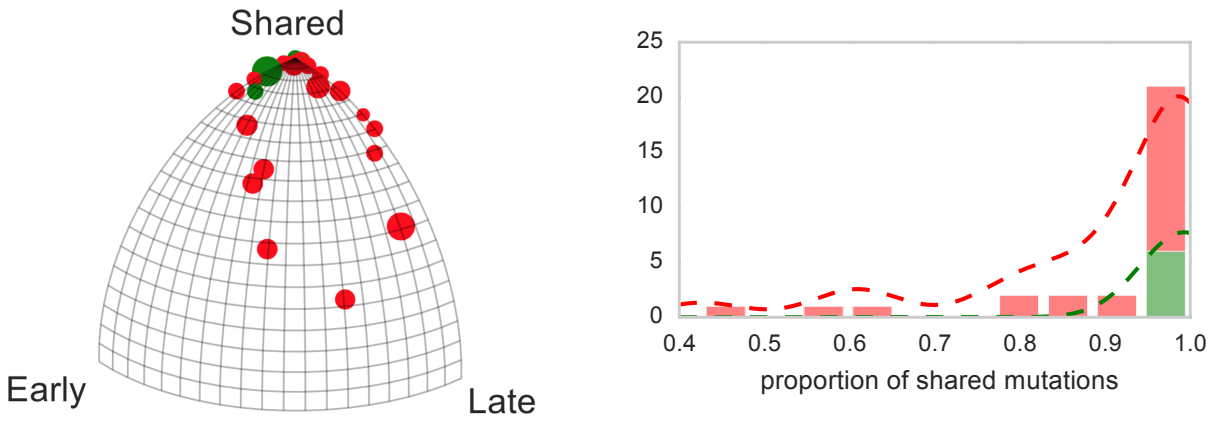


Figure 5.5: **Different evolutionary patterns observed in CLL:** (Left) In patients that did not receive chemotherapy the tumor exome did not change. Tumors from these patients, represented in green, shared most mutations along different time points. However, under therapy (shown in red) mutations before therapy were not present after treatment, and new mutations were acquired. (Right) The proportion of mutations shared between time points is significantly lower among patients treated with chemotherapy ($p = 0.049$, log-rank test).

under different therapeutic strategies [214].

A recent genomic study [120] performed whole exome sequencing on 160 CLL cases covering the spectrum of clinical courses the disease can take. This data established a space of recurrent alterations, which was then used to genotype 18 patients for whom two time points were available. Of these 18 patients, 10 of 12 treated with chemotherapy underwent clonal evolution compared to only 1 of 6 receiving no treatment according to the authors of the study. We combine the 18 patients of [120] with those of a similar study [187] wherein 3 CLL patients received chemotherapy and were sequenced at multiple time points. The multiple time points for the 3 patients studied in [187] are decomposed into all combinatorial triplets. Therefore, there are a total of 18 phylogenetic trees inferred from [120] and 12 phylogenetic trees inferred from [187]. In Figure 5.5, we map this data to $\mathbb{P}\Sigma_3$ and color based on treatment status.

From a clinical standpoint, a central question is whether treatment promotes evolution of the cancer and whether there is strong evidence for avoiding cytotoxic therapy in patient management. Quite clearly the distribution of 6 untreated patients resembles a pattern (in green) forms a tight cluster, indicating that tumors from patients who did not received ther-

apy are stable genetically, sharing most of the mutations. However, in red are represented the histories of tumors of 15 patients under therapy, presenting some mutations at different times that are not shared across different samples. The ratio between the number of mutations that are exclusive in the early branch versus the ones that are share with other phases is represented in right hand side of Figure 5.5. A number close to zero indicates a genetically stable tumor.

To assess how different are the clonal histories of tumors from untreated vs treated patients, we studied the distance between the centroids of the two populations regarded as points in $\mathbb{P}\Sigma_3$. The 95% CI for the distance between the centroids of the treated / untreated groups is (0.15, 0.36), under 1000-fold bootstrap resampling. The analogous intervals for untreated / untreated and treated / treated are (0.01, 0.14) and (0.02, 0.16) respectively. This analysis shows that the centroids of these clinically distinct sets of patients are well-resolved, supporting the idea that untreated tumors are more stable than treated ones, where district mutations can appear along the evolution of the tumor.

5.2.2 Relapsed Glioma

As another application, we examined low grade gliomas (LGG), a set of tumors of the central nervous system most often involving astrocytes or oligodendrocytes. They are distinguised from high grade gliomas (III, IV), such as glioblastoma multiforme, by the absence of anaplasia and have a more favorable prognosis. Surgery alone is not considered curative for LGG and patients are typically treated with adjuvant radiation therapy, chemotherapy, or both. If a patient relapses, the tumor may be observed to have a higher grade at that time. Johnson *et al.* studied a cohort of 23 LGG patients who relapsed, many of whom were treated with the chemotherapeutic agent temozolomide (TMZ) [102]. Whole exome sequencing was performed on tumor tissue at diagnosis and at relapse in an effort to characterize the evolution of recurrent glioma.

TMZ is an alkylating agent that directly damages the cellular genome, and accordingly

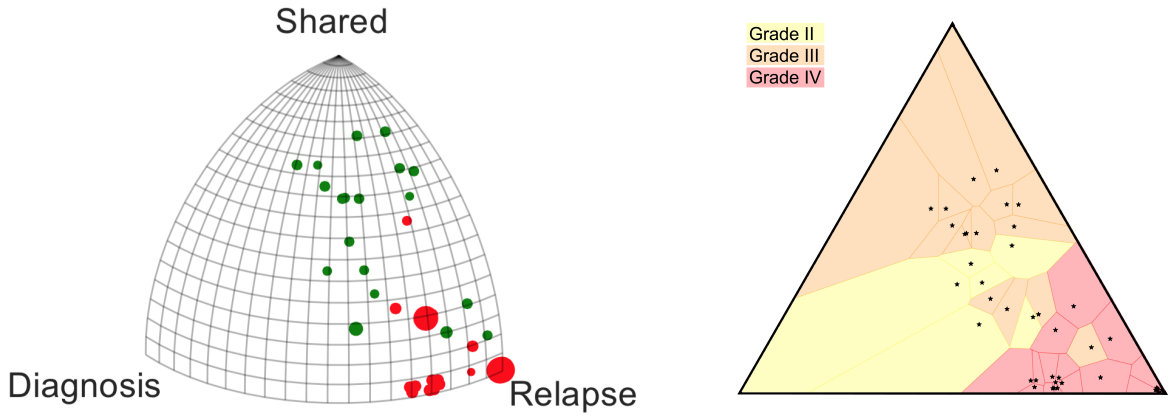


Figure 5.6: **Effects of temozolomide treatment in relapsed glioma:** Exome sequencing was performed both at diagnosis and at relapse in 23 glioma patients, allowing for 46 phylogenetic trees to be inferred (spatial replicates in certain cases). (Left) Patients treated with TMZ are colored red, and the size of a point denotes the total number of mutations observed. There is a clear tendency of TMZ-treated patients to localize in a particular corner of the space and to exhibit more mutations associated to the therapy. (Right) Less obvious is the association between the shape of patient's tree and the histologic grade of the tumor at relapse, displayed as a Voronoi tessellation.

we see that the subset of patients that were treated with TMZ show a greater acquisition of relapse-specific mutations. After projecting the trees to $\mathbb{P}\Sigma_3$, we find the 95% confidence interval for distance between centroids of the treated / untreated groups is (0.31, 0.48), under 1000-fold bootstrap resampling. The analogous intervals for untreated / untreated and treated / treated are (0.02, 0.13) and (0.02, 0.16) respectively. TMZ treatment status defines two statistically well-resolved sub-populations of patients with respect to the shape of their evolutionary behavior, an observation recently reinforced by the larger study of [215]. Also of interest is the apparent correlation between the geometry of the phylogenetic tree and the histologic grade at relapse. A 1-nearest neighbor classifier of this trinary observable (grade II, III, or IV at relapse) yields 85% accuracy under two-fold cross-validation, and the accuracy does not improve with larger values of k. The tesselation of the space associated to a 1-nearest neighbor classifier is known as a Voronoi diagram, and we have colored the cells in accordance with the grade at relapse 5.6.

This example shows that in the simple case of trees with three leaves, evolutionary tumor

histories are visibly different under different therapeutic regimes. Moreover, we see that evolutionary trajectory can be associated with prognosis, as measured by grade at relapse.

5.2.3 Metastatic pancreatic cancer

Cancer of the exocrine pancreas accounts for roughly 85% of all pancreatic malignancies and is the 4th leading cause of cancer-related deaths in the United States. In a recent study, [28] 13 cases of widely metastatic pancreatic ductal adenocarcinoma (PDAC) were studied at autopsy using a genome-wide detection method of structural rearrangements. The anatomic sites represented among the metastases include liver, lung, diaphragm, adrenal glands, peritoneum, and omentum.

We are interested in evolutionary histories involving distinct anatomical regions. To cast this data as quadruplets of successive anatomic sites of disease, we consider the hypothesis that the liver should represent the first metastatic location of PDAC. There is direct anatomical communication between the exocrine pancreas and the liver, via the common bile duct, while many of the other metastatic sites are only reachable via hematogenous spread of cancer cells. Furthermore, the liver receives a large fraction of cardiac output and might therefore be responsible for seeding the various more distant sites via the blood. For these reasons we are interested in differentiating between metastases to the liver vs. other sites. We partition the large number of samples per patient into the following disjoint subsets: normal tissue (1), primary pancreatic tumor (1), liver metastases (~ 5), non-liver metastases (~ 5).

All combinatorial 4-trees are inferred from this data and their mapping to $\mathbb{P}\Sigma_4$ is visualized in Figure 5.7. We denote the normal tissue sample by N, the primary pancreatic tumor by P, the liver metastases by LM, and the non-liver metastases by nLM. Contrary to our hypothesis that liver metastases give rise to metastases in other tissues, we find that the centroid of the data corresponds to a tree with branched ancestry between LM and nLM. Furthermore, we observe that there is no branching in the progression from normal tissue

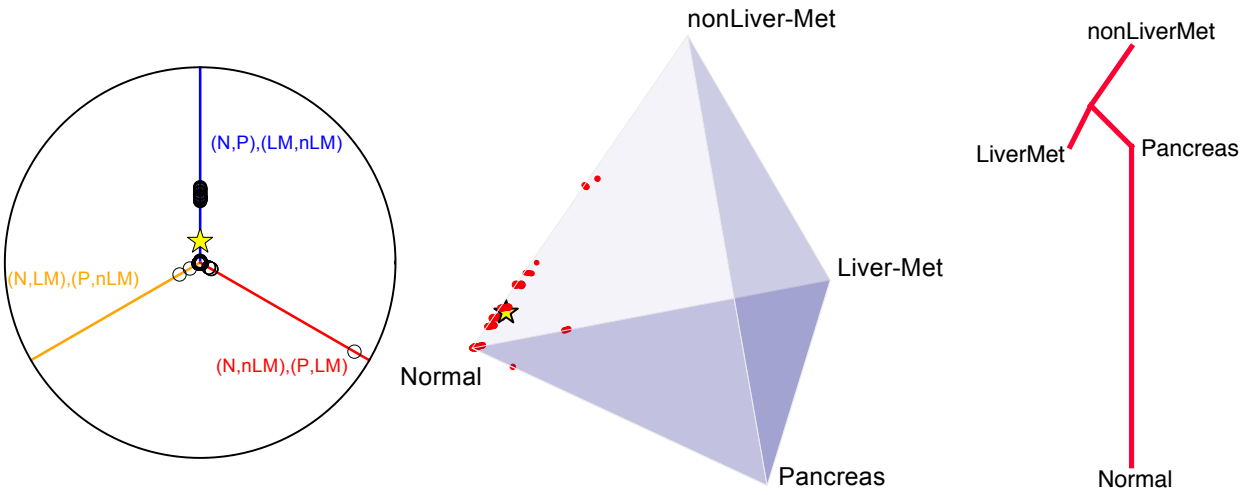


Figure 5.7: **Clonal structure of PDAC metastases:** Both linear and branching behavior observed in 10 cases of metastatic PDAC. A strong tendency toward (N,P),(LM,nLM) topology in is seen in the star plot on the left. The majority of genetic alterations are acquired at the primary tumor stage. Evolution to LM and nLM do not appear to be linearly related. The centroid of the distribution is represented as a gold star, and its associated phylogenetic tree is visualized.

to primary disease to metastatic potential. In other words, the trajectory leading to the common ancestor of LM and nLM is a linear one.

5.3 Xenografted Tumors Sampled Across Passages

The recent development of single cell transcriptomics and genomics is providing an opportunity to study the role of clonal heterogeneity in tumors [145, 56, 162] and to identify small, previously uncharacterized cell populations [85]. The single cell approach to studying complex populations brings with it new challenges associated with the large number of sampled genomes. Another rapidly maturing technology in the modeling of tumor evolution is that of patient-derived xenografts. Patient-derived xenografts (PDX) are generated by transplanting tumor tissue into immunodeficient mice, serially engrafting in new mice as each host animal expires. This provides an *in vivo* platform for drug screening as well as longitudinal monitoring of tumor adaptation and clonal dynamics.

We take advantage of recently published PDX data from breast cancer patients, where

single-nucleus deep-sequencing was performed across a lineage of host animals engrafted with a primary lesion of triple-negative breast cancer [56]. The data are comprised of normal tissue, a sample from the primary tumor, and three subsequent mouse passages. Somatic mutation calls revealed 55 informative sites of substitution in this lineage, and these variants were assigned to eight distinct cellular populations based on bulk sequencing.

Only bulk sequencing data were available from the primary tumor and matched normal tissue, while the three xenograft passages were sequenced at single-nucleus resolution. For each cellular fraction we randomly sampled a single nucleus from the first, second, and fourth PDX passages (27, 36, 27 nuclei respectively were available). This preparation of the data implies five sequential time points along the tumor's history: benign germline genome, genotype at diagnosis of primary tumor, genotype at first PDX, genotype at second PDX, and genotype at fourth PDX. The combinatorial possibilities in the latter three time points give rise to a large forest of phylogenetic trees.

We examined the difference in heterogeneity between phylogenetic trees constructed on the first four time points and those constructed on last four time points. The distributions trees from both time windows are visualized in Figure 5.8 as point clouds in $\mathbb{P}\Sigma_4$. Consistent linear evolution is seen from primary tumor through the first two xenograft passages, however we observe significant heterogeneity of tumor clones upon the fourth mouse passage. The first time window (purple) is completely contained within the topology corresponding to linear evolution, unlike the second (gold) which is centered on the origin and extends into all three possible topologies. The point cloud for the second time window displays a higher standard deviation than the first (10.49 vs. 8.69), and its centroid is essentially a star tree. Centroids and variances are computed in Σ_4, prior to rescaling of branch lengths. The high degree of genotypic heterogeneity giving rise to the second time window distribution is suggestive of a clonal replacement event between the time points of Xenograft 2 (X2) and Xenograft 4 (X4). Many of the prevalent alterations before X4 disappear during the final passage, and many new mutations rise to dominance. These results raise interesting questions about the

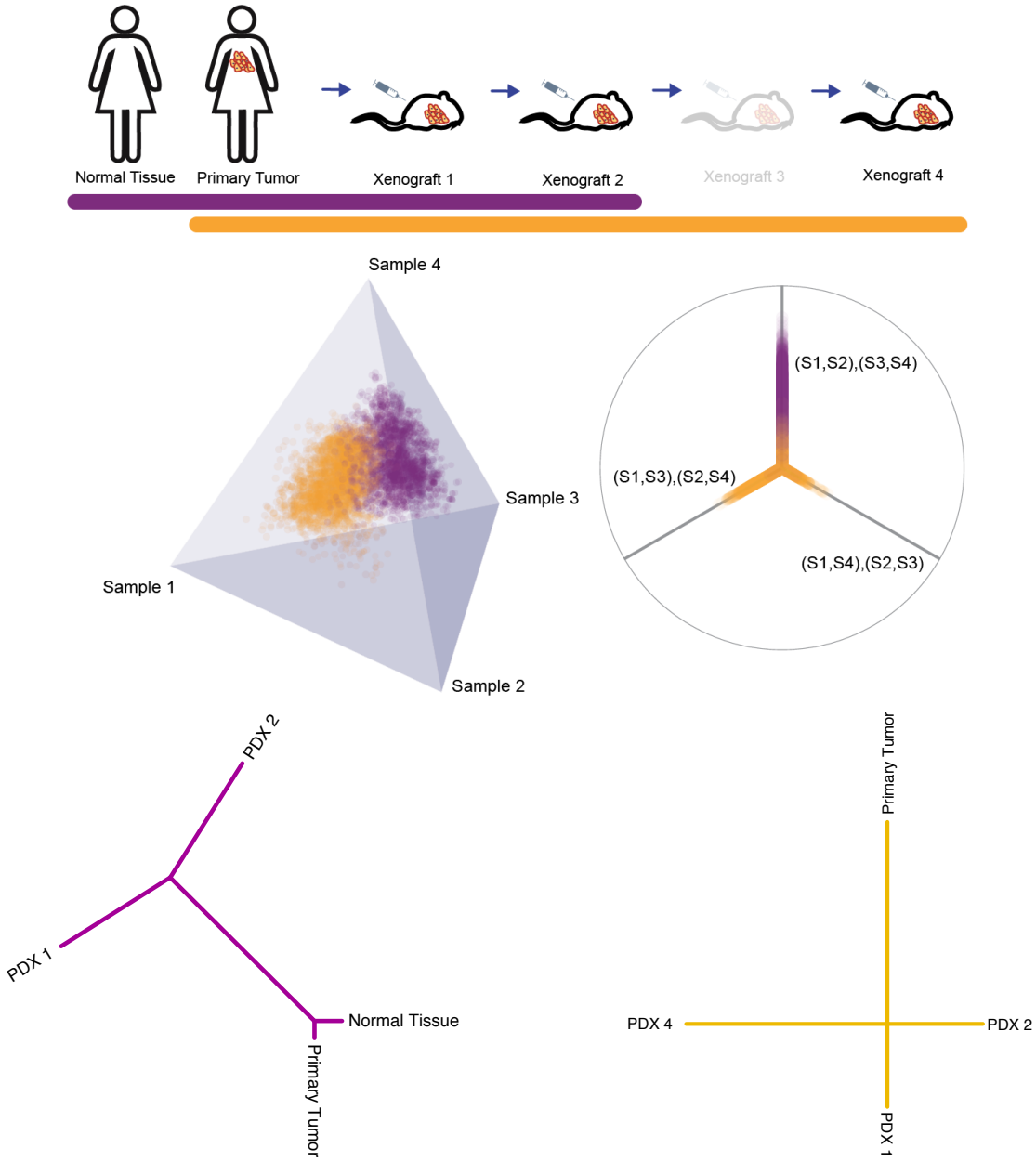


Figure 5.8: **Emerging clonal heterogeneity in patient-derived xenograft:** Single cell analysis of tumor evolution in a breast cancer derived xenograft model. Single-nucleus deep-sequence data was obtained from mouse passages 1, 2, and 4, while only bulk sequencing data was available from the primary tumor and matched normal tissue. This data is used to generate two distributions of four-leaved trees, shown in purple and gold in $\mathbb{P}\Sigma_4$. The former space displays lower standard deviation than the latter, whose centroid is a star tree.

long-term fidelity of PDX vehicles to the genetics of their ancestral primary tumors, which theoretically they serve to mimic.

5.4 H3N2 Influenza Hemagglutinin Sampled Seasonally

In this section, we describe an analysis of dynamics in the circulating hemagglutinin sequences of seasonal H3N2 influenza. Influenza A is an RNA virus that annually infects approximately 5–10% of adults and 20–30% of children, leading to more than half a million flu associated deaths [156]. Vaccination against the virus remains a major way of preventing morbidity. However, the virus genome evolves rapidly, changing the antigenic presentation of proteins that are in the envelope of the virus, mostly hemagglutinin (HA). These continuous antigenic changes, often referred as antigenic drift, can lead to failures in vaccine effectiveness. The design of the influenza vaccine is based on collected isolates of previous years, leaning heavily on the results of hemagglutinin inhibition (HI) assays to detect drift variants. The great majority of H3N2 isolates are only analyzed antigenically via HI assay, with approximately 10% of viruses undergoing genetic sequencing of the HA segment [179]. Relatively small genetic changes in the genome of the virus can cause drastic antigenic changes. Influenza vaccine failures can be associated with the emergence of new clones with novel antigenic properties that have replaced recent circulating strains.

A different phenomenon that can lead to significant antigenic changes is reassortment. The influenza virus genome consists of eight single-stranded RNA segments, two of which code the antigenic surface glycoproteins, hemagglutinin (HA) and neuraminidase (NA). When two different viruses co-infect the same host, they can generate progeny containing segments from both parental strains. Changes in the constellation of segments could introduce dramatic genetic and antigenic changes. Reassortments could occur between different viruses infecting the same host or even, more rarely, viruses that are typically found in

different hosts. Introducing viral segments from non-human reservoirs has led to major pandemics over the past century [170, 169] and was particularly associated with the emergence of the 2009 H1N1 pandemic [208, 196]. In 1968, the reassortment of then-circulating H2N2 with avian strains created H3N2 viruses, which have since been infecting human population. Reassortment of seasonal strains can also contribute to vaccine failure as low frequency hemagglutinin segments could combine with highly transmissible strains. In particular, reassortment of two H3N2 clades during the 2002-2003 season resulted in a major epidemic and higher incidents of vaccine failure in the succeeding season [70].

In this section, we analyze how the emergence of a novel subclone could be identified by unusual/unexpected tree structures and we develop a genomic predictor of vaccine effectiveness. We study the recent history of influenza A H3N2 using 1,089 sequences of hemagglutinin collected in the United States between 1993 and 2016 (Figure 5.9). Genomic data is downloaded from the GISAID EpiFlu database, and aligned with MUSCLE [54] using default parameters. Vaccine effectiveness figures were drawn from the meta analysis of [86].

First, we used tree dimensionality reduction to obtain visualizations of the flu evolutionary profile for exploratory data analysis. Using 1,089 full length sequences of hemagglutinin collected in New York state between 1993 and 2016, we relate HA sequences from one season to those from the preceding ones. We randomly select sets of HAs such that a single isolate is drawn from each of three, four, or five consecutive seasons to form a temporal window. Neighbor-joining with a Hamming metric is used to generate unrooted trees from the temporally ordered tuples of HA isolates. The case of length five windows is illustrated in Figure 5.10, where we superimpose each temporal slice onto the same moduli space. If the current viruses are most similar to viruses circulating in the immediately preceding season, one should expect an unrooted tree topology relating $(1,2),3,(4,5)$ branches. Deviations from this topology indicate unexpected genetic relationships. Figure 5.10 confirms that the vast majority of points land along the topologies most compatible with linear evolution of HA. Certain windows yield well resolved clusters of trees, while others are dispersed point

Table 5.1: **Seasonal H3N2 influenza data collected:** 23 seasons' of seasonal influenza hemagglutinin (HA) segments are collected from NY state isolates. We annotate each season with the CDC estimate for vaccine effectiveness (V.E), as well as the antigenic clusters (A.C) inferred from PREDAC [53]. Certain seasons contained more than one circulating A.C. in NY state. Near the end of each flu season a particular strain is selected for the vaccine formulation of the subsequent season, and we annotate the antigenic clusters for the strains historically chosen.

