1-1-2015

Detecting Selection on Noncoding Nucleotide Variation: Methods and Applications

Yang Ding
University of Pennsylvania, dingyang@sas.upenn.edu

Detecting Selection on Noncoding Nucleotide Variation: Methods and Applications

Abstract
There has been a long tradition in molecular evolution to study selective pressures operating at the amino-acid level. But protein-coding variation is not the only level on which molecular adaptations occur, and it is not clear what roles non-coding variation has played in evolutionary history, since they have not yet been systematically explored. In this dissertation I systematically explore several aspects of selective pressures of noncoding nucleotide variation:

The first project (Chapter 2) describes research on the determinants of eukaryotic translation dynamics, which include selection on non-coding aspects of DNA variation. Deep sequencing of ribosome-protected mRNA fragments and polysome gradients in various eukaryotic organisms have revealed an intriguing pattern: shorter mRNAs tend to have a greater overall density of ribosomes than longer mRNAs. There is debate about the cause of this trend. To resolve this open question, I systematically analysed 5' mRNA structure and codon usage patterns in short versus long genes across 100 sequenced eukaryotic genomes. My results showed that compared with longer ones, short genes initiate faster, and also elongate faster. Thus the higher ribosome density in short eukaryote genes cannot be explained by translation elongation. Rather it is the translation initiation rate that sets the pace for eukaryotic protein translation. This work was followed by modelling studies of translation dynamics in a yeast cell.

Chapter 3 concerns detecting selective pressures on the viral RNA structures. Most previous research on RNA viruses has focused on identifying amino-acid residues under positive or purifying selection, whereas selection on RNA structures has received less attention. I developed algorithms to scan along the viral genome and identify regions that exhibit signals of purifying or diversifying selection on RNA structure, by comparing the structural distances between actual viral RNA sequences against an appropriate null distribution. Unlike other algorithms that identify structural constraints, my approach accounts for the phylogenetic relationships among viral sequences, as well the observed variation in amino-acid sequences. Applied to Influenza viruses, I found that a significant portion of influenza viral genomes have experienced purifying selection for RNA structure, in both the positive- and negative-sense RNA forms, over the past few decades; and I found the first evidence of positive selection on RNA structure in specific regions of these viral genomes.

Overall, the projects presented in these chapters represent a systematic look at several novel aspects of selection on noncoding nucleotide variation. These projects should open up new directions in studying the molecular signatures of natural selection, including studies on interactions between different layers at which selection may operate simultaneously (e.g. RNA structure and protein sequence).

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Biology

DETECTING SELECTION ON NONCODING NUCLEOTIDE VARIATION: METHODS AND APPLICATIONS

Yang Ding

A DISSERTATION

in

Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2015

Supervisor of Dissertation:

Joshua B. Plotkin
Professor of Biology

Graduate Group Chairperson:

Michael A. Lampson
Associate Professor of Biology

Dissertation Committee:

Joshua B. Plotkin, Professor of Biology, University of Pennsylvania
Paul S. Schmidt, Associate Professor of Biology, University of Pennsylvania
Paul D. Sniegowski, Professor of Biology, University of Pennsylvania
R. Scott Poethig, Professor of Biology, University of Pennsylvania
Brian D. Gregory, Assistant Professor of Biology, University of Pennsylvania
Junhyong Kim, Professor of Biology, University of Pennsylvania

DETECTING SELECTION ON NONCODING NUCLEOTIDE VARIATION: METHODS AND APPLICATIONS

ACKNOWLEDGEMENT

I would first like to thank my advisor, Joshua Plotkin for his unconditional support in the past five years. Whenever I encountered a difficult situation in my PhD study such as the collapse of my results due to a newly discovered bug in my code, he would always assure me everything would be fine, and we could look at the results together and find a way to fix it. I also had the unthinkable freedom in the past five years to pursue whatever research projects I found most interesting. For these and many other aspects I am deeply grateful to him. I would also like to thank my thesis committee members, Professors Paul Schmidt, Paul Sniegowski, Scott Poethig, Brian Gregory and Junhyong Kim, for taking the time to give me insightful suggestions and comments on my research, which have substantially improved many aspects of this dissertation. I also would like to thank many professors I took courses or interacted with during my PhD, an incomplete list including Professors Larry Shepp, Junhyong Kim, Rick Bushman, Michael Steele, Itay Goldstein and Christopher Chen. Their knowledge and wisdom changed the way I think about the world around me, as well as how to conduct scientific research.