Season Start Year	Season End Year	# of Hemagglutinin Isolates	Single Circulating A.C.?	Selected Vaccine Strain A.C.	CDC Vaccine Effectiveness
1993	1994	49	Yes	SD93	38%
1994	1995	48	Yes	JH94	25%
1995	1996	26	No	WU95	45%
1996	1995	50	No	WU95	28%
1997	1995	50	No	Minor-31	-17%
1998	1995	78	No	Minor-31	34%
1999	2000	70	No	SY97, PA99	43%
2000	2001	1	Yes	SY97, PA99	81%
2001	2002	84	Yes	SY97, PA99	55%
2002	2003	16	No	SY97, PA99	33%
2003	2004	82	Yes	FU02	12%
2004	2005	86	Yes	CA04	10%
2005	2006	54	No	WS05	21%
2006	2007	30	No	WS05	52%
2007	2008	41	Yes	BR07	37%
2008	2009	68	Yes	BR07	60%
2009	2010	2	Yes	PE09	56%
2010	2011	6	Yes	PE09	60%
2011	2012	9	Yes	PE09, SW13	47%
2012	2013	52	Yes	PE09, SW13	49%
2013	2014	15	No	SW13	51%
2014	2015	81	No	SW13	23%
2015	2016	20	Yes	HK14	59%

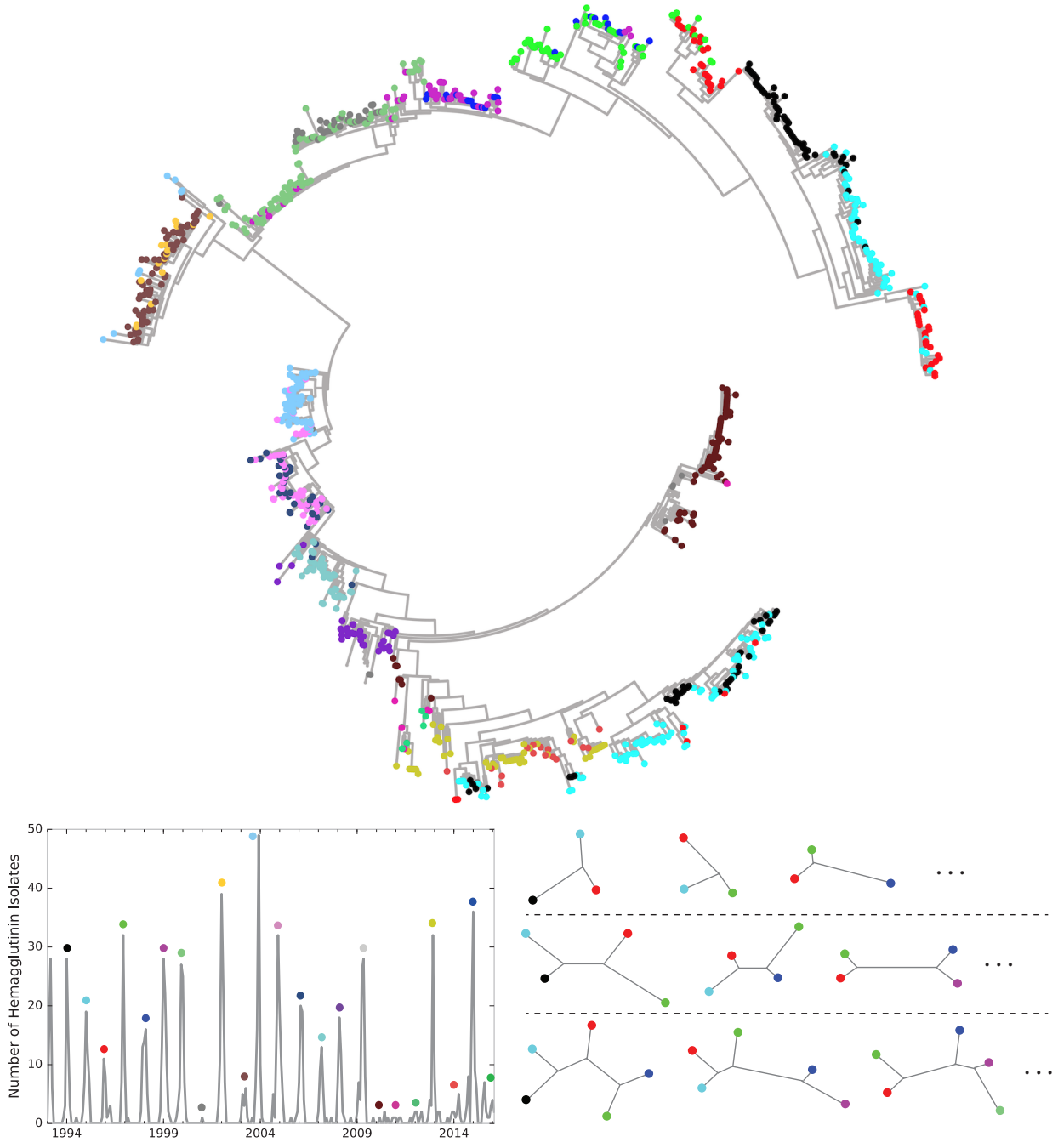


Figure 5.9: **H3N2 Hemagglutinin (HA) isolates 1993–2016:** Identifying statistical patterns in large phylogenies is often difficult. (Top) Overall phylogenetic tree inferred from 1,089 sequences, collected in New York state, spanning 24 influenza seasons. (Bottom) There is inter-season variability in the number of H3N2 isolates collected, and we generate sequences of lower dimensional trees by randomly selecting a single HA per season within a temporal window. This procedure decomposes the overall phylogeny into distributions of smaller trees.

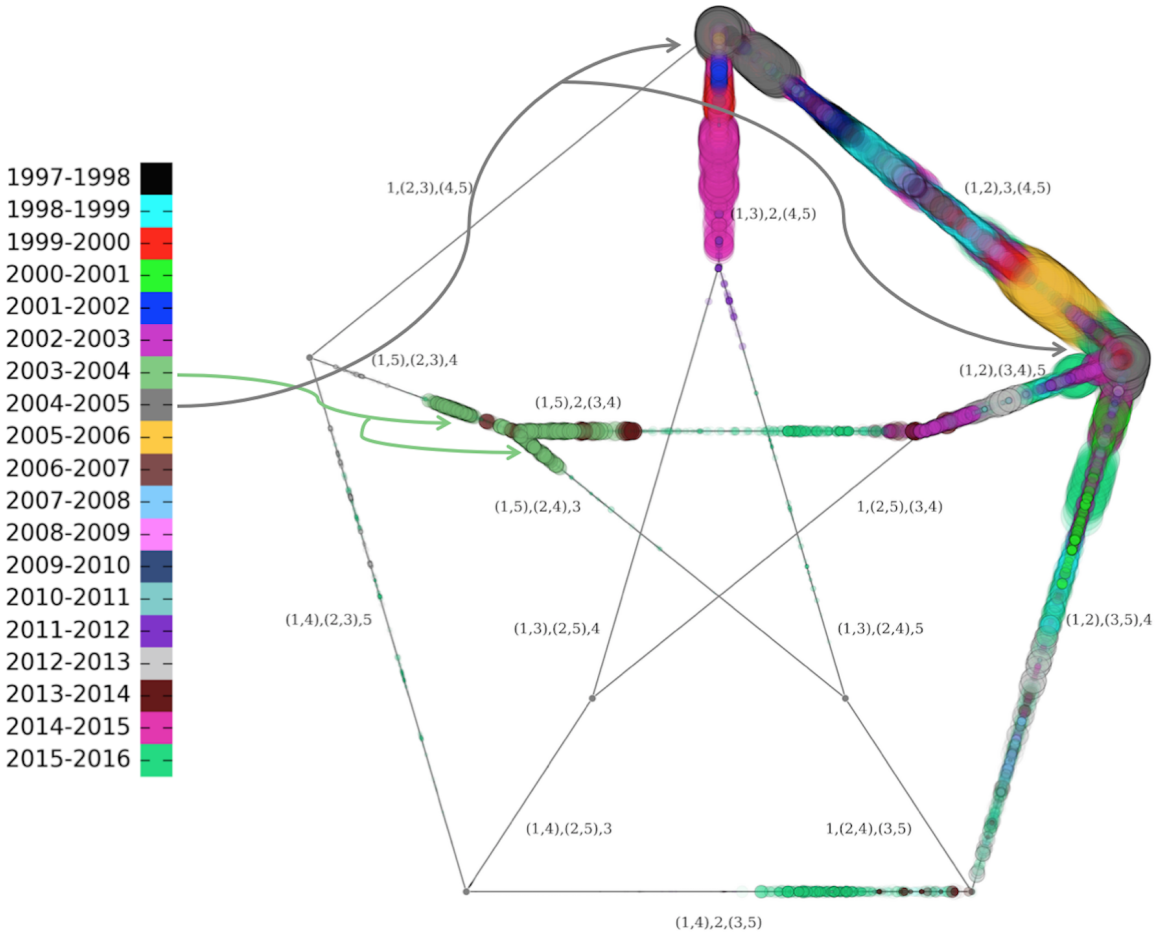


Figure 5.10: **Temporally windowed subtrees in $\mathbb{P}\Sigma_5$:** Using a common set of axes for projective tree space, we superimpose the distributions of trees derived from windows five seasons long. 1,089 full-length HA segments (H3N2) were collected in New York state from 1993 to 2016. Trees are colored by their most recent season, and point size encodes the magnitude of the cone coordinate. Two consecutive seasons of poor vaccine effectiveness are 2003-2004 and 2004-2005, highlighted with green and gray arrows respectively. The green distribution strongly pairs the 1999-2000 and 2003-2004 strains, hinting at a reemergence.

clouds. Either scenario might be indicative of elevated diversity in the HA segment or a clonal replacement event underway. The window ending in the 2003-2004 season shows a clear reemergence of strains in 2003-2004 that were genetically similar to those circulating in the 1999-2000 season [88].

Next, we used tree dimensionality reduction based on windowing to generate a predictor for vaccine effectiveness. A natural hypothesis is that elevated HA genetic diversity in circulating influenza predicts poor vaccine performance in the subsequent season. Distribution

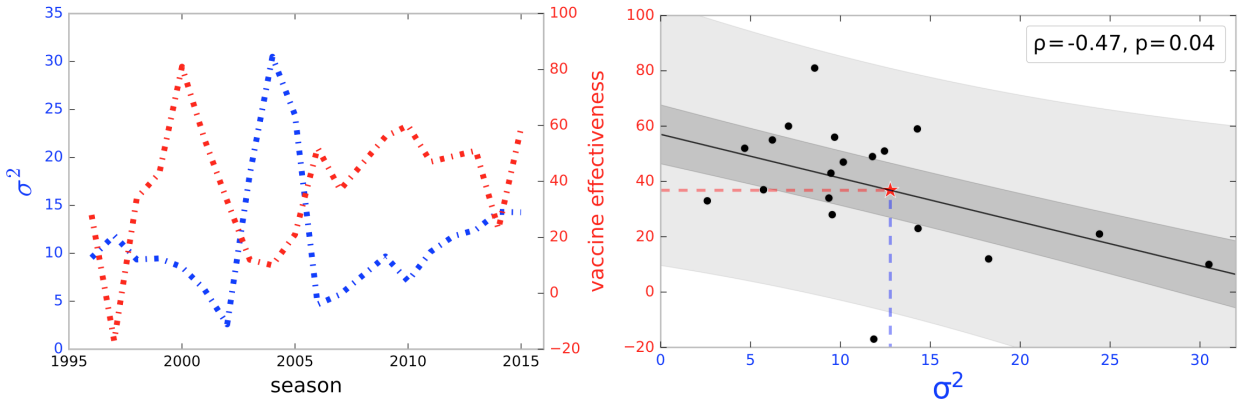


Figure 5.11: **Diversity in recent circulating HA predicts vaccine failure:** Negative correlation observed between vaccine effectiveness in season $t+1$ and the variance in trees generated from seasons $(t, t-1, t-2)$.

features that may intuitively predict future vaccine performance include the variance and the number of clusters in the point cloud. However, given our limited number of temporal windows, too rich a feature space runs the risk of overfitting, so we focused simply on the variance. In Figure 5.11 we illustrate the prediction of vaccine effectivenss using the variance of the distribution of trees generated by a lagging window of length 3. In our notation, a window labeled year y would include the flu season of $(y-1, y)$ and preceding years. The vaccine effectiveness figures represent season $(y, y+1)$. It is clear, both from the left and right panels, that lower variance in a temporal window predicts increased future vaccine effectiveness, with a Spearman correlation of -0.52 and p-value of 0.02. The lone outlier season came in 1997-1998 [86], when the vaccine effectiveness was lower than expected. In this season the dominant circulating strain was A/Sydney/5/97 while the vaccine strain was A/Wuhan/359/95. The analysis can be carried out with length-4 or length-5 windows to yield a similar result. Noteworthy is the fact that this association rests only on aligned nucleotide sequence, making no direct use of HA epitope or HI assay data. The correlation between variance of tree distributions and vaccine effectiveness allows us to estimate the influenza vaccine effectiveness for future seasons based on genomic data. In particular, for the 2016-2017 season, the variance in tree space is 12.77, corresponding to an approximate effectiveness of 36%.

We can also use our approach to retrospectively examine the annual W.H.O. decisions to either keep or change the H3N2 componenet of the Northern hemisphere vaccine, and whether the choice resulted in superior or inferior vaccine effectiveness in the following season. A coarse-grained view of the antigenic features of our HA isolates can be obtained using the work of [53], who defined a clustering of antigenic phenotype and trained a naive Bayes model, that maps HA protein sequence to these labels. We begin by labeling our phylogenetic trees using the antigenic cluster (A.C.) assignments of the classifier, and selecting those seasons in which more than one A.C. is observed. The goal is to define a mapping between features of the different A.C. distributions and the change in vaccine effectiveness of the next season relative to the present. Figure 5.12 indicates that in 9 of the 19 seasons for which we have data, only a single A.C. was observed. In this case it does not make sense to ask whether a change in H3N2 vaccine strain should have considered for the subsequent season, since we detect no antigenic diversity. The other 10 seasons are represented as vectors comprised of 4 features: distance between the centroids of the older / newer A.C. distributions, standard deviation of the older A.C. distribution, standard deviation of the newer A.C. distribution, whether the vaccine strain for the subsequent season was changed from that of the current season. We associated a binary label to each of the 10 seasons: whether the change in vaccine effectiveness was positive or negative.

Using a heavily restricted vocabulary of logical and arithmetic operators, we exhaustively search for a decision tree mapping the feature vectors to the binary label, based on a fitness function maximizing area under the receiver operating characteristic [184]. A decision rule that achieves perfect classification on this data set is depicted in Figure 5.12. The limited size of the data set means that caution is warranted when interpreting the results. Nonetheless, the results do suggest that if a vaccine strain is unchanged, a higher variance in the old A.C. distribution predicts improvement in vaccine effectiveness ($\Delta V.E. > 0$), while if a vaccine strain is changed, then the new A.C. distribution being well-resolved from the old A.C. predicts $\Delta V.E. > 0$.

Table 5.2: **Distributions of trees in multi-A.C. seasons:** 10 of the 19 possible three-season windows terminate on a time point with more than one circulating A.C. observed. 20,000 trees are sampled at random for each temporal window (from all combinations possible given the set of three seasons' HA isolates). For the purposes of statistical characterization of these distributions of trees we do not rescale the branch lengths and prefer to work in affine BHV space. We are interested in basic summary statistics of the point clouds, such as the distance between their means and their variances.

Final Season of Window	$\lvert\mu_{red} - \mu_{black}\rvert$	σ_{red}	σ_{black}	Vaccine Updated?	$\Delta_{t+1,t}V.E.$
1996-1997	1.927	11.503	11.696	No	-45%
2013-2014	3.756	14.817	13.791	Yes	-28%
2002-2003	28.825	2.403	14.730	No	-21%
1995-1996	4.212	8.774	9.298	Yes	-17%
2006-2007	11.386	4.410	5.269	No	-15%
1998-1999	2.413	9.797	9.309	No	+9%
2005-2006	3.242	4.109	4.447	Yes	+31%
2014-2015	7.027	12.969	14.182	Yes	+36%
1999-2000	3.941	9.910	8.384	Yes	+38%
1997-1998	7.254	9.052	10.300	Yes	+58%

5.5 Conclusions

We began this chapter with schematic visualizations of the projective tree space as it would pertain to serial cancer samples. The three types of genomic data from clonal processes we that analyzed were bulk sequencing of human tumors (CLL, glioma, PDAC), single cell analysis of a BRCA xenograft model, and finally seasonal sampling of H3N2 infleunza. The human tumor vignettes demonstrate the utility of analyzing clinically labeled data in $\mathbb{P}\Sigma_m$, as distances between point clouds can be easily appreciated. The patient-derived xenograft analysis highlights a possible clonal replacement event between the second and fourth animal passages, captured by an increase in variance of the point cloud and a translation away from a definite topology to being centered at the origin. This was the first data set that required the use of the tree dimensionality reduction operation. Our final data set came from 23 seasons' of H3N2 hemagglutinin (HA) collection and sequencing in NY state. We demonstrate an inverse correlation between the variance of distributions of smaller HA trees

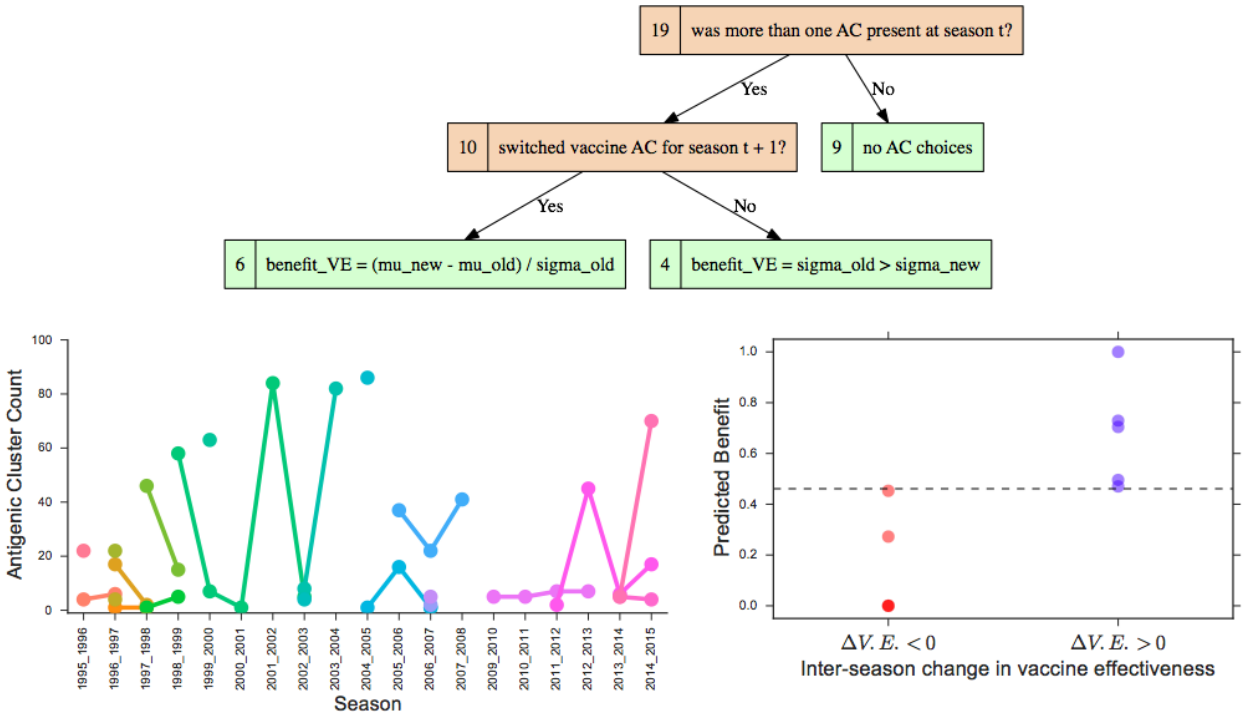


Figure 5.12: **Stratification of trees on predicted antigenic cluster:** Trees are categorically labeled using the predicted [53] antigenic cluster (A.C.) of their most recent isolate. (Top) A simple decision rule is fit [184] to the 10 seasons in which more than one A.C. is observed, explaining the change in vaccine effectivness (V.E.) in the following season. (Bottom) Colors encode different A.C. labels associated to the HA sequences collected over time. We also plot the predicted V.E. change given by the decision rule against the historical data, for the 10 relevant seasons.

and the influenza vaccine effectiveness and attempt to link our purely genomic analysis with data on each sequence's predict antigenic cluster (A.C.). A promising future application of our tree space method lies in the tracking of affinity maturation of antibodies.

Part III

Directed Clonal Evolution of Short RNA Genomes

Chapter 6

Evolving RNA Geometries For Ribosomal Frameshifting

In this chapter we continue our study of clonal processes, while moving from naturally occurring genomes to synthetic RNA genomes. Selection, reproduction, and mutation are all recapitulated *in vitro* to yield a platform for directed evolution. The fitness objective of this directed evolution requires inducing a slippage of reading frame in the translating ribosome. We thus evolve RNAs capable of very high efficiency -1 programmed ribosomal frameshift starting from a randomized RNA scaffold. Our approach brings together the combinatorial complexity of a classical *in vitro* RNA selection and the biochemical complexity of a cell lysate. High-throughput sequencing is leveraged as a fitness read-out for a population of high fitness RNAs, and a computational analysis is established for determining optimal frameshifting pseudoknot geometry and nucleotide composition. An application in the rational design of ligand-responsive riboswitches is highlighted.

Material presented in this chapter is published, wholly or in part, in: [9] in collaboration with A.V. Anzalone, A.J. Lin, R. Rabadan, V.W. Cornish

6.1 Programmed Ribosomal Frameshift: Natural and Engineered

The ribosome coordinates the biosynthesis of proteins from mRNA templates according to a standard translational program. While the ribosome typically executes translation uniformly and with high fidelity [232], in some cases the program is temporarily altered in order to change the protein output of a given gene [78, 66]. This reprogramming endows the translation apparatus with expanded synthetic capabilities, enabling the expression of proteins containing non-canonical amino acids or the regulated expression of multiple distinct protein products from a single mRNA transcript. Some forms of translation reprogramming have been directly adopted for biotechnology, including internal ribosome entry sites (IRES) [139] and co-translational cleaving 2A peptides [46]. While significant progress has been made in these areas, other modes of translation reprogramming remain largely unexplored despite their potential applications to synthetic biology.

Many reprogramming mechanisms utilize cis-acting RNA elements embedded within mRNAs. Recently, other RNA-based gene expression frameworks have emerged as powerful tools for engineering biological systems [95]. Over two decades of SELEX and related *in vitro* selection experiments [57, 210, 175] have yielded synthetic RNA molecules, termed aptamers, that bind to diverse ligands [100, 17]. These aptamers have been coupled to RNA-based expression platforms to construct ligand-controlled gene regulatory tools such as allosteric ribozymes [200, 114, 205, 115, 125]. These RNA devices have been utilized in cellular computation [222], regulation of gene expression [12], and phenotypic control [33, 76]. The apparent modularity of device construction suggests that other RNA gene expression frameworks could be exploited to engineer new classes of RNA devices with distinct regulatory opportunities.

We identify -1 programmed ribosomal frameshifting (-1 PRF) as a potentially powerful gene regulatory mechanism for RNA device engineering. Eukaryotic -1 PRF signals

contained within mRNA transcripts are composed of two principal features: (i) a heptanucleotide slippery site where the frameshift event occurs, with the general sequence X-XXY-YYZ (dashes indicate original frame; X denotes any nucleotide; Y denotes A or U; Z denotes A, C or U); and (ii) a downstream stimulatory RNA structure, typically a hairpin or pseudoknot [24]. When encountering a -1 PRF signal in an mRNA, a fraction of translating ribosomes slip back by a single nucleotide, placing the translation apparatus in the -1 reading frame. This alters the amino acid composition of the polypeptide that is synthesized downstream of the frameshift site.

-1 PRF has been well-studied in retroviruses such as HIV, where it serves to establish a precise ratio of Gag to Gag-Pol proteins [98]. Regardless of variation in mRNA transcript levels or translational activity, the stoichiometry of frameshift to non-frameshift protein products remains constant for a given -1 PRF signal. While viral -1 PRF signals have fixed frameshift activities, it may be possible to engineer frameshift signals to respond to environmental ligands [96].

6.2 Design and Read-Out of *In Vitro* Evolution

The general strategy is to contruct an iterative scheme for *in vitro* mutation, reproduction, and selection acting on a particular RNA construct. These three forces are sufficient to ensure evolution of the population. The RNA molecules of this population should be considered independent genomes, each conveniently fitting within a single Illummina read.

An mRNA library was designed such that only active -1 PRF signals form mRNA-peptide fusions and become enriched. Starting from a prokaryotic riboswitch scaffold [230], 14 of 35 nucleotides were randomized to generate 2.68×10^8 sequence variants (Figure 6.1). A library of this size is easily accommodated by the mRNA display technology, which allows for upwards of 10^{14} input sequences [127]. To enrich active -1 PRF stimulatory elements, the library was encoded downstream of the heptanucleotide slippery site U-UUA-AAC and

an in-frame UAG termination codon. mRNA display templates were translated in vitro in rabbit reticulocyte lysate and purified based solely on the presence of the peptide epitope tags. Unlike traditional mRNA display, which uses the purified mRNA-peptide fusions for subsequent selections, the functional selection for -1 PRF activity is complete at this stage of the cycle.

After three rounds of in vitro selection, assaying of selection products in a dual-fluorescent protein (dual-FP) reporter in S. cerevisiae revealed enrichment for active -1 PRF stimulatory elements with in vivo efficiencies of up to 30% (Figure 6.2). Moreover, flow cytometry of individual clones revealed that the ratio of fluorescent proteins remained constant for a given population of cells harboring the same -1 PRF signal, irrespective of total protein synthesis. As a result, two populations of yeast with frameshifting efficiencies that differ by only three- to fourfold are highly resolvable, despite overall expression levels that span several orders of magnitude.

High-throughput sequencing of the selection products is performed using the Illumina HiSeq platform (Columbia Genome Center). A computational pipeline will be described below to identify promising sequence motifs for downstream engineering applications. Briefly, sequences were grouped on the basis of compatibility with different hairpin- type (H-type) pseudoknot geometries, assessed for structure subtype enrichment, and then finally clustered into motifs based on primary sequence identity. The differential abundances of sequence variants within a motif were used to identify nucleotide preferences at variable sites and mutation intolerant positions. This analysis provides strong support for further engineering of -1 PRF motifs, and it forms the basis for rational switch engineering (Figure 6.2).

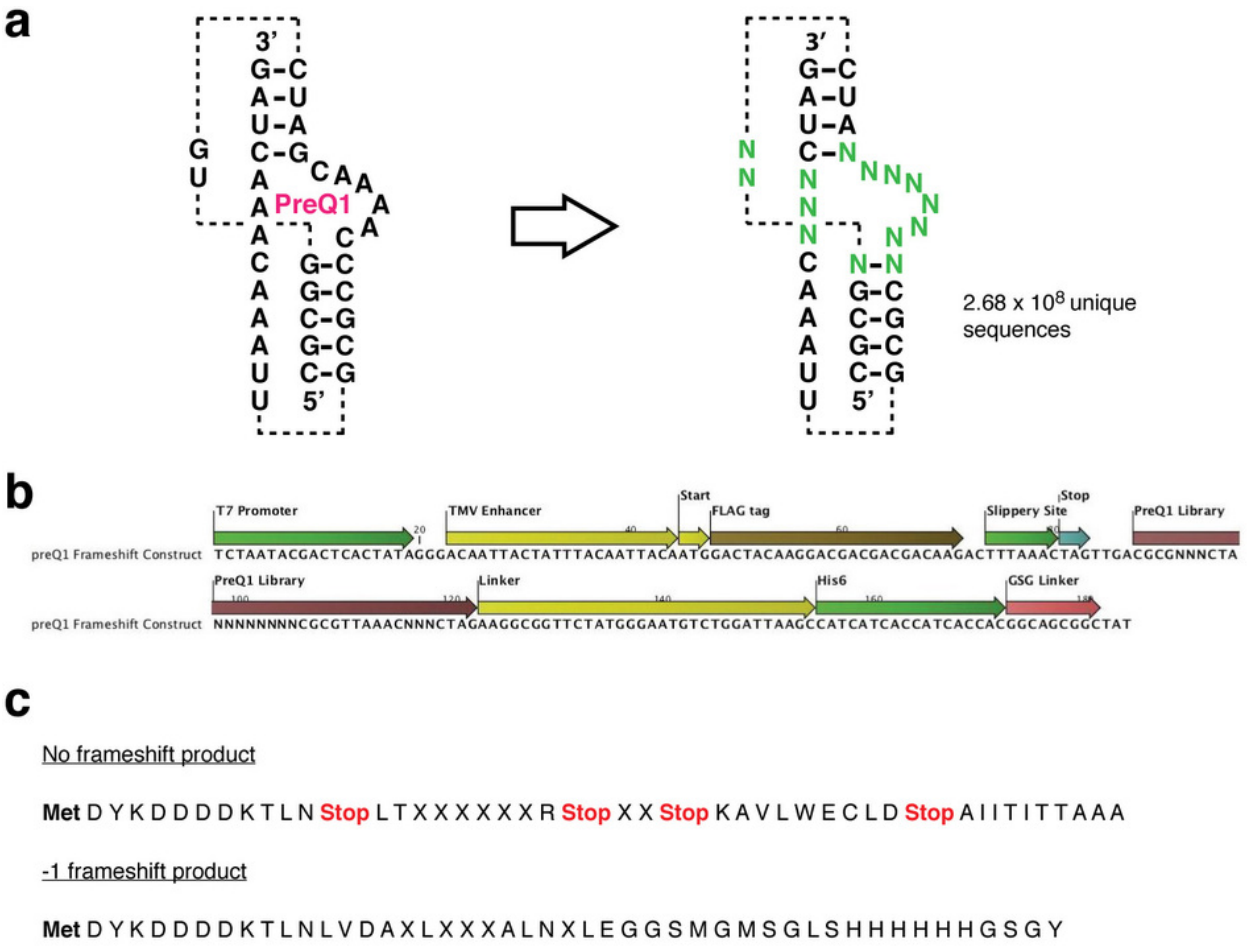


Figure 6.1: **Selection construct for PRF:** (a) The library is constructed based on a previously characterized prokaryotic PreQ1 riboswitch scaffold [230]. Nucleotides surrounding the ligand binding pocket are randomized, generating a library of $\sim 2.68 \times 10^8$ variants. (b) Annotated sequence of the selection construct in DNA form. (c) The translation product resulting from maintenance of the original reading frame (upper) or -1 frameshift at the heptanucleotide slippery site (lower).

6.3 Detecting Efficient Frameshifting Pseudoknot Compositions

A high-level description of our computational approach, to accompany panel (d) of Figure 6.2, is as follows. The scaffold and its 14 variable positions define a restricted region of sequence space ($\sim 2.68 \times 10^8$). We constructed a pseudoknot (PK) feature space based on a combination of the nucleotide constraints and our user-defined segment constraints,

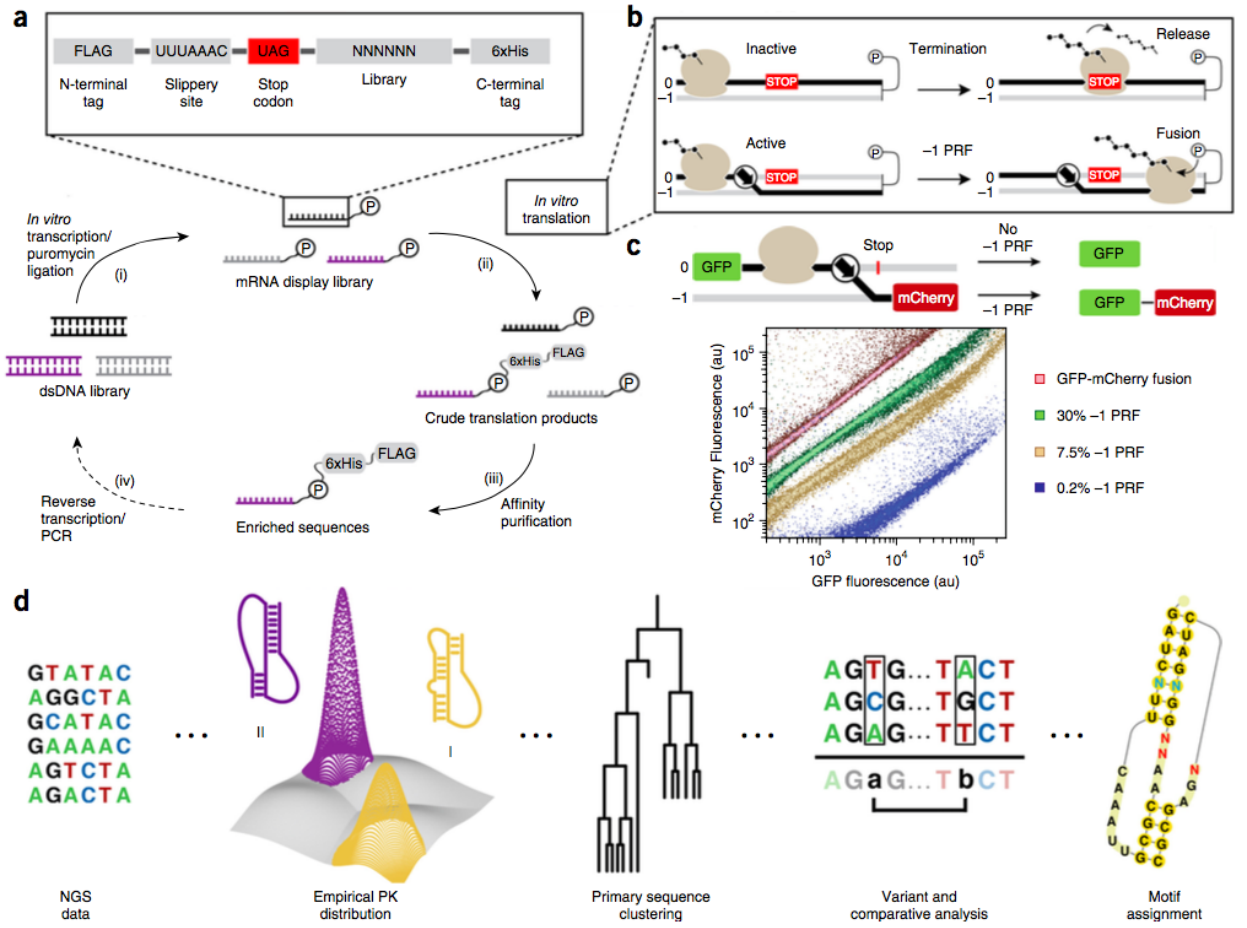


Figure 6.2: **Overview of directed evolution experiment:** (a) The four stages of the mRNA display selection cycle are shown. (b) Translation reprogramming selection principle. Ribosomes that terminate translation upstream of the designated fusion point will fail to produce mRNA-peptide fusions (upper). Frameshifting enables bypass of encoded stop codons (lower). (c) Dual-FP reporter assay in S. cerevisiae. The frameshift variant is cloned between a green fluorescent protein (GFP) and the red fluorescent protein variant mCherry. The ratio of FP signals reflects bulk -1 PRF efficiency. Flow cytometry of individual clones harboring -1 PRF stimulatory elements of varying efficiencies is shown ('au' stands for 'arbitrary units'). (d) High-throughput sequencing workflow for library selection products. Selected sequences are grouped into pseudoknot (PK) families, analyzed for post-selection enrichment, and clustered based on primary sequence identity. Motifs can be analyzed by comparative analysis, or individual sequences can be analyzed for single and pair-wise nucleotide changes.

generating a set of 2,068 individual PK features (Figure 6.3). We enumerate the full set of unique sequences theoretically present in the initial library and calculate their compatibilities with the 2,068 PK features. The feature with the greatest amount of base pairing (including G/U) is associated to each sequence. PK compatibilities are similarly calculated for the sequences observed in the *in vitro* selection products. Highly enriched PK features could then be measured by comparing the two distributions. Of the top 10% of PK features by enrichment, those with the highest absolute representation in the selection library defined the broad motif categories. Within a set of PK-compatible sequences, a greedy clustering algorithm is used to divide the set of sequences into the final motifs. The modes of the different motif sets are further characterized in terms of the abundances and entropies of their immediate neighborhood in sequence space. In particular, the modes from the highest abundance motifs are assessed for single-nucleotide and pairwise nucleotide variant sensitivity to reveal potential base pair position and tertiary interactions within a given PK geometry.

A more granular description of the pipeline for processing and analyzing the sequencing products will now be given, with the final goal being the identification of high fitness motifs evolved in the RNA population. Step 1 of the pipeline beings by processing the raw sequencing fastq file and generates a file containing every unique sequence and its associated read count. The HiSeq run yielded roughly 56 million sequencing reads after processing, from which >3 million unique sequences are recovered. Abundances (or read counts) of these sequences ranged from 1 read to >800,000 reads. 50% of the total sequencing reads are accounted for by the top 6,752 most abundant unique sequences. We set out to identify -1 PRF motifs from this dataset with a specific focus on RNA pseudoknots (PKs), which are known to promote -1 PRF. To assign a PK structure to each sequence, we chose to assess compatibility with PK configurations without accounting for energetic contributions. Because this approach is tailored to our restricted scaffold, it is computationally more efficient than absolute prediction. Notably, this approach often leads to PK assignments that are in agreement with pseudoknot prediction algorithms (such as pKiss [99]). We created

a hairpin-type (H-type) PK feature space using the starting library's sequence constraints and our imposed structural constraints. Every H-type PK can be defined by the lengths of its segments: (a) stem 1; (b) loop 1; (c) stem 2; (d) loop 2; and (e) loop 3 (Figure 6.3). To establish the PK position within the library scaffold, we also define the length of the segment of unpaired nucleotides preceding (5' to) the PK, the length of the segment downstream of (3' to) the PK, and fix the total length to 44 nucleotides.

Using the library scaffold's constraints and our imposed constraints, a total of 2,068 PK features are generated. The top 20,000 most abundant sequences from the high-throughput sequencing, as well as every theoretical starting library sequence, are evaluated for their compatibility with the PK feature space (step 2 of the pipeline). A sequence is deemed compatible with a given PK feature if it supported the base pairs present within the PK feature (allowable base pairs include the standard Watson-Crick A-U, U-A, G-C, and C-G, along with the wobble G-U and U-G pairs). Of the PK features supported by a given sequence, the feature with the most base pairs is chosen as the assigned PK feature for that sequence. In the case of a tie, both PK features are assigned to the sequence. From the theoretical starting library, 1,507 PK features are selected by at least 1 sequence.

In step 3 of the pipeline, we determined the enrichment of each PK feature by calculating its pre-selection probability and comparing it to its post-selection probability (Figure 6.4). Explicitly, pre-selection probability is defined as the total number of sequences from the theoretical starting library that are compatible with the PK feature divided by the total starting sequences ($\sim 2.68 \times 10^8$). Post-selection probability is defined as the total number of sequencing reads that fit to the PK feature (number of unique sequences in that feature multiplied by the mean read-count of sequences within that PK feature) divided by the total sequencing reads ($\sim 5.6 \times 10^7$). The enrichment factor (EF) is then defined as the ratio of the post-selection probability to the pre-selection probability. Each PK feature is ranked according to its EF, and PK features that showed low occupancy (low total read mass) are removed. Then, PK features ranked within the top 10% of EFs are nominated for primary

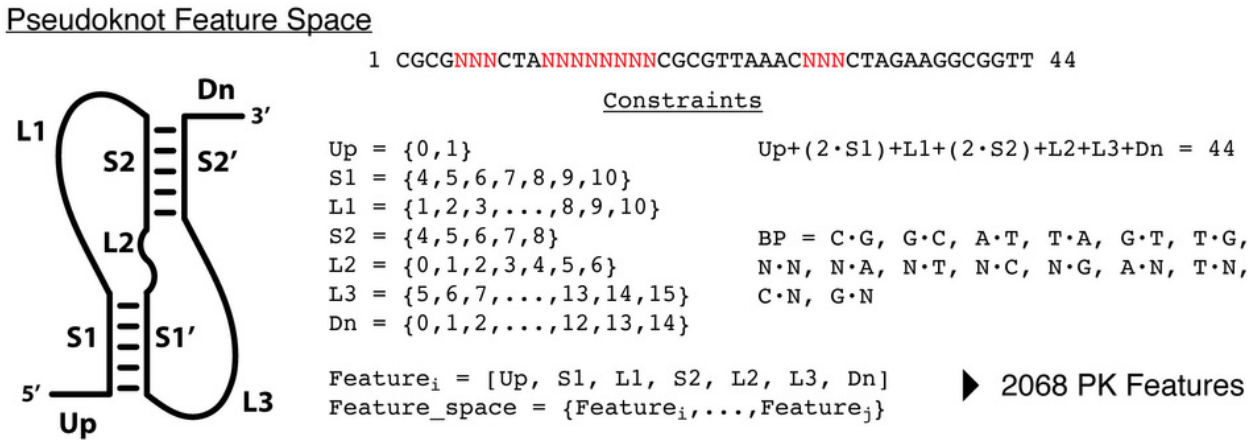


Figure 6.3: **Degrees of freedom for pseudoknots:** H-type pseudoknots can be defined by the lengths of their segments, specifically their loops (L1, L2, and L3) and stems (S1 and S2). To fix the position of the pseudoknot within the library scaffold, an upstream (Up) and downstream (Dn) segment is also defined. The sum of these strand lengths (stem segments are double stranded) must equal the scaffold's total length of 44 nucleotides. All combinations of stand lengths that obey the above constraints are generated, and then assessed for base pairing compatibility using the scaffold sequence (where N can take on any nucleotide identity). The resulting PK features that satisfy these requirements define the set of all possible PKs. For this set of constraints and scaffold there are 2,068 PK features.

sequencing clustering analysis. Primary sequence clustering is performed using a greedy algorithm, resulting in families of sequences that comprise the motifs. Clusters containing less than 5 sequences are discarded or ignored. From this analysis 115 clusters are nominated as final -1 PRF motifs, the top few cases of which are presented in Table 6.1.

6.4 Engineering Frameshifting Riboswitches

Previous studies [230, 93] *in vitro* and in mammalian cell culture demonstrated the feasibility of small-molecule-regulated -1 PRF using metabolite sensing transcriptional riboswitches that adopt frameshift stimulatory pseudoknot conformations in the presence of their cognate ligands. However, the ligand-binding domains of these bacterial riboswitches are integral components that cannot be exchanged with other ligand-binding RNA aptamer domains and are not easily modified to recognize entirely new ligands. As a result, no general design strategy currently exists for assembling synthetic -1 PRF devices that respond exclusively

Table 6.1: **Predicted frameshifting PK geometries:** The motif in red is chosen for downstream engineering applications. The rows of this table do not describe the evolved population in terms of nucleotide sequence, but rather in terms of compatibility with a particular PK geometry. The set of RNAs supporting each PK class are described in terms of their size (N_{seq}), average fitness ($\overline{copy\_number}$), and localization in sequence space (normalized entropy in sequence neighborhood $H_{neighbor}$). Confirmation of predicted PK structure matching the compatibility class is performed with the pKiss software [99].

PK Compatibility	MFE Structure of Sequence Mode	N_{seq}	$\overline{C.N.}$	$H_{neighborhood}$
0|4|6|7|1|7|8	`[[[[......{{{{{{{.]]]].......}}}}}}}.....`	20	22593	0.335
0|4|6|7|1|6|9	`[[[[......{{{{{{{.]]]].......}}}}}}}.....`	5	18737	0.494
0|4|2|8|4|6|8	`[[[[..{{{{{{{{....]]]].......}}}}}}}}.....`	124	17966	0.551
0|4|2|7|5|7|8	`[[[[..{{{{{{{.....]]]].......}}}}}}}.....`	87	15640	0.684
0|4|3|7|4|6|9	`[[[[...{{{{{{{....]]]].......}}}}}}}.....`	127	11725	0.753
0|4|7|6|1|7|9	`[[[[.......{{{{{{.]]]].......}}}}}}.....`	6	10129	0.610
0|6|1|7|2|6|9	`[[[[[[.{{{{{{{..]]]]]].......}}}}}}}.....`	16	8308	0.679
0|4|5|7|2|14|1	`[[[[.....{{{{{{{..]]]].............}}}}}}}.`	12	6454	0.585
0|6|1|6|3|7|9	`[[[[[[.{{{{{{...]]]]]].......}}}}}}.....`	128	6178	0.852
0|4|5|7|2|14|1	`[[[[.....{{{{{{{..]]]].............}}}}}}}.`	9	6065	0.579
0|5|5|7|0|7|8	`[[[[[.....{{{{{{{]]]]].......}}}}}}}.....`	6	5435	0.717
0|5|4|7|1|14|1	`[[[[[....{{{{{{{.]]]]].............}}}}}}}.`	17	5193	0.725
0|7|2|6|0|15|1	`[[[[[[[..{{{{{{{]]]]]]].............}}}}}}.`	19	4991	0.852
0|4|3|7|4|6|9	`[[[[...{{{{{{{....]]]].......}}}}}}}.....`	6	4893	0.742
0|4|3|7|4|6|9	`[[[[...{{{{{{{....]]]].......}}}}}}}.....`	31	4671	0.830
0|6|1|7|0|8|9	`[[[[[[.{{{{{{{]]]]]].......}}}}}}}.....`	11	4586	0.797
0|7|2|6|0|15|1	`[[[[[[[..{{{{{{{]]]]]]].............}}}}}}.`	391	4156	0.867
1|6|9|5|5|6|1	`.[[[[[[.........{{{{{.....]]]]]].....}}}}}.`	459	3867	0.914
0|6|1|6|3|7|9	`[[[[[[.{{{{{{...]]]]]].......}}}}}}.....`	21	3437	0.810
0|4|3|7|4|6|9	`[[[[...{{{{{{{....]]]].......}}}}}}}.....`	11	3125	0.808

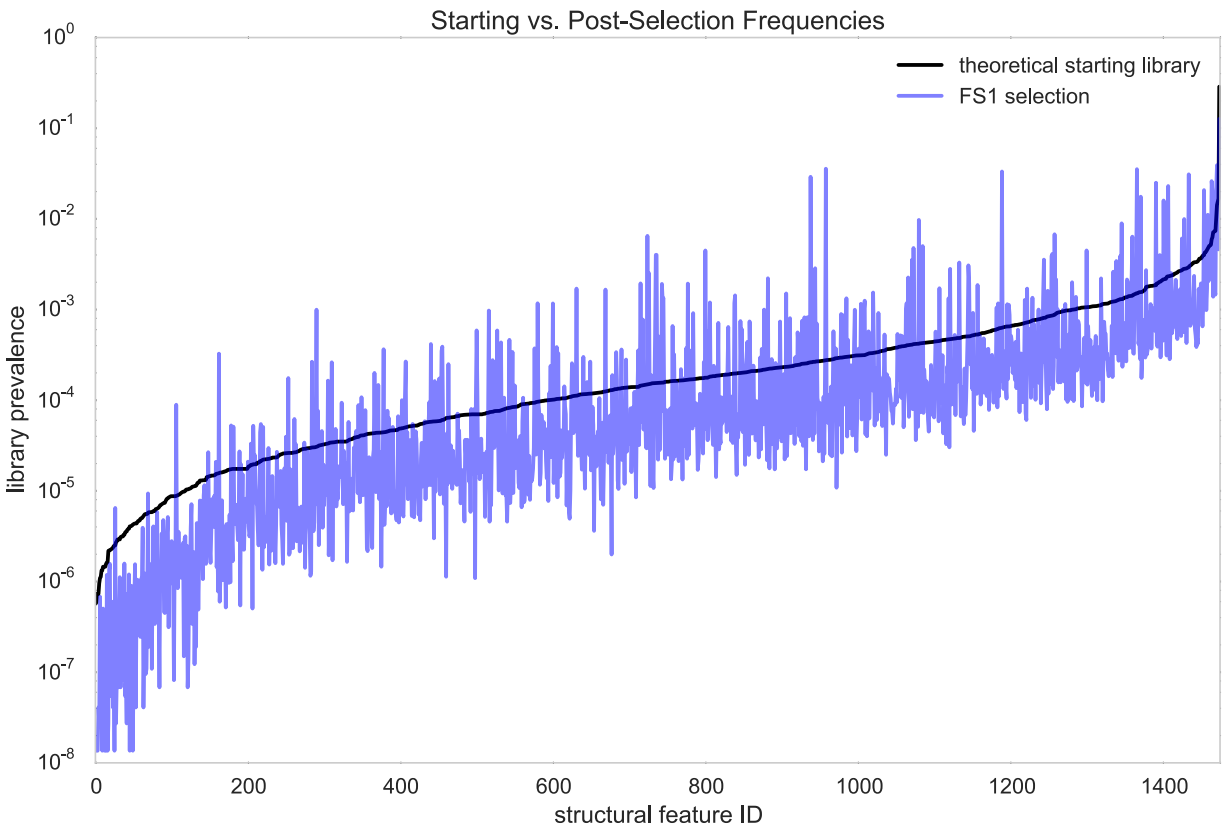


Figure 6.4: **Enrichment among PK compatibility equivalence classes:** The PK feature probability is calculated from the theoretical starting library (assuming equal distribution of sequences) by dividing the number of sequences within a PK feature by the starting library size ($\sim 2.68 \times 10^8$). Post selection probability is computed by dividing the number of sequencing reads contained within the PK feature by the total sequencing reads ($\sim 6 \times 10^7$).

to an orthogonal small molecule of choice.