I would like to thank my lab members and friends Oana Carja, Premal Shah, David McCandlish, Mitchell Newberry, Jakub Otwinowski, Davorka Gulisija, Alex Stewart, Jia He, Ami Tiyaboonchai, Swathi Ayloo, Tanya Singh, Matias Matiesco, Christel Chehoud, Alexandra Brown, Michael Warner, Kevin Hong, Tianpu Zhang, Kaixuan Yang, Guyue Li, among many others. Without their emotional support I would not have survived my PhD.

In the end I would like to thank my family, parents and grandparents. I grew up in a family where my parents and grandparents from both sides are all schoolteachers or agricultural experts, so I inherited from them the love for nature and respect for knowledge. As a child I was encouraged to buy and read countless number of books, and one of my favorite genre was popular science books. That was the early inspiration for me to later start a career in scientific research. I thus would like to dedicate my dissertation to family.

ABSTRACT

DETECTING SELECTION ON NONCODING NUCLEOTIDE VARIATION: METHODS AND APPLICATIONS

There has been a long tradition in molecular evolution to study selective pressures operating at the amino-acid level. But protein-coding variation is not the only level on which molecular adaptations occur, and it is not clear what roles non-coding variation has played in evolutionary history, since they have not yet been systematically explored. In this dissertation I systematically explore several aspects of selective pressures of noncoding nucleotide variation:

The first project (Chapter 2) describes research on the determinants of eukaryotic translation dynamics, which include selection on non-coding aspects of DNA variation. Deep sequencing of ribosome-protected mRNA fragments and polysome gradients in various eukaryotic organisms have revealed an intriguing pattern: shorter mRNAs tend to have a greater overall density of ribosomes than longer mRNAs. There is debate about the cause of this trend. To resolve this open question, I systematically analysed 5' mRNA structure and codon usage patterns in short versus long genes across 100 sequenced eukaryotic genomes. My results showed that compared with longer ones, short genes initiate faster, and also elongate faster. Thus the higher ribosome density in short eukaryote genes cannot be explained by translation elongation. Rather it is the translation initiation rate that sets the pace for eukaryotic protein translation. This work was followed by modelling studies of translation dynamics in a yeast cell.

Chapter 3 concerns detecting selective pressures on the viral RNA structures. Most previous research on RNA viruses has focused on identifying amino-acid residues under positive or purifying selection, whereas selection on RNA structures has received less attention. I developed algorithms to scan along the viral genome and identify regions that exhibit signals of purifying or diversifying selection on RNA structure, by comparing the structural distances between actual viral RNA sequences against an appropriate null distribution. Unlike other algorithms that identify structural constraints, my approach accounts for the phylogenetic relationships among viral sequences, as well the observed variation in amino-acid sequences. Applied to Influenza viruses, I found that a significant portion of influenza viral genomes have experienced purifying selection for RNA structure, in both the positive- and negative-sense RNA forms, over the past few decades; and I found the first evidence of positive selection on RNA structure in specific regions of these viral genomes.

Overall, the projects presented in these chapters represent a systematic look at several novel aspects of selection on noncoding nucleotide variation. These projects should open up new directions in studying the molecular signatures of natural selection, including studies on interactions between different layers at which selection may operate simultaneously (e.g. RNA structure and protein sequence).

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1

NONE

Chapter 2

Chapter 3

NONE

Chapter 4

NONE

LIST OF FIGURES

Chapter 1

NONE

Chapter 2

Chapter 3

Chapter 4

NONE

Chapter One

Introduction

The role of natural selection in molecular evolution

Genetic drift and natural selection are the two principal forces that shape the evolutionary history of cellular organisms and viruses. Genetic drift is a stochastic force that causes the random fluctuations in allele and phenotypic frequencies without conferring any fitness advantages, whereas natural selection influences allele frequencies deterministically by differential reproduction. Identifying genomic sites that are subject to natural selection has profound intellectual and practical implications.

First, there has been a long-standing debate in evolutionary biology about the relative roles of genetic drift and natural selection in shaping observed molecular variation(Nei, Suzuki, and Nozawa 2010; Fay 2011; Barrett and Hoekstra 2011). Charles Darwin advocated natural selection as the main force in shaping evolutionary history in his On the Origin of Species (Darwin 1859), although, admittedly at the time he was not aware of or discussing molecular phenotypes. Since then, many instances of adaptation on the morphological level have been studied and documented in detail. With the dawn of molecular biology and the development and application of a series of techniques to measure protein sequence variation (Hubby and Lewontin 1966; Harris 1966) in the 1960s, however, a surprisingly large amount of protein variation was observed among different species, as well as among individuals within a single species. This posed serious challenges to Darwin's selectionist view, and motivated Motoo Kimura to develop his