To engineer ligand-responsive devices, we pursued a modular strategy of rationally coupling -1 PRF signals to small- molecule-binding RNA aptamers. For the stimulatory element, we chose sequence illustrated in Figure 6.5 and the top row of Table 6.1. This sequence was the second most abundant in the population at the time of sequencing, it displayed high -1 PRF efficiency in yeast (30%), and it has a confidently predicted pseudoknot fold. As aptamer domains, we chose the theophylline [100] and neomycin [218] binding aptamers based on their previous successes in in vivo applications [33].

In the OFF switch design (Figure 6.6), the RNA aptamer and pseudoknot sequences

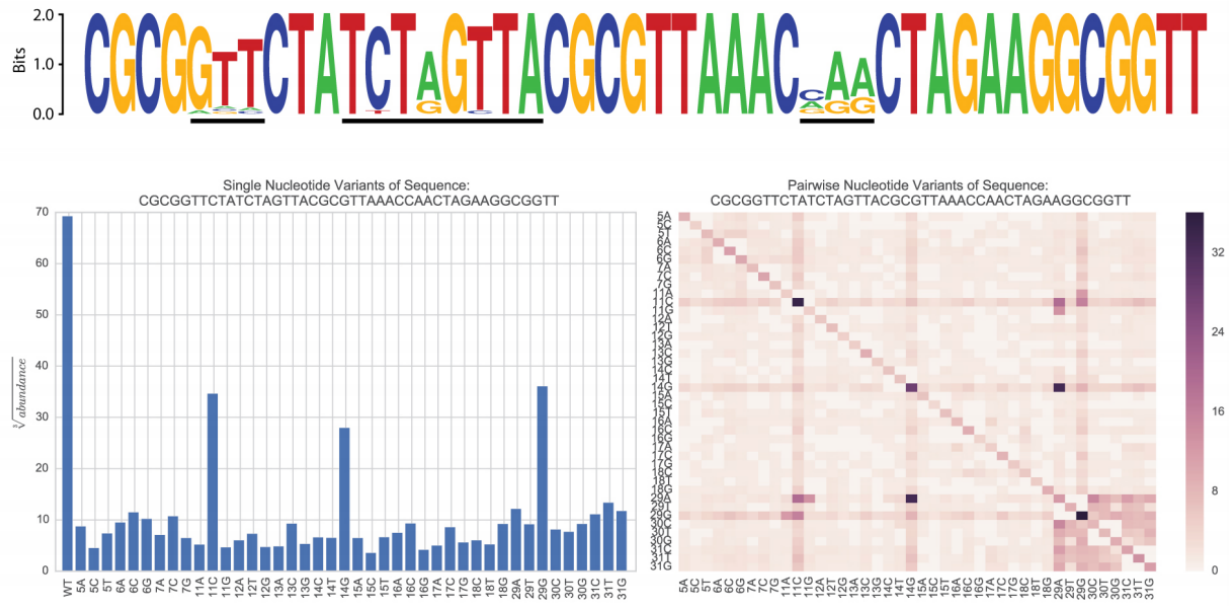


Figure 6.5: **Top sequence for PK selection:** The analysis pipeline outputs various statistics for each RNA (see Table 6.1), as well as a sequence logo [43] (underlined segments represent regions of randomized nucleotides in the starting library). Illustrated above is our selection of the best PK sequence in the population. It can be analyzed for single or pairwise nucleotide variants to identify mutation tolerant or intolerant positions, and reveal potential mutual information.

overlap and thus compete for folding. In the absence of ligand, the active pseudoknot predominates and stimulates high -1 PRF activity. However, in the presence of ligand, the folded aptamer is stabilized by ligand binding energy and disrupts the pseudoknot, resulting in lowered -1 PRF activity. We designed several constructs by varying the length and composition of the aptamer stems. As a general trend, we found that increasing the thermodynamic stability of the aptamer lowered frameshift activity. This simple approach led to the discovery of a high-performing theophylline OFF switch that displays a sevenfold reduction in -1 PRF in the presence of ligand.

To engineer ON switches, a 'switching hairpin' was introduced to compete with the pseudoknot folding and to reduce basal -1 PRF levels. By design, structural rearrangements stimulated by ligand-aptamer recognition serve to destabilize the switching hairpin, leading to coincident refolding of the pseudoknot and restoration of -1 PRF activity (Figure 6.6). Several constructs were created to tune the relative stabilities of the ON and OFF states by

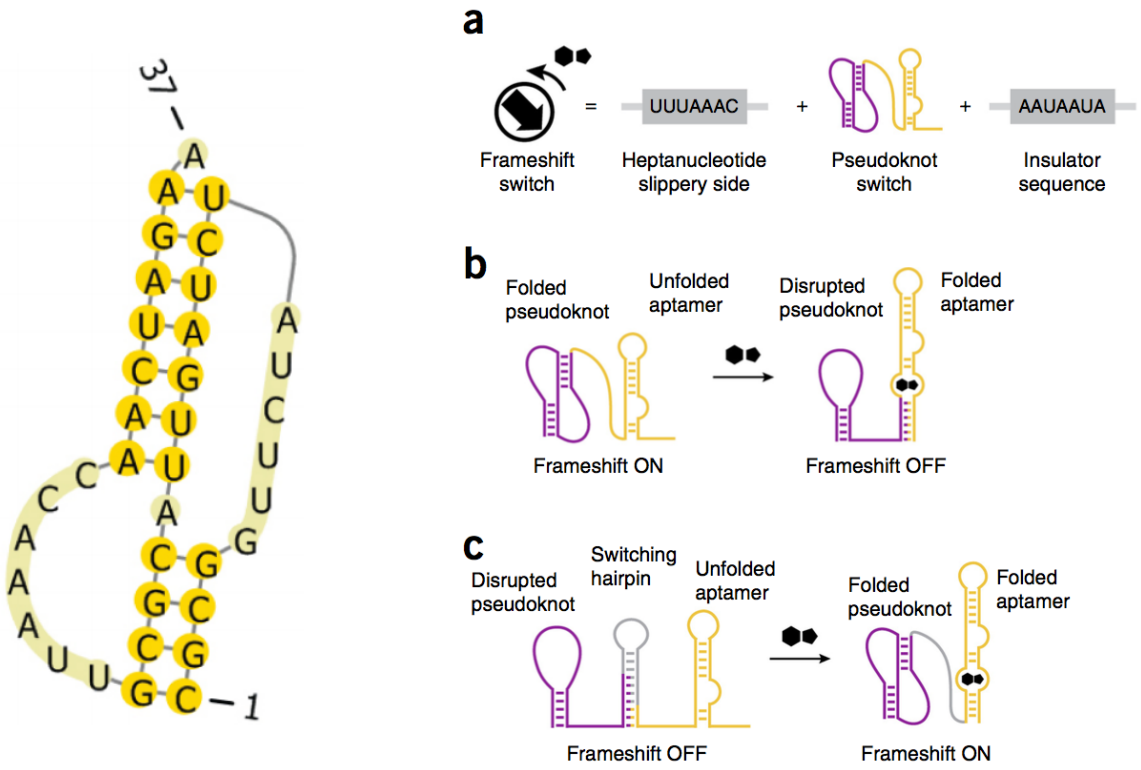


Figure 6.6: **Ligand-responsive on/off conformations:** (Left) The PK structure selected for downstream riboswitch engineering. This structure is described in more granular detail in the top row of Table 6.1. (Right) (a) Architecture of frameshift switch devices. (b) OFF switch design. In the absence of ligand, the stimulatory pseudoknot (purple) is energetically dominant, producing high frameshift levels. Ligand binding induces aptamer (gold) folding, which disrupts the pseudoknot structure, leading to lowered frameshift levels. (c) ON switch design. A switching hairpin (gray) is installed to disrupt the pseudoknot and lower basal frameshifting. In the presence of ligand, the aptamer folds and destabilizes the switching hairpin, allowing the pseudoknot to re-fold and restore frameshift activity.

varying the lengths and compositions of the aptamer stems and switching hairpins. These constructs were assessed using the NUPACK [231] RNA secondary structure prediction and tested experimentally in the dual-FP assay to correlate hairpin and aptamer stability to switch activity. With minimal optimization, this approach led to efficient ON switches that respond to theophylline (5.9-fold) or neomycin (4.2-fold).

6.5 Conclusions

-1 Programmed ribosomal frameshift (PRF) is a naturally occurring phenomenon with great promise in bioengineering applications. In this chapter we evolved high efficiency PRF structures via a randomized starting RNA population and an *in vitro* selection approach. Due to the computational expense of performing pseudoknot predictions, we devised a parametrization of all the pseudoknot geometries accessible from the PreQ1 scaffold and calculated sequence compatibility with each of those geometries. Our procedure for finding enriched, high efficiency genotypes entailed explicitly calculating the probability distribution over the pseudoknot geometries pre- and post- evolution. The genotype with the highest fitness was chosen for the construction of ligand-responsive riboswitches, which ultimately were used to build RNA logic gates.

Chapter 7

Evolving Sequence Contexts That Promote Stop Codon Readthrough

In this chapter we describe a directed evolution approach for the discovery of stop codon readthrough *cis* stimulatory element. The starting population of RNA genomes are less biased than in Chapter 6, as they contain no scaffolding and a full 75 randomized positions. High-throughput sequencing is leveraged as a fitness read-out for a population of candidate RNAs, where copy number in the overall library is a proxy for RNA fitness. A machine learning model is the trained using sequence and structural features to discern high efficiency readthrough motifs from decoy RNAs. An application in the prediction of stop codon readthrough in human 3' UTR sequences is explored. U (uracil) and T (thymine) are used somewhat interchangeably throughout the chapter, with the understanding that the original sequences are RNA while their characterization by deep sequencing requires the construction of a cDNA library.

Material presented in this chapter will be published, wholly or in part, in the manuscript "Large scale profiling of stop codon readthrough RNA signals and their identification in human 3'-UTRs" (in preparation) in collaboration with A.V. Anzalone, A.J. Lin, R. Rabadan, V.W. Cornish

7.1 Stop Codon Readthrough: Natural and Engineered

Alternative modes of translation are common in the world of viruses, whose limited genome size implies extreme seletion pressures at each position. Reprogramming canonical translation can expand the coding capacity of a nucleic acid. Non-canonical translation generally divies into two categories: initation-related mechanisms (internal ribosome entry, leaky scanning, non-AUG initiation, reinitiation) and elongation mechanisms (frameshifting and readthrough). Stop codon readthrough (RT) is especially interesting because it is not limited to RNA viruses, having recently been observed in fungi [72] and metazoa [104]. Both *cis*-factors (sequence and structural elements on the transcript) and *trans*-factors (cytoplasmic proteins or metabolites) may be involved in the RT phenomenon, complicating the dissection of exact biochemical mechanisms. Even in the simpler case of *cis*-factors, we know from viruses that RT can be mediated by sequence (tobacco mosaic virus) or structure (murine leukemia virus). Structural motifs are, of course, more difficult to detect purely from the reference genomes of organisms.

The first bonafide RT-promoting motif was discovered in the tobacco mosaic virus (TMV) [164] and related plant viruses [191], implying that RT may be a physiological feature of eukaryotic biology. The consensus TMV motif is purely sequence-based, CARYYA, where R = {A,G} and Y = {C,T}. Structural motifs have also been discovered and characterized, for example a stem-loop 8 nucleotides downstream of a UGA stop codon in Colorado tick fever virus [143] or a pseudoknot 8 nucleotides downstream of a UAG stop codon in Moloney murine leukemia virus [221]. In general UGA is believed to be the most leaky stop codon and UAA the least. VEGFA, a human gene with a UGA stop codon, was recently shown [59] to undergo RT to produce an anti-angiogenic isoform relevant to colon adenocarcinoma.

As the list of genes and species exploiting RT expands, we are faced with the important technical challenge of predicting RT from sequence and ultimately understanding its

biochemical determinants. In the fields of bioengineering and chemical biology it is of great interest to incorporate unnatural amino acids [126] at particular codons. Elucidating the natural mechanisms for incorporating near-cognate amino acids at stop codons (the very definition of RT) would seemingly empower the engineering of translation more effectively. Two existing methodologies for identifying *cis*-RT elements are based on primary sequence features alone: the first calculates under-utilized nucleotide contexts surrounding stop codons [219], while the other detects sequence conservation and protein coding potential downstream of stop codons in the 3' UTR [104]. Both of the these approaches make strong assumptions, such as a) evolutionary conservation of the 3' extension or b) even that RT is a rare event whose signature should be depletion of depletion of certain nucleotide contexts from negative selection. In this chapter we prefer to take the approach of evolving RT from a population of randomized starting RNAs. We leverage the same mRNA display technology detailed in Chapter 6, combining it with a eukaryotic cell lysate for an *in vitro* translation system.

7.2 Design and Read-Out of *In Vitro* Evolution

To enrich eukaryotic RT motifs by *in vitro* selection, we design a strategy based on mRNA display, which covalently links translation products to their encoding mRNAs [174]. We demonstrated in the previous chapter that mRNA display can be used to identify eukaryotic -1 programmed ribosomal frameshifting motifs from large libraries of RNA sequence variants when translated in rabbit reticulocyte lysate [9]. With such a strategy, libraries of RNA variants can be cycled through multiple rounds of selection, enriching at each stage for the desired reprogramming behavior (Figure 7.1). Because selection relies on a distance-dependent puromycin fusion reaction [127], mRNA display efficiently selects for translation reprogramming signals that enable the ribosome to bypass in-frame termination codons.

We first validate the mRNA display selection platform by establishing depletion of stop

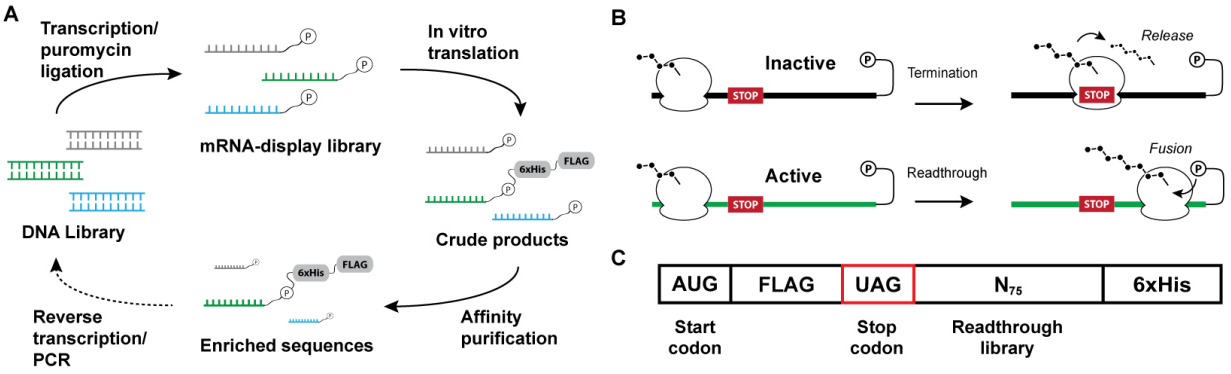
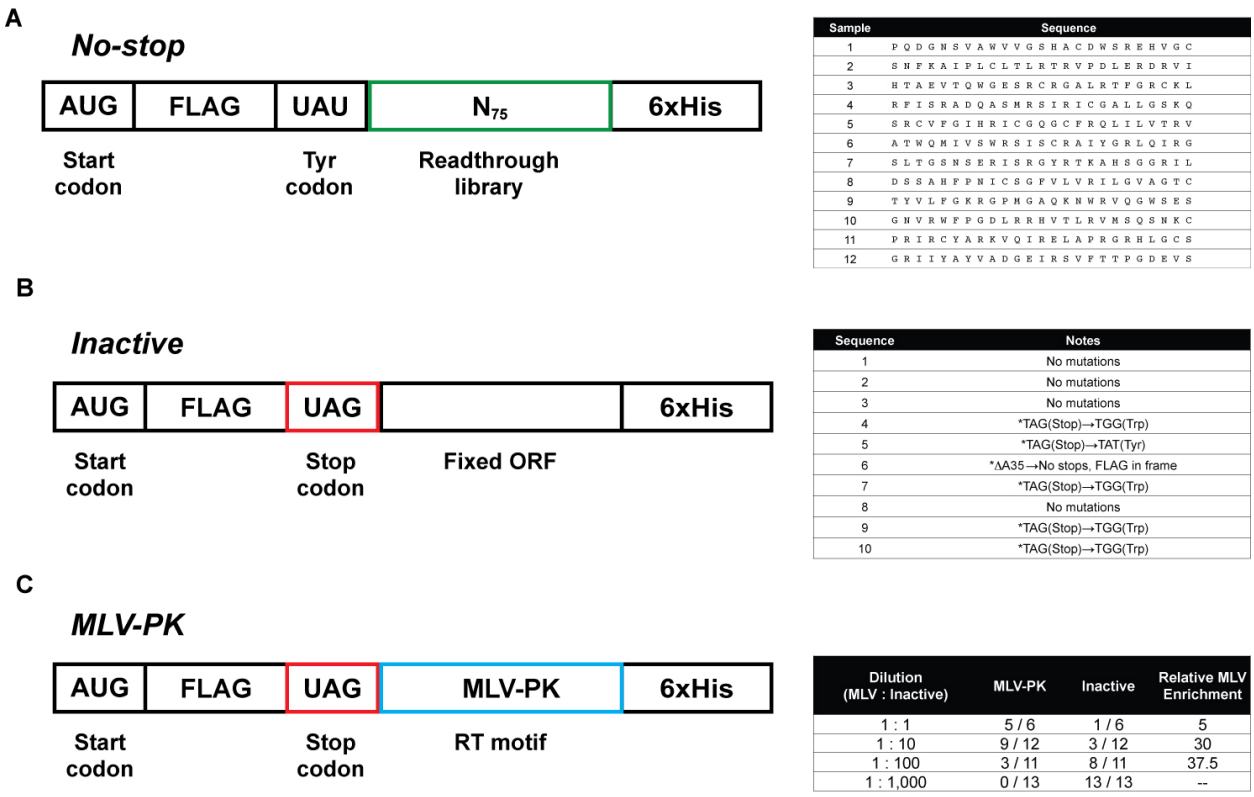


Figure 7.1: **Directed evolution conditions and the selection construct:** Three rounds of selection, purification, and reproduction are performed upon an initial population of randomized RNAs. A rabbit reticulocyte lysate is used as the *in vitro* translation system. (A) Purification is achieved via α-FLAG antibody as well as a nickel column. Reproduction is implemented via a noisy PCR to permit the acquisition of mutations during expansions following population bottlenecks. (B) Selection is based upon the RNA's ability to induce a stop codon readthrough event, which permits the full translation product including hexa-His tag to be produced. A puromycin reaction is used to covalently link the RNA genotype with its peptide product. (C) The selection construct is highly unbiased in that it contains 75 fully randomized nucleotides 3' to the stop codon.

codon containing sequences. A single round of selection is performed on a library of mRNAs containing 75 randomized nucleotides (70% probability of at least one in-frame stop codon). Consistent with prior studies [35], we observe complete depletion of stop codon containing mRNAs in sampled selection products (Figure 7.2A). To ascertain the fate of mRNA transcripts that contain a programmed stop codon but no readthrough signal, one round of selection was performed on a sequence containing a single in frame UAG stop codon. Impressively, sequences containing mutations to the stop codon were efficiently enriched, with more than half of the sampled sequences mutating around the in frame stop codon (Figure 7.2B).

To demonstrate that authentic reprogramming signals can be enriched from a background of inactive sequences, mock selections were performed using the murine leukemia virus (MLV) stimulatory pseudoknot [92, 221]. Selections were conducted by varying the input ratio of the active MLV mRNA to the inactive control mRNA. The results revealed that the active MLV sequence is significantly enriched at an estimated level of approximately 30-fold during



Sample	Sequence
1	PQDGNSVAWVVGSHACDWSREHVGC
2	SNFKAIPLCLTLRTRVPDLERDRVI
3	HTAEVTQWGESRCRGALRTFGRCKL
4	RFISRADQASMRSIRICGALLGSKQ
5	SRCVFGIHRICGQGCFRQLILVTRV
6	ATWQMIVSWRSISCRAIYGRLQIRG
7	SLTGSNSERISRGYRTKAHSGGRIL
8	DSSAHFPNICSGFVLVRILGVAGTC
9	TYVLFGKRGPMGAQKNWRVQGWSES
10	GNVRWFPGDLRRHVTLRVMSQSNKC
11	PRIRCYARKVQIRELAPRGRHLGCS
12	GRIIYAYVADGEIRSVFTTPGDEVS

Sequence	Notes
1	No mutations
2	No mutations
3	No mutations
4	*TAG(Stop)→TGG(Trp)
5	*TAG(Stop)→TAT(Tyr)
6	*ΔA35→No stops, FLAG in frame
7	*TAG(Stop)→TGG(Trp)
8	No mutations
9	*TAG(Stop)→TGG(Trp)
10	*TAG(Stop)→TGG(Trp)

Dilution (MLV : Inactive)	MLV-PK	Inactive	Relative MLV Enrichment
1 : 1	5 / 6	1 / 6	5
1 : 10	9 / 12	3 / 12	30
1 : 100	3 / 11	8 / 11	37.5
1 : 1,000	0 / 13	13 / 13	--

Figure 7.2: **Directed evolution control enrichments:** (A) No sequences contain in-frame stop codons after a single round of evolution, demonstrating purifying selection. (B) With no randomized positions 3' in the vicinity of the stop codon, the easiest fitness-conferring mutation for the contruct is to alter the stop codon itself. Profound enrichment of mutations at the stop codon is seen after one round of evolution, demonstrating strong positive selection. (C) The MLV pseudoknot structure is a bonafide RT promoting element, and its enrichment is measurable in a larger population of inactives at different starting concentrations.

a single round of selection (Figure 7.2C). Together, these results confirm that the in vitro mRNA display selection is capable of responding to RT motifs, and that the selection strategy is capable of enriching active RT elements from a pool composed largely of inactive sequences.

We next set out to enrich RT motifs *de novo* from a large library of randomized RNA sequences. Because RT stimulating elements can exist in the nucleotides directly adjacent to the stop codon as well as several nucleotides downstream as hairpin or pseudoknot structures, we choose to encode a library of 75 randomized nucleotides directly adjacent to a UAG codon. Though this vast theoretical library size ($\sim 10^{45}$) has exceedingly low coverage from the 10^{13} unique RNA molecules submitted to the first round of selection, we anticipate that diverse

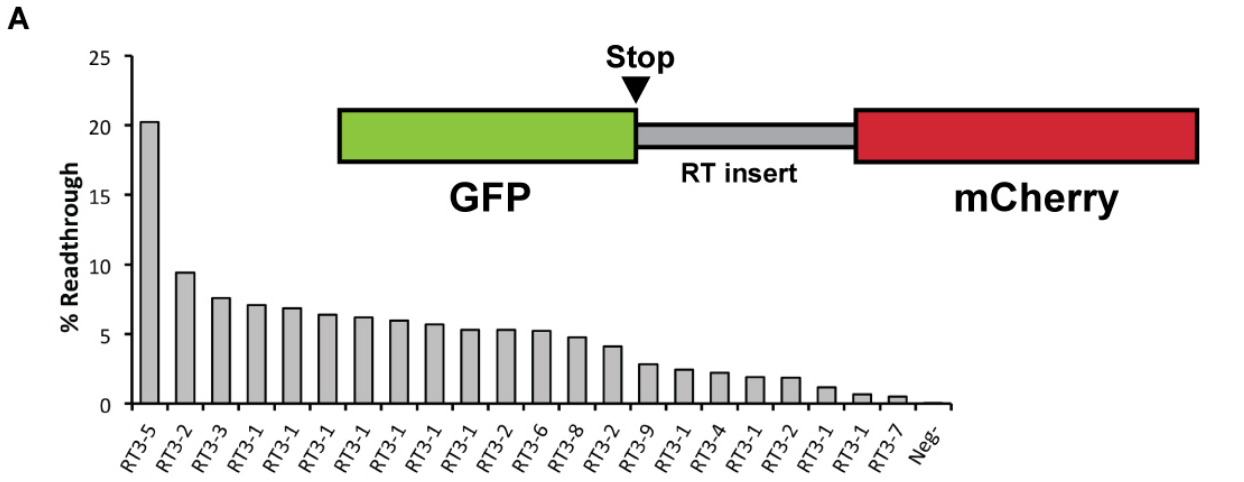
stimulatory structures will be accessible from this library.

After the initial round of selection, sequences containing mutations to the stop codon were enriched, as was observed in the earlier control selections. To prevent further enrichment of these sequences, primers used for PCR amplification were designed to contain the entire 5' segment of the selection construct leading up to and including the stop codon. Following three rounds of selection, cloning and sequencing of a small sampling of the library shows enrichment of a hexanucleotide motif directly adjacent to the UAG stop codon with a consensus sequence of CAAYYA. Significantly, this conforms to the well-described TMV motif and the previously identified CARNBA motif from yeast [142], suggesting that sequences promoting stop codon readthrough have been enriched.

To study the RT activity of the selection products *in vivo*, we implement a dual-fluorescent protein (dual-FP) reporter assay in saccharomyces cerevisiae, analogous to the reporter described in the previous chapter (Figure 7.3A). Post-selection library members are cloned between GFP and mCherry ORFs, and readthrough activity is assessed by the ratio of GFP and mCherry fluorescence signals. Sampled yeast colonies demonstrate a range of RT activity, with the most efficient sequence reaching nearly 20% RT (Figure 7.3B). Notably, while many sequences contain the CAAYYA motif, these sequences do not always exhibit equivalent readthrough efficiencies. This suggests that factors beyond the hexanucleotide motif are contributing to RT levels in the enriched sequences.

7.3 Modeling Readthrough Efficiency from Sequence and Structure

In order to probe the full population structure of RNAs after 3 rounds of evolution, we perform Illumina sequencing at a depth of $\sim 10^8$ reads with 100bp read length. At this length we capture the entire RNA genome on a single read, which preempts the issue of haplotyping mutations and dramatically increases our statistical power. Interestingly, a sig-



B

Sequence	%RT
CAATTAGCCT CCAGCATTCC TGGGAAAAGT CGTGGAGTTA CCTCGCATGG TGATGTTGCT CTCGTGTCAC GTACG	20.2
CAATTAGAAC GGTCTCCCAA TATTCAGAAT GTCGGGGGAC CTGGGAGCCT GGTAGAGCCA GGTGTTACGC CTCTG	9.4
CAATTAGTAC CAGGCATGGA TCGGCAACTA TTCGATCCTG CGCTACGTTG GGAGGCTGGT AATCGGTCTG CACGG	7.6
CAATCAAGGA GCCGGATATT CCCTGCTGAT CAGAGCGGCT TCATTTACCG CACAGCGCTT GTGGCCAATG GTGAG	7.1
CAACTAGCTC CGCCAGCGGG TATTGCACAT GATGAACCAG TGTGTAGTTC AGCTAGTGGT GTGGCGTCTC GGAGG	6.9

Figure 7.3: **Individual colonies assayed post-selection:** A dual fluorescent protein reporter cassette is used to assay the RT efficiencies of indiviual 3' sequence contexts. This plot confirms that highly enriched sequences from the directed evolution experiment are functional *in vivo*, as we detect non-zero levels of the mCherry channel for multiple constructs. The fluroescence detection lower limit in this reporter system is ∼0.2%.

nificant fraction of the reads encode an internal methionine followed by a FLAG-like peptide sequence which we interpret as a site for internal initiation of translation downstream of the stop codon. Such an adaptation would be highly fitness-conferring since it would enable that genotype to survive affinity purifications and propagate to subsequent generations. These sequences are separated in our computational analysis and will form a negative training set, or "decoy" set. Essentially these are sequences that likely have no ability to promote RT and are cheating the selection mechanism.

The first and easiest analysis is primary sequence, where we calculate enriched or depleted codons along the length of the contruct. The first codon downstream of the stop (position +1 to +3) shows profound enrichment for CAA, with CAG, CCA, and CTA also highly represented. The second codon is highly enriched for all combinations of YYA, along with

CGA. The third codon shows some enrichment for CAG. These results, as well as the correlations between codons at different positions, are plotted in Figure 7.4. The edge weights in the network visualization encode the degree of overlap between codons at each position, and we see that the strongest correlation is between 1_CAA and 2_TTA or 2_TCA. This clearly recapitulates the naturally occurring TMV motif in the first 6 nucleotide positions, but with the added resolve epistatic interactions in the proximal nucleotide context. We extend our calculations of codon usage beyonf the 3 most proximal codon positions, and the traces can be seen in Figure 7.5. It is known that cytosine at the third position is not preferred in mammalian codons [60], and our hypothesis of likely bias in the codon usage of the *in vitro* translation system is confirmed by the data. The traces all share a 4-letter periodicity with a wobble cytosine associated with decreased fitness. The over-represented codons correspond to the amino acids aspartic acid or glutamic acid (GA*), glycine (GG*), and valine (GU*).

Next we examine the ensemble of RNA secondary structures formed in the high copy number RT genotypes as well as the decoys. Unlike in Chapter 6 where we focused on H-type pseudoknots, in this analysis we focus exclusively on stem-loop prediction which is significantly more efficieny computationally. We use the ViennRNA 2.0 software [129] to perform stem prediction (the "rnafold" utility was used) with default parameters. It is worth mentioning that we first append the 60 nucleotides of fixed RNA scaffold downstream of the 75 randomized positions because applying RNAfold, since it is quite common for the randomized library to base pair with fixed sequence downstream. The predicted structures are represented in the standard dot-brack format, where '.' denotes an unpaired nucleotide and '(' and ')' denote base paired positions in the sequence. RNAfold also predicts the minimum free energy (MFE) for each ensemble of structures, the probability of base pairing at each position, and the entropy at each position. Summary distributions on these quantities are displayed in Figure 7.6, where we stratify the RNAs into the decoy set (negative examples, cheated the selection mechanism) and the RT set (proper high-efficiency RT motifs). The

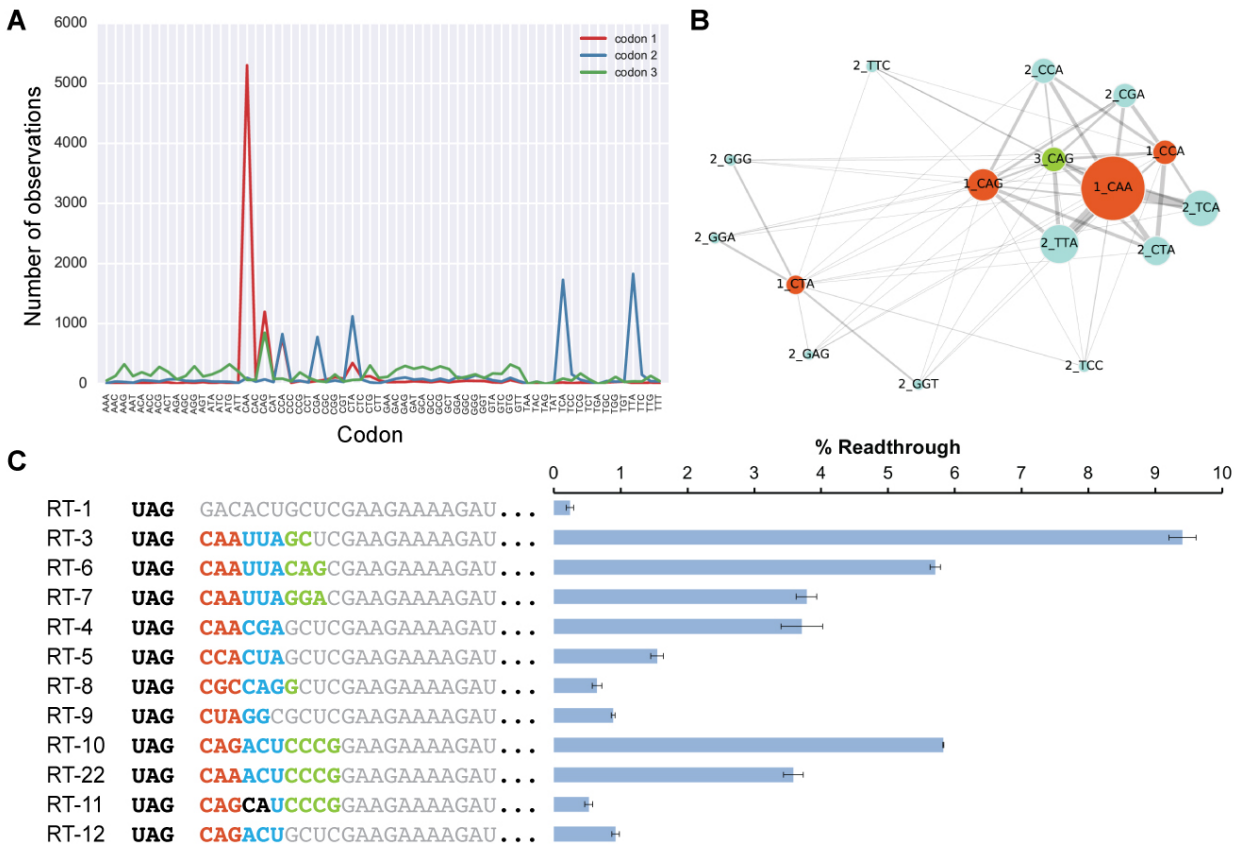


Figure 7.4: **Sequence enrichment proximal to stop codon:** (A,B) A statistical analysis of the first 9 nucleotides immediately downstream of the stop codon reveals striking enrichments and correlations. The first 3 positions are dominated by the motif CAA, while the next 3 positions are highly enriched for YYA. (C) Targeted experimental perturbations in the sequence neighborhood of the motifs generated by the primary sequence analysis.

distributions are quite resolved for the MFE, with the RT set possessing more stable stems and therefore lower MFE. Similarly, the distribution of positional base pair probabilities is more peaked near 1.0 for the RT set than the decoy set. Most interesting though, is the position at which stem formation starts across the two sets. The decoy set has an unremarkable distribution of start positions, monotonically decreasing as we move along the genome. The RT set, on the other hand, shows a characteristic spike in probability around the 8th nucleotide. We recall that the structures in Colorado tick fever virus and Moloney murine leukemia virus started at the 8th nucleotide. This association is not surprising as this position is close to the mRNA entry tunnel on the ribosome, as determined by ribosome

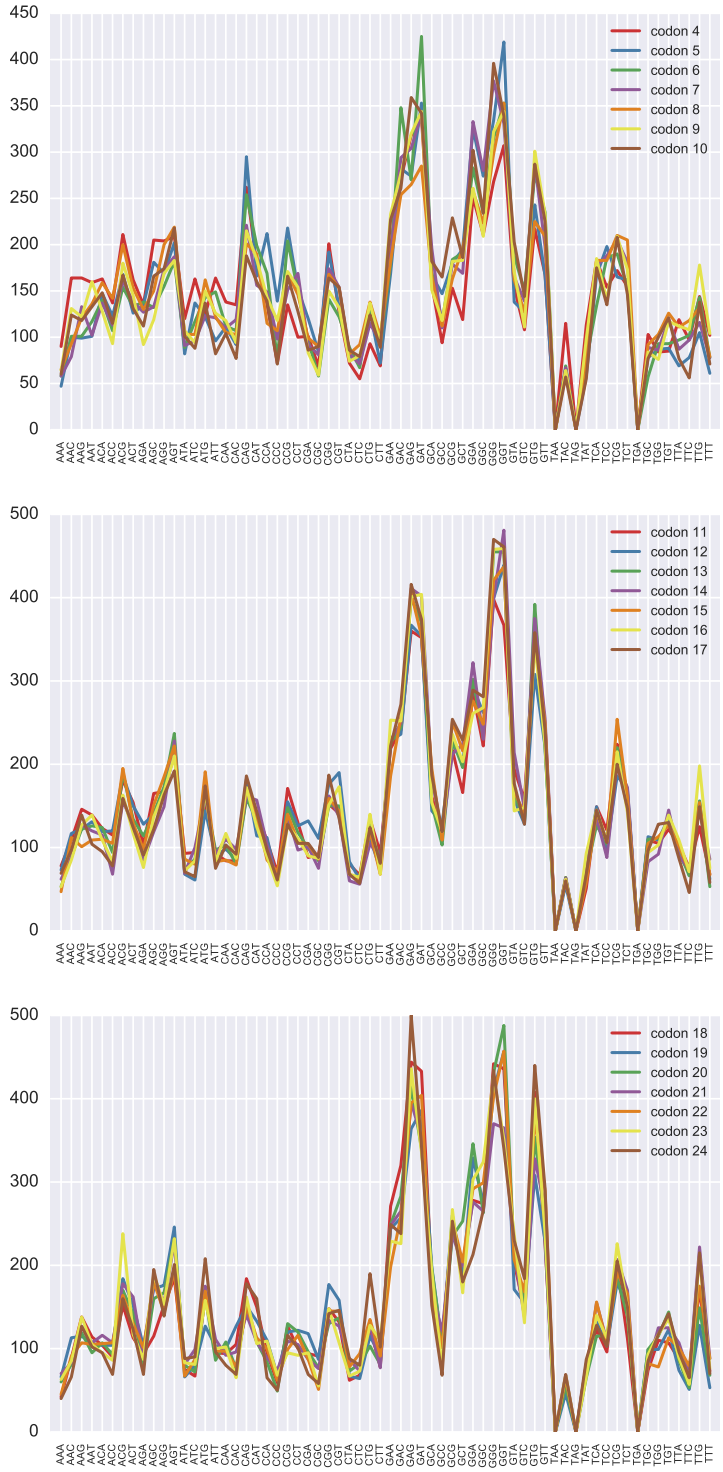


Figure 7.5: **Codon bias along the construct:** We expect some degree of codon bias in the data due to varying abundances of amino acid specific tRNA molecules in the mammalian cell lysate. The over-represented codons correspond to the amino acids aspartic acid, glutamic acid, glycine, and valine.

footprinting [121], and therefore may pose a steric hindrance to translation. To further explore this hypothesis, we plot the positional base-pairing probability and the positional entropy over the length of the randomized genome in Figure 7.7. The decoy set is plotted in dashed lines and displays no interesting behavior. The RT set plotted in solid lines, on the other hand, shows marked divergence of entropy and base-pairing probability between approximately nucleotide positions 8-20. This is range of positions we expect stable stem formation to be taking place, and perhaps driving RT.

In preparation for training a machine learning model to discern the RT and decoy sets of genomes, we must define a relevant feature space. Our exploratory analysis thus far indicates the importance of the nucleotides in the proximal 3' positions (likely within the ribosome tunnel), in addition to stem formation in the cytosolic portion of the RNA further downstream. We settle on a hybrid sequence/structural space of features, in which the first 6 positions are encoded in a position-specific-nucleotide manner and the characteristics of a potential stem are summarized with 6 numbers. The stem features include 1) starting position, 2) length, 3) %GC base pairs, 4) average base pair probability, 5) variance of base pair probability, and 6) number of bulge nucleotides. These features are depicted in Figure 7.8.

Whether each genome is contained in the decoy set or the RT set defines a binary labeling of the data and permits the training of a binary classifier. Analogous to the development of Pegasus in Chaper 2, we train an ensemble of gradient boosted decision trees to discriminate the label from the sequence and structural features. The model achieves an AUROC of 0.99 on held out data under 10-fold cross validation after 100 boosting rounds using depth-3 decision trees as base learners. The most informative feature is 1_C. Within the top 10 features are all six structural features, in addition to nucleotide identities 1,2,3,6 of the classic TMV motif. The definition of feature importance scores can be found in Chapter 2, along with an overview of gradient tree boosting. The training performance is depicted in Figure 7.9 and the decision tree from the first boosting round is depicted in Figure 7.10.

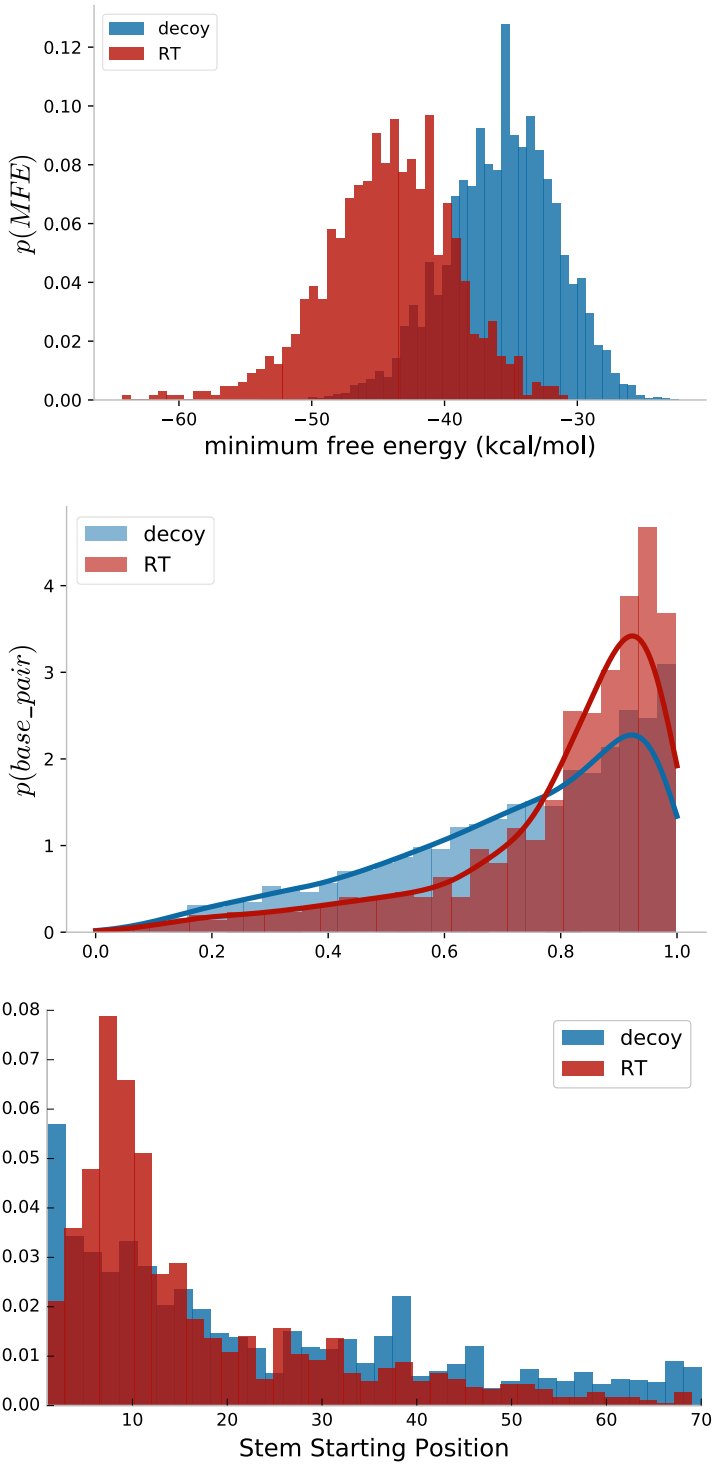


Figure 7.6: **Structural predictions for RT and decoy sequences:** Three key quantities resolve the positive RT examples from the negative decoy examples. In order of plotting, these are: minimum free energy (MFE) of predicted stem, average base pairing probability along predicted stem, and the starting nucleotide (position) for the predicted stem.

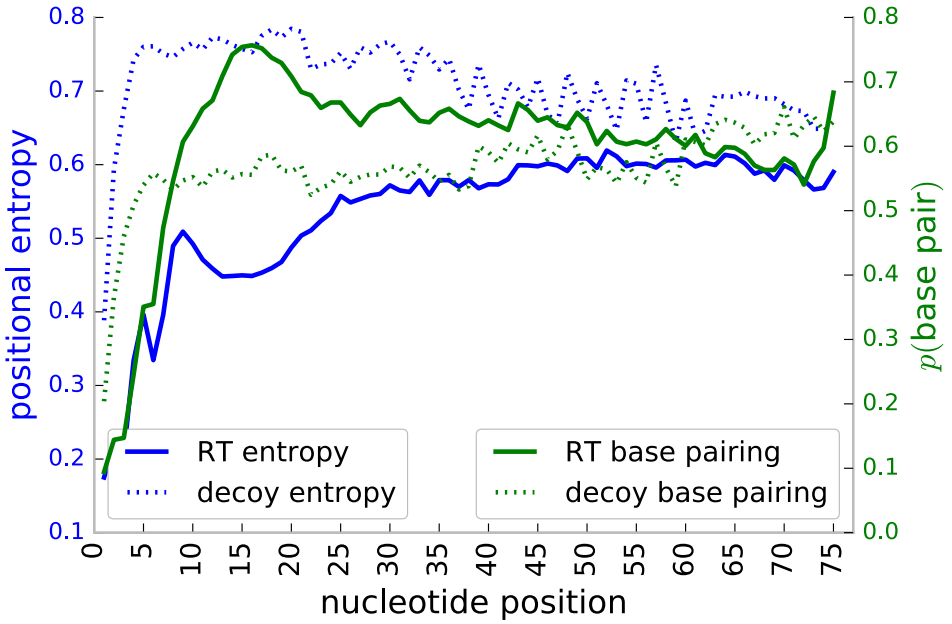


Figure 7.7: **Positional entropy and base pairing at mRNA entry tunnel:** To further explore the signal we observe in base pairing and stem formation (Figure 7.6), we compare the positional entropy and base pairing probability along the RNA sequence. The decoy set shows no interesting behavior for either quantity, while the RT curves strongly diverge around the 8th nucleotide. A high degree of stem formation is initiated in this region, corresponding to a rise in base pairing probability and a decrease in entropy at that locus. This region of the RNA may be close to the entry tunnel in the eukaryotic ribosome, given the typical side of mRNA footprints [121].

The shading of the nodes in the decision tree denotes the concentration of positive (orange) vs. negative (white) labels.

7.4 Predicting Readthrough in Human Transcripts from 3'UTR Context

A benefit of our feature space construction is that it generalizes to RNA sequences outside of our randomized library. The most obvious data on which to apply the trained classifier are 3' UTR sequences from human (and model organism) transcriptomes. We collect all annotated human 3' UTR coordinates from the UCSC table browser [105] and extract their

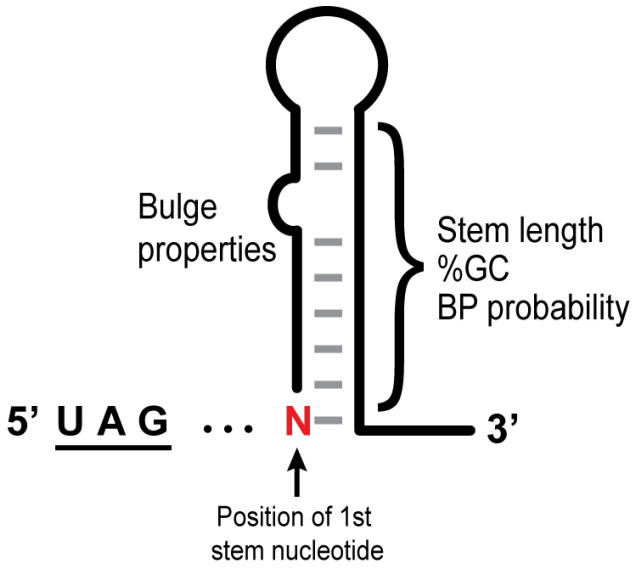


Figure 7.8: **Structural component of RNA feature space:** Based on our exploratory analysis of stable stem formation proximal to the ribosome tunnel, we define a feature encoding that captures potential structural determinants of RT efficiency.

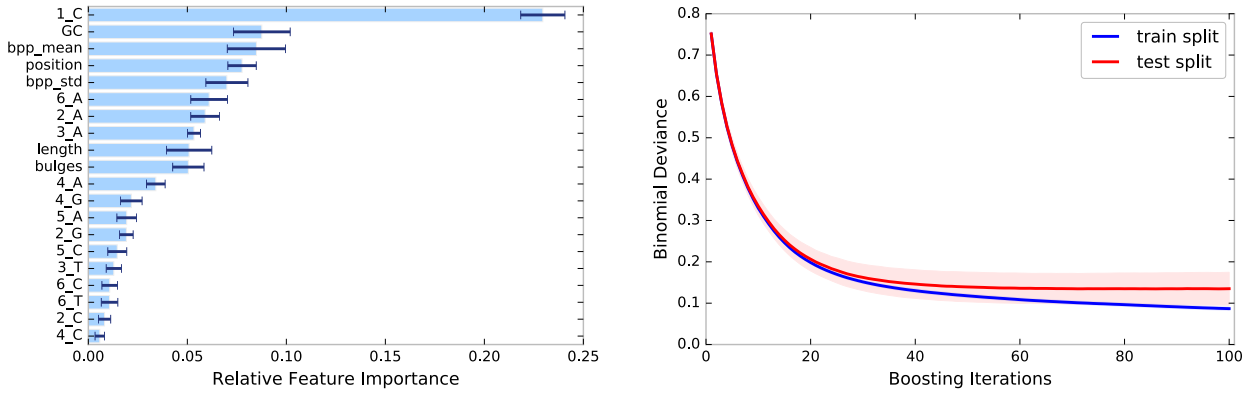


Figure 7.9: **Boosting results for binary RT classification:** (Left) Feature importances, as defined in Chapter 2, are plotted in descending order. We recover an intuitive ranking that includes key nucleotides of the known TMV sequence motif, as well as structural features. (Right) The loss function is plotted over the boosting rounds for both the training split and the test split (10-fold cross validation).

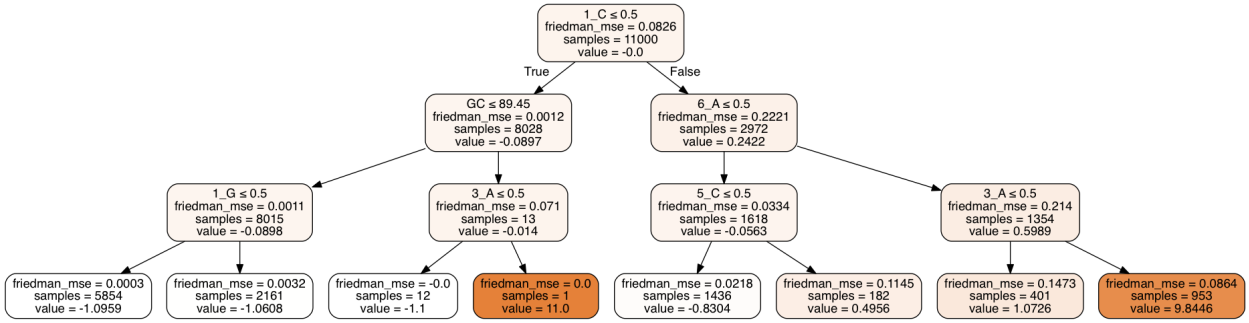


Figure 7.10: **Decision tree from first boosting round:** The base learners in the boosting algorithm are depth-3 decision trees and the one visualized here comes from the first boosting round. As expected from the feature importance scores, the question 1_C is at the root of the decision tree, denoting the binary-valued statement "the first nucleotide is C." We note that most of the structural features are not invoked in the first boosting round.

corresponding nucleotide sequence from the hg19 genome assembly using bedtools [168]. UTR sequences are trimmed to the 135 nucleotides 3' to the stop codon, and hairpin structure prediction is performed using RNAfold [129] with default parameters. Position-specific nucleotide information and predicted structures are combined into the same feature vector representation described for the model training above. The trained classifier is applied to the featurized UTRs and the distribution of predicted RT probabilities is visualized in Figure 7.1. As expected, the vast majority of transcripts are predicted to have no RT, while a narrow tail of higher scoring UTRs are interesting from the perspective of uncovering novel regulation of translation in humans. Using the dual fluorescent reporter described above, we have validated certain predictions (Table 7.1) from the trained classifier as well as positive and negative controls from the literature. Validations are performed in transfected human (HEK293) cells, with a lower limit of RT detection of approximately 0.5%.

A recent study [130] relied on PhyloCSF [124] to detect regions of sequence conservation in the immediate vicinity of a UGA stop codon, and subsequently used a luciferase assay in HEK293 cells to validate predicted genes. The main results of their work were a recurrent RT-promoting CTAG motif immediately 3' to the stop codon and the confirmation of *in vivo* RT in the transcripts of MAPK10, AQP4, OPRK1, and OPRL1 (guiding our choice of controls). They measured no RT experimentally in one of their computationally predicted genes, ACP2,

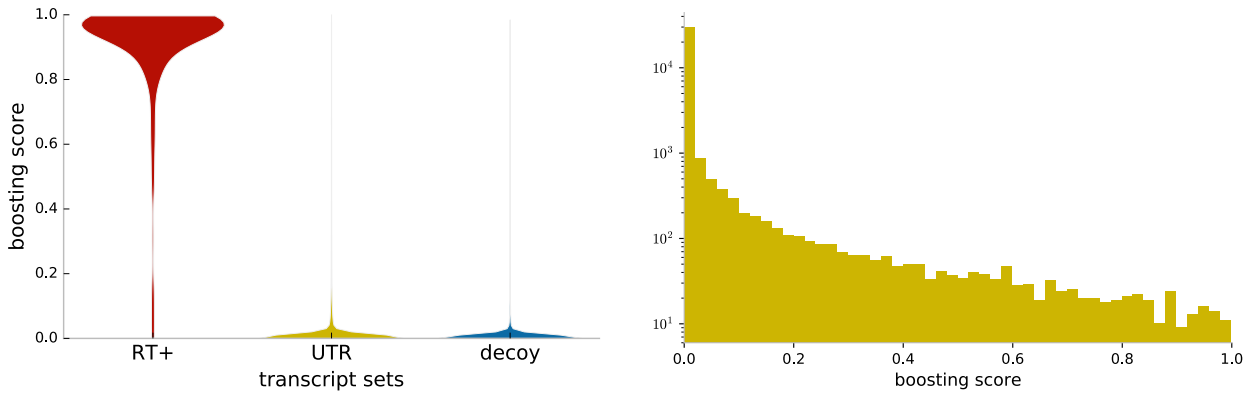


Figure 7.11: **Application of trained classifier:** (Left) The 0.99 AUROC of the classifier on held out data during training does not indicate the size of the classification margin. We demonstrate the high degree of resolution in the labels for the known positive (RT+) and negative (decoy) training examples, while also plotting the predictions for the human 3' UTR data. (Right) We plot the distribution of scores associated to all annotated human 3' UTR sequences on a log scale to demonstrate the relatively small set of UTR contexts that are scored highly in the eyes of the classifier.

which was thus therefore intended to be our negative control. Our computational method also scores ACP2 among the highest of all human transcripts (Table 7.1), but surprisingly we experimentally observe a 2.43% RT efficiency – well above our detection limit. The key difference with the previous study's measurement was their choice of validation construct which truncated a critical 3' stem, while our construct extends far enough into the UTR to capture the entire structural motif. ACP2 should therefore have been a true positive rather than a false positive of the PhyloCSF analysis, and our methodology illuminates the precise region of the UTR that mediates the RT.

Strikingly, we also detect greater than 5% RT efficiency in the transcripts of the vitamin D receptor (VDR). The alternative translational isoform of VDR produced by RT contains an additional 66 amino acids, 17 of which are prolines (26%). This predicted isoform was directly visualized on western blotting of N-terminally FLAG-tagged VDR constructs. This extended translational isoform is not described in the literature at present, thus raising questions about its stability *in vivo*. In terms of a potential biological role, C-terminal proline-rich tails have been implicated in nuclear localization [161], which could imply a

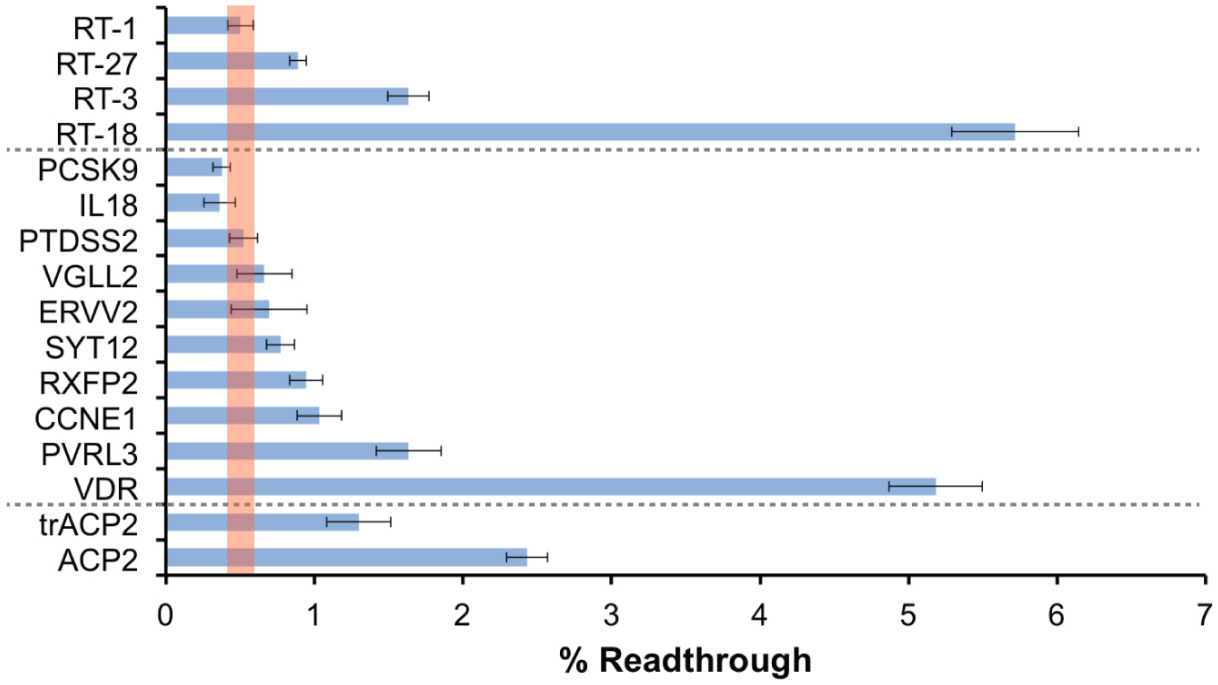


Figure 7.12: **Validation of RT efficiency in human cell culture:** HEK293 cells were transfected with constructs for selected mRNAs along with their 3' UTR. RT-1 represents the negative control, and its confidence interval for RT efficiency constitutes the lower detection limit for evaluating RT efficiencies of the other constructs. Our ACP2 construct is distinguised from trACP2 which is the shorter, truncated construct assayed previously in [130] and found to have no RT. Detailed, feature-level information for the human gene constructs is contained in Table 7.1.

testable hypothesis that the extended VDR would predominantly be observed in the nucleus as opposed to the cytoplasm.

7.5 Conclusions

Stop codon readthrough (RT) is increasingly being appreciated as a physiological regulatory mechanism of translation. Both primary sequence and secondary structure contribute to promoting RT, although the exact biochemical basis of the phenomenon is not known. We employ a large, randomized population of RNAs and an *in vitro* selection to drive the clonal evolution of the populations toward high efficiency RT. A machine learning model is trained to discriminate high efficiency RT from decoy RNAs, and our feature space contruction is

Table 7.1: **Predicted readthrough in human 3' UTRs:** Transcripts from all human genes are assessed for the efficiency of their 3' UTR context in inducing RT. We tabulate a small subset of RNAs that are nominated for experimental validation via the reporter assay described above in human HEK293 cells. Informative features of transcripts are tabulated, including the initial 10 nucleotide sequence context beyond the stop codon, as well as parameters of predicted stem structure such as starting position, length, GC content, and average predicted base-pairing probability. MAPK10, AQP4, and OPRL1 are included as positive controls from an earlier study [130], however ACP2 is specifically reported in [130] as having no RT. A novel discovery in this panel appears to be VDR with its greater than 5% RT.

Score	Gene	Stop Codon	First 10nt	$Start_{stem}$	Len_{stem}	GC_{stem}	$\overline{BPP}_{stem}$	RT_{HEK293}
0.989961	RXFP2	TAG	CAATCATTTT	10	11	54.5%	0.795	0.95%
0.973489	ACP2	TGA	CAACCACTCA	7	16	68.8%	0.985	2.43%
0.962823	SYT12	TAG	CAACCAGGGC	14	17	58.8%	0.992	0.77%
0.958299	PVRL3	TAG	CAACCAGGGC	7	21	33.3%	0.946	1.64%
0.929773	VGLL2	TGA	TCTGCTGACC	14	16	62.5%	0.989	0.66%
0.880812	PCSK9	TGA	CAGCCCCATC	11	23	69.6%	0.723	0.38%
0.866885	CCNE1	TGA	CCACCCCATC	21	22	59.1%	0.889	1.03%
0.740971	IL18	TAG	CTATTAAAAT	12	11	72.7%	0.906	0.36%
0.629644	VDR	TGA	CTAGGACAGC	25	11	81.8%	0.984	5.18%
0.557524	PTDSS2	TGA	CCTGGGCCGT	2	21	76.2%	0.99	0.52%
0.226695	MAPK10	TGA	CTAGCCGCCT	9	9	77.7%	0.509	5.77%
0.115449	AQP4	TGA	CTAGAAGATC	13	30	43.3%	0.556	4.77%
0.084206	OPRL1	TGA	CTAGGCGTGG	20	10	60.0%	0.898	20.16%
0.007830	ERVV2	TGA	GACAGAGCAA	8	19	57.9%	0.985	0.70%

broadly applicable to naturally occurring 3' UTRs. Two biochemical principles that emerge are that stem formation near the entrance of the ribosome tunnel seems to correlate with RT, and that the TMV motif of CARYYA in the first 6 positions is the dominant RT-promoting sequence element. Predicted RT efficiency has been calculated for all human 3' UTRs and much of the future work on this project will involve experimental validation of highly scored transcripts. The novel observation of an extended translational isoform of the calcitriol receptor, VDR, also offers many opportunities for downstream characterization. If 5% of endogenous calcitriol receptors are in fact RT isoforms, and if the 66 amino acid C-terminal extension confers some differential function, then our methodology would have uncovered a new wrinkle in endocrine physiology.

Chapter 8

Impressions

Large scale sequencing is becoming a foundational technology in the pipelines of both basic and applied life science projects. This work highlights its utility in uncovering evolutionary associations between mutation and disease, as well as its role in evolving nucleic acids for engineering purporses. In Part I we discussed the detection of oncogenic gene fusions and tumor-associated viruses in the context of whole transcriptome sequencing. In Part II we explored approaches for visualization and statistical summarization of clonal evolution in the context of tumor and viral genome surveillance. In Part III we illustrated an engineering approach to clonal evolution, wherein populations of nucleic acids were iteratively evolved for their ability to reprogram protein synthesis.

The study and exploitation of genetic evolution has become more quantitative in recent decades, with the simultaneous revolutions in molecular biology and digital computing of the latter 20th century driving the trend. The most obvious consequence of biology's transformation to a quantitative discipline is the rising demand for data analysis infrastructure, particularly in the area of high-throughput sequencing. Part I showcased two high-throughput data analysis approaches that typify current needs in translational genomics. Today's technologies for deep molecular profiling of genetic material are providing data faster than ever before, but there is more to quantitative biology than just data. As this work made clear, vast

nucleotide-resolution data sets can often point to surprisingly simple mathematical objects, and the degrees of freedom of these objects define intuitive spaces for biological hypothesis formation. In Part II the objects were somewhat abstract, phylogenetic trees and geometric spaces thereof, while in Parts I and III the objects were more concrete, namely DNA fusions and RNA pseudoknots and hairpins.

A crucial step in the development of predictive models in biology is identifying relevant mathematical structures and relevant degrees of freedom in general. I am confident that computational approaches in this spirit will continue to find fertile ground in life science's open problems.

Bibliography

1. F. Abate*, S. Zairis*, E. Ficarra, A. Acquaviva, C. H. Wiggins, V. Frattini, A. Lasorella, A. Iavarone, G. Inghirami, R. Rabadan, *BMC systems biology* **8**, 97 (2014).

2. F. Abate, A. Acquaviva, G. Paciello, C. Foti, E. Ficarra, A. Ferrarini, M. Delledonne, I. Iacobucci, S. Soverini, G. Martinelli, *et al.*, *Bioinformatics* **28**, 2114–2121 (2012).

3. F. Abate, M. Todaro, J.-A. van der Krogt, M. Boi, I. Landra, R. Machiorlatti, F. Tabbò, K. Messana, C. Abele, A. Barreca, *et al.*, *Leukemia* **29**, 1390–1401 (2015).

4. F. Abate, M. R. Ambrosio, L. Mundo, M. A. Laginestra, F. Fuligni, M. Rossi, S. Zairis, S. Gazaneo, G. De Falco, S. Lazzi, *et al.*, *PLoS Pathog* **11**, e1005158 (2015).

5. F. Abate, A. C. da Silva-Almeida, S. Zairis, J. Robles-Valero, L. Couronne, H. Khiabanian, S. A. Quinn, M.-Y. Kim, M. A. Laginestra, C. Kim, *et al.*, *Proceedings of the National Academy of Sciences*, 201608839 (2017).

6. J. Adams, A. Harris, C. Pinkert, L. Corcoran, W. Alexander, S. Cory, R. D. Palmiter, R. Brinster, *Nature* **318**, 533–538 (1985).

7. A. Alexandrov, *Schr. Forschungsinst. Math. Berlin* **1**, 33–84 (1957).

8. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *Journal of molecular biology* **215**, 403–410 (1990).

9. A. V. Anzalone, A. J. Lin, S. Zairis, R. Rabadan, V. W. Cornish, *Nature methods* **13**, 453–458 (2016).

10. A. Arvey, A. I. Ojesina, C. S. Pedamallu, G. Ballon, J. Jung, F. Duke, L. Leoncini, G. De Falco, E. Bressman, W. Tam, *et al.*, *Blood* **125**, e14–e22 (2015).

11. K. Atteson, presented at the Mathematical Hierarchies and Biology, Proceedings of a DIMACS Workshop, November 13-15, 1996, pp. 133–148.

12. S. Ausländer, P. Stücheli, C. Rehm, D. Ausländer, J. S. Hartig, M. Fussenegger, *Nature methods* **11**, 1154–1160 (2014).

13. D. Barden, L. Huiling, M. Owen, *Electronic journal of probability* **18**, 1–25 (2013).

14. M. Barreira, S. Fabbiano, J. R. Couceiro, E. Torreira, J. L. Martínez-Torrecuadrada, G. Montoya, O. Llorca, X. R. Bustelo, presented at the.

15. N. Beerenwinkel, R. F. Schwarz, M. Gerstung, F. Markowetz, *Systematic biology* **64**, e1–e25 (2015).

16. A. I. Bell, K. Groves, G. L. Kelly, D. Croom-Carter, E. Hui, A. T. Chan, A. B. Rickinson, *Journal of General Virology* **87**, 2885–2890 (2006).

17. C. Berens, A. Thain, R. Schroeder, *Bioorganic & medicinal chemistry* **9**, 2549–2556 (2001).

18. M. Bernstein, V. D. Silva, J. C. Langford, J. B. Tenenbaum, *Graph Approximations to Geodesics on Embedded Manifolds*, Stanford University technical report, 2000.

19. L. J. Billera, S. P. Holmes, K. Vogtmann, *Advances in Applied Mathematics* **27**, 733–767 (2001).

20. L. Breiman, J. Friedman, C. J. Stone, R. A. Olshen, *Classification and regression trees* (CRC press, 1984).

21. C. W. Brennan, R. G. Verhaak, A. McKenna, B. Campos, H. Noushmehr, S. R. Salama, S. Zheng, D. Chakravarty, J. Z. Sanborn, S. H. Berman, *et al.*, *Cell* **155**, 462–477 (2013).

22. M. R. Bridson, in *Group theory from a geometric viewpoint* (World Scientific, 1991), pp. 373–464.

23. M. R. Bridson, A. Haefliger, *Metric spaces of non-positive curvature* (Springer, Berlin, New York, 1999).

24. I. Brierley, *Journal of General Virology* **76**, 1885–1892 (1995).

25. P. Buneman, *Journal of Combinatorial Theory, Series B* **17**, 48–50 (1974).

26. D. Burago, Y. Burago, S. Ivanov, *A course in metric geometry* (American Mathematical Society, 2001).

27. X. R. Bustelo, *Small GTPases* **5**, e973757 (2014).

28. P. J. Campbell, S. Yachida, L. J. Mudie, P. J. Stephens, E. D. Pleasance, L. A. Stebbings, L. A. Morsberger, C. Latimer, S. McLaren, M.-L. Lin, *et al.*, *Nature* **467**, 1109–1113 (2010).

29. Y. Cao, E. M. Janssen, A. W. Duncan, A. Altman, D. D. Billadeau, R. T. Abraham, *The EMBO journal* **21**, 4809–4819 (2002).

30. G. Cario, U. z. Stadt, A. Reiter, K. Welte, K.-W. Sykora, *British journal of haematology* **110**, 537–546 (2000).

31. M. Carrara, M. Beccuti, F. Lazzarato, F. Cavallo, F. Cordero, S. Donatelli, R. A. Calogero, *BioMed research international* **2013** (2013).

32. Y. Chang, E. Cesarman, M. S. Pessin, F. Lee, *et al.*, *Science* **266**, 1865 (1994).

33. Y. Y. Chen, M. C. Jensen, C. D. Smolke, *Proceedings of the National Academy of Sciences* **107**, 8531–8536 (2010).

34. R. Chiarle, C. Voena, C. Ambrogio, R. Piva, G. Inghirami, *Nature Reviews Cancer* **8**, 11–23 (2008).

35. G. Cho, A. D. Keefe, R. Liu, D. S. Wilson, J. W. Szostak, *Journal of molecular biology* **297**, 309–319 (2000).

36. Q.-L. Choo, G. Kuo, A. J. Weiner, L. R. Overby, D. W. Bradley, M. Houghton, *Science* **244**, 359 (1989).

37. J. M. Churko, G. L. Mantalas, M. P. Snyder, J. C. Wu, *Circulation research* **112**, 1613–1623 (2013).

38. J. Clark, S. Merson, S. Jhavar, P. Flohr, S. Edwards, C. Foster, R. Eeles, F. L. Martin, D. Phillips, M. Crundwell, *et al.*, *Oncogene* **26**, 2667–2673 (2007).

39. U. Consortium *et al.*, *Nucleic acids research*, gkr981 (2011).

40. P. S. Costello, A. E. Walters, P. J. Mee, M. Turner, L. F. Reynolds, A. Prisco, N. Sarner, R. Zamoyska, V. L. Tybulewicz, *Proceedings of the National Academy of Sciences* **96**, 3035–3040 (1999).

41. R. Crescenzo, F. Abate, E. Lasorsa, M. Gaudiano, N. Chiesa, F. Di Giacomo, E. Spaccarotella, L. Barbarossa, E. Ercole, M. Todaro, *et al.*, *Cancer cell* **27**, 516–532 (2015).

42. P. Crespo, K. E. Schuebel, A. A. Ostrom, J. S. Gutkind, X. R. Bustelo, *Nature* **385**, 169 (1997).

43. G. E. Crooks, G. Hon, J.-M. Chandonia, S. E. Brenner, *Genome research* **14**, 1188–1190 (2004).

44. R. Dalla-Favera, M. Bregni, J. Erikson, D. Patterson, R. C. Gallo, C. M. Croce, *Proceedings of the National Academy of Sciences* **79**, 7824–7827 (1982).

45. S. S. Dave, K. Fu, G. W. Wright, L. T. Lam, P. Kluin, E.-J. Boerma, T. C. Greiner, D. D. Weisenburger, A. Rosenwald, G. Ott, *et al.*, *New England Journal of Medicine* **354**, 2431–2442 (2006).

46. P. de Felipe, L. E. Hughes, M. D. Ryan, J. D. Brown, *Journal of Biological Chemistry* **278**, 11441–11448 (2003).

47. L. de Leval, P. Gaulard, *Blood* **123**, 2909–2910 (2014).

48. M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. Del Angel, M. A. Rivas, M. Hanna, *et al.*, *Nature genetics* **43**, 491–498 (2011).

49. S. L. Devadoss, D. Huang, D. Spadacene, *SIAM Journal on Discrete Mathematics* **28**, 1508–1514 (2014).

50. R. Diaz-Uriarte, *bioRxiv*, 069500 (2016).

51. E. W. Dijkstra, *Numerische mathematik* **1**, 269–271 (1959).

52. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, *Bioinformatics* **29**, 15–21 (2013).

53. X. Du, L. Dong, Y. Lan, Y. Peng, A. Wu, Y. Zhang, W. Huang, D. Wang, M. Wang, Y. Guo, *et al.*, *Nature communications* **3**, 709 (2012).

54. R. C. Edgar, *Nucleic acids research* **32**, 1792–1797 (2004).

55. H. Edgren, A. Murumagi, S. Kangaspeska, D. Nicorici, V. Hongisto, K. Kleivi, I. H. Rye, S. Nyberg, M. Wolf, A.-L. Borresen-Dale, *et al.*, *Genome biology* **12**, R6 (2011).

56. P. Eirew, A. Steif, J. Khattra, G. Ha, D. Yap, H. Farahani, K. Gelmon, S. Chia, C. Mar, A. Wan, *et al.*, *Nature* (2014).

57. A. D. Ellington, J. W. Szostak, *nature* **346**, 818 (1990).

58. M. A. Epstein, B. G. Achong, Y. M. Barr, *The Lancet* **283**, 702–703 (1964).

59. S. M. Eswarappa, A. A. Potdar, W. J. Koch, Y. Fan, K. Vasu, D. Lindner, B. Willard, L. M. Graham, P. E. DiCorleto, P. L. Fox, *Cell* **157**, 1605–1618 (2014).

60. A. Fedorov, S. Saxonov, W. Gilbert, *Nucleic acids research* **30**, 1192–1197 (2002).

61. A. L. Feldman, G. Vasmatzis, Y. W. Asmann, J. Davila, S. Middha, B. W. Eckloff, S. H. Johnson, J. C. Porcher, S. M. Ansell, A. Caride, *Genes, Chromosomes and Cancer* **52**, 1097–1102 (2013).

62. J. Felsenstein, *cladistics* **5**, 6 (1989).

63. J. Felsenstein, *Inferring phylogenies* (Sinauer associates Sunderland, 2004), vol. 2.

64. H. Feng, M. Shuda, Y. Chang, P. S. Moore, *Science* **319**, 1096–1100 (2008).

65. A. Feragen, M. Owen, J. Petersen, M. M. Wille, L. H. Thomsen, A. Dirksen, M. de Bruijne, presented at the Information Processing in Medical Imaging, pp. 74–85.

66. A. E. Firth, I. Brierley, *Journal of General Virology* **93**, 1385–1409 (2012).

67. K. Fischer, Y. Kong, H. Nishina, K. Tedford, L. Marengere, I. Kozieradzki, T. Sasaki, M. Starr, G. Chan, S. Gardener, *et al.*, *Current Biology* **8**, 554–S3 (1998).

68. R. A. Fisher, *The genetical theory of natural selection* (Oxford University Press, 1930).

69. P. Flicek, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, *et al.*, *Nucleic acids research*, gkr991 (2011).

70. C. for Disease Control, Prevention, *MMWR. Morbidity and mortality weekly report* **53**, 8 (2004).

71. V. Frattini, V. Trifonov, J. M. Chan, A. Castano, M. Lia, F. Abate, S. T. Keir, A. X. Ji, P. Zoppoli, F. Niola, *et al.*, *Nature genetics* **45**, 1141–1149 (2013).

72. J. Freitag, J. Ast, M. Bölker, *Nature* **485**, 522–525 (2012).

73. J. H. Friedman, *Annals of statistics*, 1189–1232 (2001).

74. J. Friedman, T. Hastie, R. Tibshirani, *The elements of statistical learning* (Springer series in statistics Springer, Berlin, 2001), vol. 1.

75. J. Friedman, T. Hastie, R. Tibshirani, *et al.*, *The annals of statistics* **28**, 337–407 (2000).

76. K. E. Galloway, E. Franco, C. D. Smolke, *Science* **341**, 1235005 (2013).

77. GATK, *Calling Variants in RNAseq* (2014; https://software.broadinstitute.org/gatk/guide/article?id=3891).

78. R. F. Gesteland, J. F. Atkins, *Annual review of biochemistry* **65**, 741–768 (1996).

79. E. Giné, J. Zinn, *The Annals of Probability*, 929–989 (1984).

80. E. Giné, J. Zinn, in *Probability and Banach spaces* (Springer Berlin Heidelberg, 1986), pp. 50–113.

81. E. Giné, J. Zinn, *The Annals of Probability*, 851–869 (1990).

82. L. Giulino-Roth, K. Wang, T. Y. MacDonald, S. Mathew, Y. Tam, M. T. Cronin, G. Palmer, N. Lucena-Silva, F. Pedrosa, M. Pedrosa, *et al.*, *Blood* **120**, 5181–5184 (2012).

83. M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, *et al.*, *Nature biotechnology* **29**, 644 (2011).

84. M. Gromov, *Metric structures for Riemannian and non-Riemannian spaces* (Birkhauser, 1981).

85. D. Grün, A. Lyubimova, L. Kester, K. Wiebrands, O. Basak, N. Sasaki, H. Clevers, A. van Oudenaarden, *Nature* **525**, 251–255 (2015).

86. V. Gupta, D. J. Earl, M. W. Deem, *Vaccine* **24**, 3881–3888 (2006).

87. M. I. Gutierrez, K. Bhatia, F. Barriga, B. Diez, F. S. Muriel, M. L. de Andreas, S. Epelman, C. Risueno, I. T. Magrath, *Blood* **79**, 3261–3266 (1992).

88. E. C. Holmes, E. Ghedin, N. Miller, J. Taylor, Y. Bao, K. StGeorge, B. T. Grenfell, S. L. Salzberg, C. M. Fraser, D. J. Lipman, *et al.*, *PLoS biology* **3**, 1579 (2005).

89. S. Holmes, *Statistical Science*, 241–255 (2003).

90. S. Holmes, *Mathematics of evolution and phylogeny. Oxford University Press, Oxford, UK*, 91–120 (2005).

91. L. Holsinger, I. Graef, W. Swat, T. Chi, D. Bautista, L. Davidson, R. Lewis, F. Alt, G. Crabtree, *Current Biology* **8**, 563–573 (1998).

92. B. Houck-Loomis, M. A. Durney, C. Salguero, N. Shankar, J. M. Nagle, S. P. Goff, V. M. D'Souza, *Nature* **480**, 561–564 (2011).

93. H.-T. Hsu, Y.-H. Lin, K.-Y. Chang, *Nucleic acids research*, gku1233 (2014).

94. W. Huang, L. Li, J. R. Myers, G. T. Marth, *Bioinformatics* **28**, 593–594 (2012).

95. F. J. Isaacs, D. J. Dwyer, J. J. Collins, *Nature biotechnology* **24**, 545–554 (2006).

96. I. P. Ivanov, R. F. Gesteland, J. F. Atkins, *Nucleic Acids Research* **28**, 3185–3196 (2000).

97. M. K. Iyer, A. M. Chinnaiyan, C. A. Maher, *Bioinformatics* **27**, 2903–2904 (2011).

98. T. Jacks, H. D. Madhani, F. R. Masiarz, H. E. Varmus, *Cell* **55**, 447–458 (1988).

99. S. Janssen, R. Giegerich, *Bioinformatics*, btu649 (2014).

100. R. D. Jenison, S. C. Gill, A. Pardi, B. Polisky, *et al.*, *SCIENCE-NEW YORK THEN WASHINGTON-*, 1425–1425 (1994).

101. Y. Jin, F. Mertens, C.-M. Kullendorff, *et al.*, *Neoplasia* **8**, 413–418 (2006).

102. B. E. Johnson, T. Mazor, C. Hong, M. Barnes, K. Aihara, C. Y. McLean, S. D. Fouse, S. Yamamoto, H. Ueda, K. Tatsuno, *et al.*, *Science* **343**, 189–193 (2014).

103. T. H. Jukes, C. R. Cantor, *et al.*, *Mammalian protein metabolism* **3**, 132 (1969).

104. I. Jungreis, M. F. Lin, R. Spokony, C. S. Chan, N. Negre, A. Victorsen, K. P. White, M. Kellis, *Genome research* **21**, 2096–2113 (2011).

105. D. Karolchik, A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler, W. J. Kent, *Nucleic acids research* **32**, D493–D496 (2004).

106. G. L. Kelly, H. M. Long, J. Stylianou, W. A. Thomas, A. Leese, A. I. Bell, G. W. Bornkamm, J. Mautner, A. B. Rickinson, M. Rowe, *PLoS Pathog* **5**, e1000341 (2009).

107. G. L. Kelly, A. E. Milner, G. S. Baldwin, A. I. Bell, A. B. Rickinson, *Proceedings of the National Academy of Sciences* **103**, 14935–14940 (2006).

108. G. L. Kelly, J. Stylianou, J. Rasaiyaah, W. Wei, W. Thomas, D. Croom-Carter, C. Kohler, R. Spang, C. Woodman, P. Kellam, *et al.*, *Journal of virology* **87**, 2882–2894 (2013).

109. G. Kelly, A. Bell, A. Rickinson, *Nature medicine* **8**, 1098–1104 (2002).

110. P. Kim, S. Yoon, N. Kim, S. Lee, M. Ko, H. Lee, H. Kang, J. Kim, S. Lee, *Nucleic acids research* **38**, D81–D85 (2010).

111. M. Kimura, *Journal of molecular evolution* **16**, 111–120 (1980).

112. M. Kimura, *The neutral theory of molecular evolution* (Cambridge University Press, 1983).

113. J. F. C. Kingman, *Stochastic processes and their applications* **13**, 235–248 (1982).

114. B. Klauser, J. Atanasov, L. K. Siewert, J. S. Hartig, *ACS synthetic biology* **4**, 516–525 (2014).

115. M. Koizumi, G. A. Soukup, J. N. Q. Kerr, R. R. Breaker, *Nature Structural & Molecular Biology* **6**, 1062–1071 (1999).

116. Y.-Y. Kong, K.-D. Fischer, M. F. Bachmann, S. Mariathasan, I. Kozieradzki, M. P. Nghiem, D. Bouchard, A. Bernstein, P. S. Ohashi, J. M. Penninger, *Journal of Experimental Medicine* **188**, 2099–2111 (1998).

117. D. C. Krakauer, J. P. Collins, D. Erwin, J. C. Flack, W. Fontana, M. D. Laubichler, S. J. Prohaska, G. B. West, P. F. Stadler, *Journal of theoretical biology* **276**, 269–276 (2011).

118. C. Küçük, B. Jiang, X. Hu, W. Zhang, J. K. Chan, W. Xiao, N. Lack, C. Alkan, J. C. Williams, K. N. Avery, *et al.*, *Nature communications* **6** (2015).

119. M. R. Kuhne, G. Ku, A. Weiss, *Journal of Biological Chemistry* **275**, 2185–2190 (2000).

120. D. A. Landau, S. L. Carter, P. Stojanov, A. McKenna, K. Stevenson, M. S. Lawrence, C. Sougnez, C. Stewart, A. Sivachenko, L. Wang, *et al.*, *Cell* **152**, 714–726 (2013).

121. L. F. Lareau, D. H. Hite, G. J. Hogan, P. O. Brown, *Elife* **3**, e01257 (2014).

122. A. Lear, M. Rowe, M. Kurilla, S. Lee, S. Henderson, E. Kieff, A. Rickinson, *Journal of virology* **66**, 7461–7468 (1992).

123. Y. Liao, G. K. Smyth, W. Shi, *Bioinformatics* **30**, 923–930 (2014).

124. M. F. Lin, I. Jungreis, M. Kellis, *Bioinformatics* **27**, i275–i282 (2011).

125. K. H. Link, L. Guo, T. D. Ames, L. Yen, R. C. Mulligan, R. R. Breaker, *Biological chemistry* **388**, 779–786 (2007).

126. C. C. Liu, P. G. Schultz, *Annual review of biochemistry* **79**, 413–444 (2010).

127. R. Liu, J. E. Barrick, J. W. Szostak, R. W. Roberts, *Methods in enzymology* **318**, 268–293 (2000).

128. Y. Liu, X. Cao, *Cellular & molecular immunology* **12**, 1 (2015).

129. R. Lorenz, S. H. Bernhart, C. H. Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, I. L. Hofacker, *Algorithms for Molecular Biology* **6**, 26 (2011).

130. G. Loughran, M.-Y. Chou, I. P. Ivanov, I. Jungreis, M. Kellis, A. M. Kiran, P. V. Baranov, J. F. Atkins, *Nucleic acids research* **42**, 8928–8938 (2014).

131. C. A. Maher, N. Palanisamy, J. C. Brenner, X. Cao, S. Kalyana-Sundaram, S. Luo, I. Khrebtukova, T. R. Barrette, C. Grasso, J. Yu, *et al.*, *Proceedings of the National Academy of Sciences* **106**, 12353–12358 (2009).

132. I. J. Majewski, L. Mittempergher, N. M. Davidson, A. Bosma, S. M. Willems, H. M. Horlings, I. de Rink, L. Greger, G. K. Hooijer, D. Peters, *et al.*, *The Journal of pathology* **230**, 270–276 (2013).

133. C. D. McFarland, K. S. Korolev, G. V. Kryukov, S. R. Sunyaev, L. A. Mirny, *Proceedings of the National Academy of Sciences* **110**, 2910–2915 (2013).

134. A. McPherson, F. Hormozdiari, A. Zayed, R. Giuliany, G. Ha, M. G. Sun, M. Griffith, A. H. Moussavi, J. Senz, N. Melnyk, *et al.*, *PLoS Comput Biol* **7**, e1001138 (2011).

135. E. Miller, M. Owen, J. S. Provan, "Polyhedral computational geometry for averaging metric phylogenetic trees", arXiv:1211.7046, 2012.

136. F. Mitelman, B. Johansson, F. Mertens, *et al.*, *Mitelman F, Johansson B, Mertens F, editors* (2012).

137. P. A. P. Moran, presented at the Mathematical Proceedings of the Cambridge Philosophical Society, vol. 54, pp. 60–71.

138. S. W. Morris, M. N. Kirstein, M. B. Valentine, K. G. Dittmer, D. N. Shapiro, D. L. Saltman, A. T. Look, *et al.*, *Science*, 1281–1281 (1994).

139. P. S. Mountford, A. G. Smith, *Trends in Genetics* **11**, 179–184 (1995).

140. G. Moussong, "Hyperbolic Coxeter groups", Ohio State University doctoral thesis, 1988.

141. I. Mrsi, *Wikimedia Commons, Maxam-Gilbert sequencing*, [Online; accessed 23-January-2018] (2013; https://commons.wikimedia.org/wiki/File:Maxam-Gilbert_sequencing_en.svg).

142. O. Namy, I. Hatin, J.-P. Rousset, *EMBO reports* **2**, 787–793 (2001).

143. S. Napthine, C. Yek, M. L. Powell, T. D. K. Brown, I. Brierley, *Rna* **18**, 241–252 (2012).

144. K. N. Naresh, M. Raphael, L. Ayers, N. Hurwitz, V. Calbi, E. Rogena, S. Sayed, O. Sherman, H. A. Ibrahim, S. Lazzi, *et al.*, *British journal of haematology* **154**, 696–703 (2011).

145. N. Navin, J. Kendall, J. Troge, P. Andrews, L. Rodgers, J. McIndoo, K. Cook, A. Stepansky, D. Levy, D. Esposito, *et al.*, *Nature* **472**, 90–94 (2011).

146. A. Neri, F. Barriga, D. M. Knowles, I. T. Magrath, R. Dalla-Favera, *Proceedings of the National Academy of Sciences* **85**, 2748–2752 (1988).

147. A. M. Newman, C. L. Liu, M. R. Green, A. J. Gentles, W. Feng, Y. Xu, C. D. Hoang, M. Diehn, A. A. Alizadeh, *Nature methods* **12**, 453–457 (2015).

148. G. Niedobitek, A. Agathanggelou, M. Rowe, E. Jones, D. Jones, P. Turyaguma, J. Oryema, D. Wright, L. Young, *Blood* **86**, 659–665 (1995).

149. P. Niyogi, S. Smale, S. Weinberger, *Discrete & Computational Geometry* **39**, 419–441 (2008).

150. F. J. Novo, I. O. de Mendíbil, J. L. Vizmanos, *BMC genomics* **8**, 33 (2007).

151. C. Nowell, *Annals of Hematology* **8**, 65–66 (1962).

152. P. C. Nowell, *Science* **194**, 23–28 (1976).

153. R. Noy, J. W. Pollard, *Immunity* **41**, 49–61 (2014).

154. T. M. W. Nye, *Annals of Statistics* **39**, 2716–2739 (2011).

155. M. D. Ogwang, K. Bhatia, R. J. Biggar, S. M. Mbulaiteye, *International journal of cancer* **123**, 2658–2663 (2008).

156. W. H. Organization, *Influenza*, http://www.who.int/biologicals/vaccines/influenza/en, 2016.

157. J. Oudejans, A. Van den Brule, N. Jiwa, P. De Bruin, G. Ossenkoppele, P. Van der Valk, J. Walboomers, C. Meijer, *Blood* **86**, 1893–1902 (1995).

158. M. Owen, J. S. Provan, *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* **8**, 2–13 (2011).

159. F. Ozsolak, P. M. Milos, *Nature reviews genetics* **12**, 87–98 (2011).

160. T. Palomero, L. Couronné, H. Khiabanian, M.-Y. Kim, A. Ambesi-Impiombato, A. Perez-Garcia, Z. Carpenter, F. Abate, M. Allegretta, J. E. Haydu, *et al.*, *Nature genetics* **46**, 166–170 (2014).

161. H. A. Pancio, N. Vander Heyden, L. Ratner, *Journal of virology* **74**, 6162–6167 (2000).

162. A. P. Patel, I. Tirosh, J. J. Trombetta, A. K. Shalek, S. M. Gillespie, H. Wakimoto, D. P. Cahill, B. V. Nahed, W. T. Curry, R. L. Martuza, *et al.*, *Science* **344**, 1396–1401 (2014).

163. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

164. H. R. Pelham (1978).

165. P. P. Piccaluga, G. De Falco, M. Kustagi, A. Gazzola, C. Agostinelli, C. Tripodo, E. Leucci, A. Onnis, A. Astolfi, M. R. Sapienza, *et al.*, *Blood* **117**, 3596–3608 (2011).

166. L. Pournin, *Advances in Mathematics* **259**, 13–42 (2014).

167. W. Qiao, G. Quon, E. Csaszar, M. Yu, Q. Morris, P. W. Zandstra, *PLoS computational biology* **8**, e1002838 (2012).

168. A. R. Quinlan, I. M. Hall, *Bioinformatics* **26**, 841–842 (2010).

169. R. Rabadan, A. J. Levine, M. Krasnitz, *Influenza and Other Respiratory Viruses* **2**, 9–22 (2008).

170. R. Rabadan, H. Robins, *Evolutionary bioinformatics online* **3**, 299 (2007).

171. G. Ramaswami, J. B. Li, *Nucleic acids research*, gkt996 (2013).

172. L. F. Reynolds, C. de Bettignies, T. Norton, A. Beeser, J. Chernoff, V. L. Tybulewicz, *Journal of Biological Chemistry* **279**, 18239–18246 (2004).

173. L. F. Reynolds, L. A. Smyth, T. Norton, N. Freshney, J. Downward, D. Kioussis, V. L. Tybulewicz, *The Journal of experimental medicine* **195**, 1103–1114 (2002).

174. R. W. Roberts, J. W. Szostak, *Proceedings of the National Academy of Sciences* **94**, 12297–12302 (1997).

175. D. L. Robertson, G. F. Joyce, *Nature* **344**, 467–468 (1990).

176. A. Robinson, S. Whitehouse, *Journal of Pure and Applied Algebra* **111**, 245–253 (1996).

177. D. F. Robinson, L. R. Foulds, *Mathematical biosciences* **53**, 131–147 (1981).

178. P. Rous, *The Journal of experimental medicine* **13**, 397 (1911).

179. C. A. Russell, T. C. Jones, I. G. Barr, N. J. Cox, R. J. Garten, V. Gregory, I. D. Gust, A. W. Hampson, A. J. Hay, A. C. Hurt, *et al.*, *Vaccine* **26**, D31–D34 (2008).

180. N. Saitou, M. Nei, *Molecular biology and evolution* **4**, 406–425 (1987).

181. S. Sander, D. P. Calado, L. Srinivasan, K. Köchert, B. Zhang, M. Rosolowski, S. J. Rodig, K. Holzmann, S. Stilgenbauer, R. Siebert, *et al.*, *Cancer cell* **22**, 167–179 (2012).

182. A. Saveliev, L. Vanes, O. Ksionda, J. Rapley, S. J. Smerdon, K. Rittinger, V. L. Tybulewicz, *Science signaling* **2**, ra83 (2009).

183. A. Sboner, L. Habegger, D. Pflueger, S. Terry, D. Z. Chen, J. S. Rozowsky, A. K. Tewari, N. Kitabayashi, B. J. Moss, M. S. Chee, *et al.*, *Genome biology* **11**, R104 (2010).

184. M. Schmidt, H. Lipson, *science* **324**, 81–85 (2009).

185. R. Schmitz, M. Ceribelli, S. Pittaluga, G. Wright, L. M. Staudt, *Cold Spring Harbor perspectives in medicine* **4**, a014282 (2014).

186. R. Schmitz, R. M. Young, M. Ceribelli, S. Jhavar, W. Xiao, M. Zhang, G. Wright, A. L. Shaffer, D. J. Hodson, E. Buras, *et al.*, *Nature* **490**, 116–120 (2012).

187. A. Schuh, J. Becq, S. Humphray, A. Alexa, A. Burns, R. Clifford, S. M. Feller, R. Grocock, S. Henderson, I. Khrebtukova, *et al.*, *Blood* **120**, 4191–4196 (2012).

188. J. Shendure, H. Ji, *Nature biotechnology* **26**, 1135–1145 (2008).

189. M. Shugay, I. O. de Mendíbil, J. L. Vizmanos, F. J. Novo, *Bioinformatics*, btt445 (2013).

190. D. Singh, J. M. Chan, P. Zoppoli, F. Niola, R. Sullivan, A. Castano, E. M. Liu, J. Reichel, P. Porrati, S. Pellegatta, *et al.*, *Science* **337**, 1231–1235 (2012).

191. J. M. Skuzeski, L. M. Nichols, R. F. Gesteland, J. F. Atkins, *Journal of molecular biology* **218**, 365–373 (1991).

192. D. D. Sleator, R. E. Tarjan, W. P. Thurston, *Journal of the American Mathematical Society* **1**, 647–681 (1988).

193. D. D. Sleator, R. E. Tarjan, W. P. Thurston, *SIAM Journal of Discrete Mathematics* **5**, 428–450 (1992).

194. A. Smith, D. Howell, R. Patmore, A. Jack, E. Roman, *British journal of cancer* **105**, 1684–1692 (2011).

195. M. Soda, Y. L. Choi, M. Enomoto, S. Takada, Y. Yamashita, S. Ishikawa, S.-i. Fujiwara, H. Watanabe, K. Kurashina, H. Hatanaka, *et al.*, *Nature* **448**, 561–566 (2007).

196. A. Solovyov, G. Palacios, T. Briese, W. I. Lipkin, R. Rabadan, *Euro surveillance: bulletin Europeen sur les maladies transmissibles= European communicable disease bulletin* **14** (2009).

197. C. Steidl, S. P. Shah, B. W. Woolcock, L. Rui, M. Kawahara, P. Farinha, N. A. Johnson, Y. Zhao, A. Telenius, S. B. Neriah, *et al.*, *Nature* **471**, 377–381 (2011).

198. M. R. Stratton, P. J. Campbell, P. A. Futreal, *Nature* **458**, 719–724 (2009).

199. K.-T. Sturm, *Heat Kernels and Analysis on Manifolds, Graphs, and Metric Spaces: Lecture Notes from a Quarter Program on Heat Kernels, Random Walks, and Analysis on Manifolds and Graphs: April 16-July 13, 2002, Emile Borel Centre of the Henri Poincaré Institute, Paris, France* **338**, 357 (2003).

200. J. Tang, R. R. Breaker, *Chemistry & biology* **4**, 453–459 (1997).

201. A. Tarakhovsky, M. Turner, S. Schaal, P. J. Mee, *et al.*, *Nature* **374**, 467 (1995).

202. D. A. Thorley-Lawson, J. B. Hawkins, S. I. Tracy, M. Shapiro, *Current opinion in virology* **3**, 227–232 (2013).

203. R. J. Tierney, C. D. Shannon-Lowe, L. Fitzsimmons, A. I. Bell, M. Rowe, *Virology* **474**, 117–130 (2015).

204. S. A. Tomlins, D. R. Rhodes, S. Perner, S. M. Dhanasekaran, R. Mehra, X.-W. Sun, S. Varambally, X. Cao, J. Tchinda, R. Kuefer, *et al.*, *science* **310**, 644–648 (2005).

205. B. Townshend, A. B. Kennedy, J. S. Xiang, C. D. Smolke, *Nature methods* **12**, 989–994 (2015).

206. C. Trapnell, L. Pachter, S. L. Salzberg, *Bioinformatics* **25**, 1105–1111 (2009).

207. C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. Van Baren, S. L. Salzberg, B. J. Wold, L. Pachter, *Nature biotechnology* **28**, 511–515 (2010).

208. V. Trifonov, H. Khiabanian, R. Rabadan, *New England Journal of Medicine* **361**, 115–119 (2009).

209. V. Trifonov, L. Pasqualucci, E. Tiacci, B. Falini, R. Rabadan, *BMC systems biology* **7**, S2 (2013).

210. C. Tuerk, L. Gold, *et al.*, *Science* **249**, 505–510 (1990).

211. M. Turner, P. J. Mee, A. E. Walters, M. E. Quinn, A. L. Mellor, R. Zamoyska, V. L. Tybulewicz, *Immunity* **7**, 451–460 (1997).

212. V. L. Tybulewicz, *Current opinion in immunology* **17**, 267–274 (2005).

213. C. Van den Bosch, *Advances in hematology* **2012** (2012).

214. J. Wang, H. Khiabanian, D. Rossi, G. Fabbri, V. Gattei, F. Forconi, L. Laurenti, R. Marasca, G. Del Poeta, R. Foà, *et al.*, *eLife* **3**, e02869 (2015).

215. J. Wang, E. Cazzato, E. Ladewig, V. Frattini, D. I. Rosenbloom, S. Zairis, F. Abate, Z. Liu, O. Elliott, Y.-J. Shin, *et al.*, *Nature genetics* **48**, 768–776 (2016).

216. Q. Wang, P. Jia, Z. Zhao, *PloS one* **8**, e64465 (2013).

217. X.-S. Wang, J. R. Prensner, G. Chen, Q. Cao, B. Han, S. M. Dhanasekaran, R. Ponnala, X. Cao, S. Varambally, D. G. Thomas, *et al.*, *Nature biotechnology* **27**, 1005–1011 (2009).

218. J. E. Weigand, M. Sanchez, E.-B. Gunnesch, S. Zeiher, R. Schroeder, B. Suess, *Rna* **14**, 89–97 (2008).

219. I. Williams, J. Richardson, A. Starkey, I. Stansfield, *Nucleic acids research* **32**, 6605–6616 (2004).

220. S. V. Williams, C. D. Hurst, M. A. Knowles, *Human molecular genetics* **22**, 795–803 (2013).

221. N. M. Wills, R. F. Gesteland, J. F. Atkins, *Proceedings of the National Academy of Sciences* **88**, 6991–6995 (1991).

222. M. N. Win, C. D. Smolke, *Science* **322**, 456–460 (2008).

223. C. R. Woese, O. Kandler, M. L. Wheelis, *Proceedings of the National Academy of Sciences* **87**, 4576–4579 (1990).

224. S. Wright, *Genetics* **16**, 97–159 (1931).

225. C.-C. Wu, K. Kannan, S. Lin, L. Yen, A. Milosavljevic, *Bioinformatics*, btt131 (2013).

226. J. Wu, S. Katzav, A. Weiss, *Molecular and Cellular Biology* **15**, 4337–4346 (1995).

227. J. Yeo, R. A. Goodman, N. T. Schirle, S. S. David, P. A. Beal, *Proceedings of the National Academy of Sciences* **107**, 20715–20719 (2010).

228. H. Y. Yoo, M. K. Sung, S. H. Lee, S. Kim, H. Lee, S. Park, S. C. Kim, B. Lee, K. Rho, J.-E. Lee, *et al.*, *Nature genetics* **46**, 371–375 (2014).

229. B. Yu, I. R. Martins, P. Li, G. K. Amarasinghe, J. Umetani, M. E. Fernandez-Zapico, D. D. Billadeau, M. Machius, D. R. Tomchick, M. K. Rosen, *Cell* **140**, 246–256 (2010).

230. C.-H. Yu, J. Luo, D. Iwata-Reuyl, R. C. Olsthoorn, *ACS chemical biology* **8**, 733–740 (2013).

231. J. N. Zadeh, C. D. Steenberg, J. S. Bois, B. R. Wolfe, M. B. Pierce, A. R. Khan, R. M. Dirks, N. A. Pierce, *Journal of computational chemistry* **32**, 170–173 (2011).

232. H. S. Zaher, R. Green, *Cell* **136**, 746–762 (2009).

233. S. Zairis, H. Khiabanian, A. J. Blumberg, R. Rabadan, presented at the International Conference on Brain Informatics and Health, pp. 528–539.

234. S. Zairis, H. Khiabanian, A. J. Blumberg, R. Rabadan, *arXiv preprint arXiv:1607.07503* (2016).

235. R. Zhang, F. W. Alt, L. Davidson, S. H. Orkin, W. Swat, *Nature* **374**, 470 (1995).

236. X. Zhao, S. Ghaffari, H. Lodish, V. N. Malashkevich, P. S. Kim, *Nature Structural & Molecular Biology* **9**, 117–120 (2002).

237. Z. Zhou, J. Yin, Z. Dou, J. Tang, C. Zhang, Y. Cao, *Journal of Biological Chemistry* **282**, 23737–23744 (2007).

Appendix: H3N2 Antigenic Replacement Distributions

The distributions of trees underlying the tabulated summary statistics and vaccine effectiveness classifier presented in Chapter 5 are visualized here as point clouds in $\mathbb{P}\Sigma_3$. All triplets of consecutive flu seasons are plotted. Trees colored in **red** have as their most recent sample a hemagglutinin (HA) isolate of the *same* antigenic cluster (A.C.) as that of the strain chosen for the subsequent season's vaccine formulation. Trees colored in **black** have as their most recent sample a HA isolate of a *different* A.C. than that of the strain chosen for the subsequent season's vaccine formulation. Certain triplets of consecutive seasons can only produce red points (9 out of 19), since there is only one circulating A.C. observed at the most recent time point. In those seasons with a single observed A.C. the vaccine strain is never updated (since there is no circulating HA evidence that a change in the vaccine is warranted). The other 10 out of 19 three-season windows possess point clouds of both colors, since their most recent time point contained more than one observed A.C.

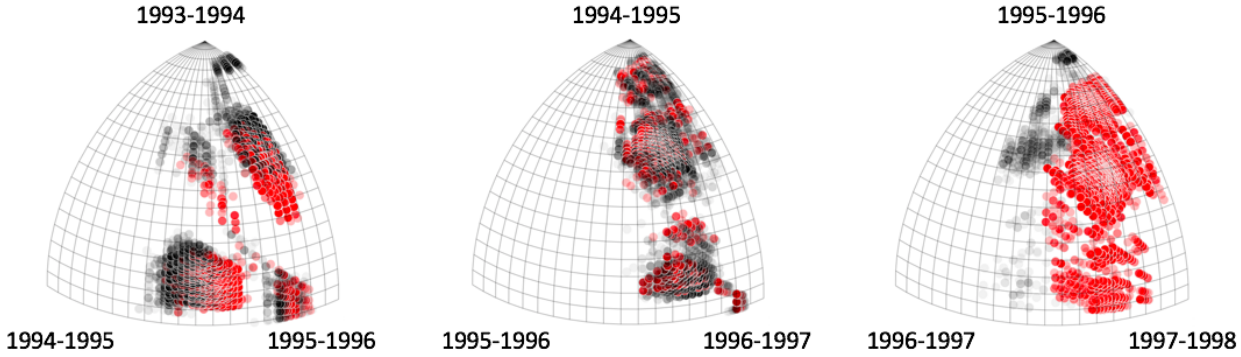


Figure 1: Three-season windows, 1993-1998.

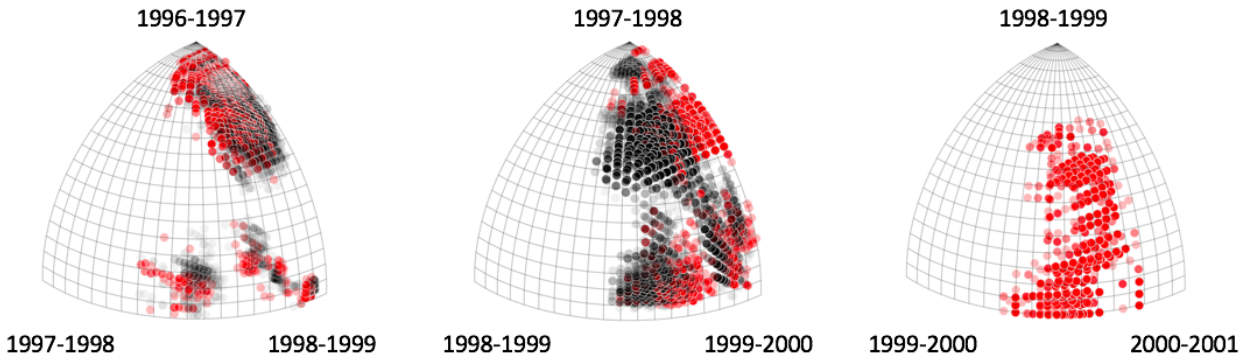


Figure 2: Three-season windows, 1996-2001.

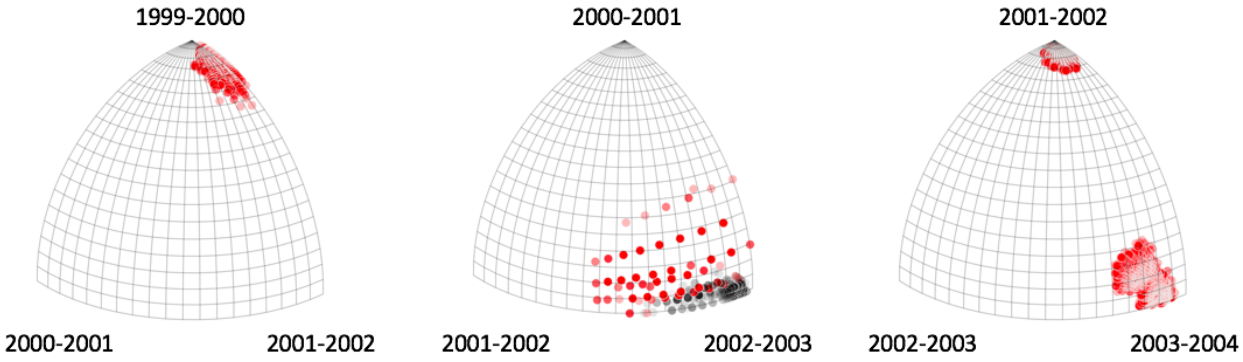


Figure 3: Three-season windows, 1999-2004.

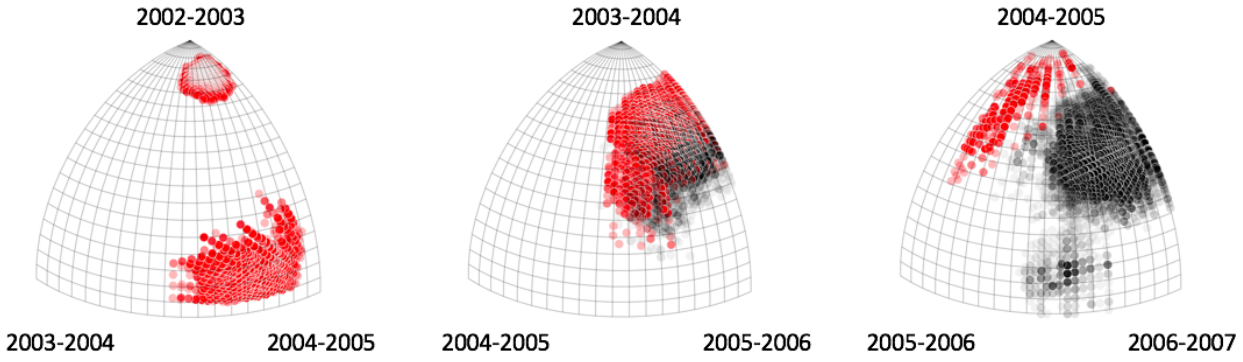


Figure 4: Three-season windows, 2002-2007.

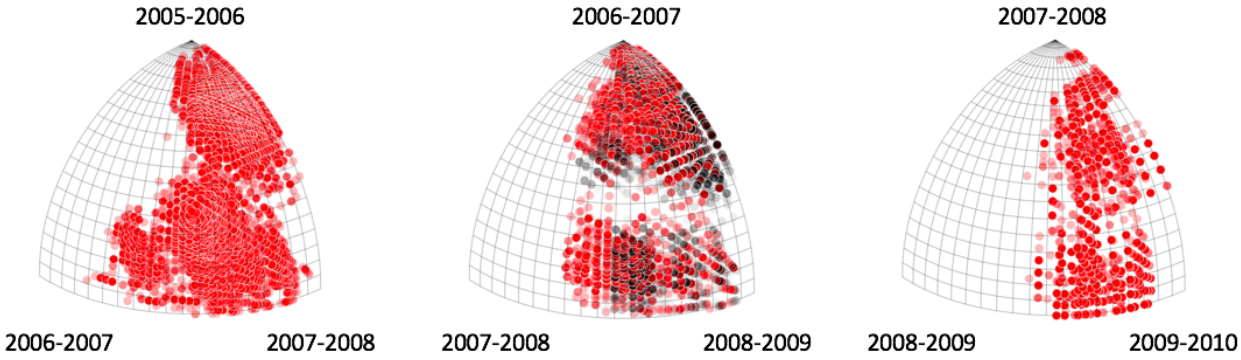


Figure 5: Three-season windows, 2005-2010.

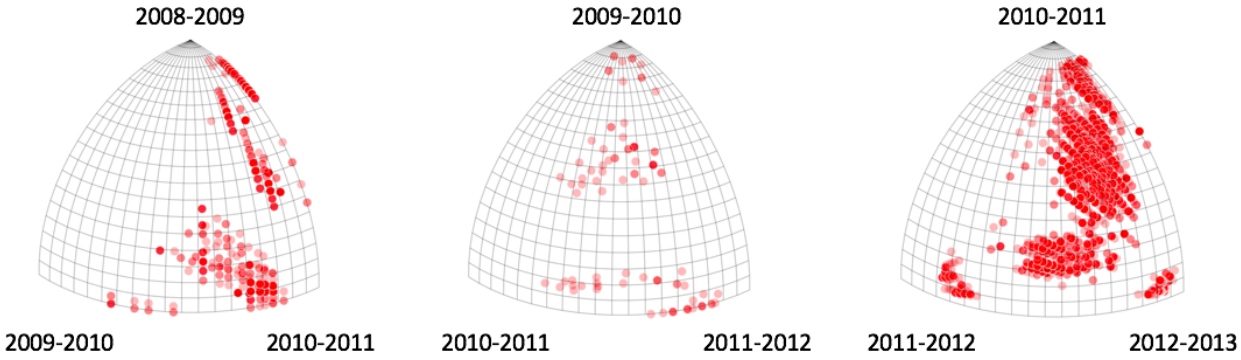


Figure 6: Three-season windows, 2008-2013.

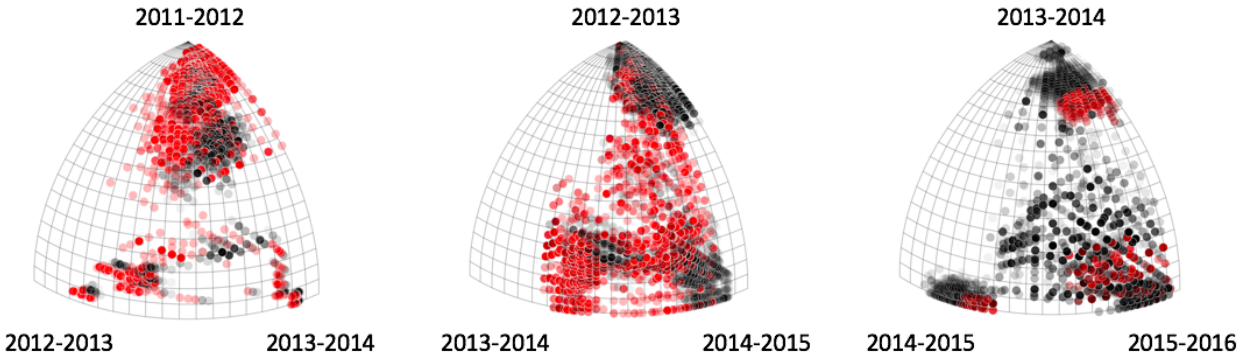


Figure 7: Three-season windows, 2011-2016.