neutral theory of molecular evolution (Kimura 1983). The neutral theory contends that most observed molecular variation is due to random fixation of neutral mutations that do not bear any selective advantages. Nonetheless, in the past decade, increasingly more convincing instances of molecular adaptation have been identified, with their fitness advantages and proximate molecular mechanism sorted out. The availability of whole genome assays and sequences from many different species and individuals has prompted many whole-genome scans (for a partial list see (Haasl and Payseur 2015)) for sites under positive selection in human lineages (Grossman et al. 2013; Lachance and Tishkoff 2013; Enard, Messer, and Petrov 2014), *Drosophila Melanogaster* (Sella et al. 2009; Langley et al. 2012; Pool et al. 2012; Fabian et al. 2012; Reinhardt et al. 2014; Bergland et al. 2014), *Arabidopsis thaliana*(Hancock et al. 2011; Huber et al. 2014), among many others. Now we know that there are many confirmed loci in the eukaryotic genomes that are undergoing adaptation, with many more candidate sites waiting to be validated.

One of the central topics in evolutionary biology remains the study of the molecular and mechanistic basis of positive selection, i.e. adaptation. How do organisms respond to new environmental pressures? Do adaptive changes mostly happen in the protein-coding genes or regulatory sequences? Do adaptations mostly come from newly arisen advantageous alleles or from standing variations? Is genetic adaptation more likely to be driven by a small number of alleles that have a large effect, or by a large number of alleles that have relatively moderate effects? These questions are still attracting much attention from evolution researchers(Hendry 2013).

Studying genomic sites under selection has practical implications as well. Sequence and structural conservation, a strong signal for negative selection, has been used extensively in the comparative genomics and RNA bioinformatics community to identify functional genetic elements. The conserved sites in the genomes can be protein-coding genes, noncoding RNAs, microRNA targets, transcription factor binding sites or other regulatory sequences.

Identifying genomic sites under selection also offers great potential for biomedical applications. Infectious pathogens have been shown to be among the strongest sources of selective pressures on human populations during local adaptation (Fumagalli et al. 2011; Karlsson, Kwiatkowski, and Sabeti 2014). For example, analyzing the genomic loci that are associated with elevated immune response against malaria can potentially offer novel therapeutic strategies for malaria treatment (Kwiatkowski 2005). Also recent research suggests that even the contemporary human population is under constant selective pressures for certain phenotypic traits (Byars et al. 2010; Stearns et al. 2010; Milot et al. 2011), many of which are probably related to human health and diseases. Understanding the selective pressures that we are currently experiencing is likely to help the treatment and prevention of these diseases.

Understanding the major adaptations during human evolution will help us recapitulate important historical events that have shaped our species. A notable example is the research showing that lactase expression in adults has independently arisen at least twice during human evolution (Bersaglieri et al. 2004; Tishkoff et al. 2007), the timing of which are coincident with the introduction of cattle domestication in Europe and the

practice of pastoralism in East Africa, respectively. Further explorations of whole-genome sequencing data from multiple human populations will undoubtedly reveal more interesting stories about important periods in human history.

Equally interesting, and just as practical, are the questions of selection pressures on microbial pathogens themselves, often mediated by host immune systems or requirements for host specificity. These questions are particularly acute for rapidly evolving viruses, which must regularly contend with immune or chemotherapeutic pressure, and whose course of evolution, in turn, may inform vaccination of drug treatment decisions.

Different modes and sources of natural selection

The simplest mode of natural selection is directional selection. There are two types of directional selection: negative (purifying) selection refers to the selective elimination of deleterious alleles, while positive selection drives evolutionary innovation by promoting the spread of beneficial alleles. The melanism of pepper moth (Cook et al. 2012) is a classic example in directional natural selection. This happened in the mid - 19th century at Manchester, England, when industrial revolution turned Manchester into an industrial city, and the tree barks were darkened by soot from the new coal-burning factories. Previously dominant light-colored pepper moths suddenly contrasted with the color of the barks, while the dark-colored moths were camouflaged well by the darkened trees. This led to increased predation of the light-colored moths by predating birds, and by 1895, the percentage of dark-colored moths in Manchester increased to 98%, and the

light-colored moths almost went extinct. Whereas the mechanistic details of this classic example of directional selection remain hotly contested, because some of the original field experiments were flawed (Majerus 1998), the example remains a classic story (if partly fictional) of directional selection in the wild.

More complicated selective scenarios include balancing selection, where multiple alleles are maintained at an appreciable frequency in the gene pool. One recent example (Bergland et al. 2014; Behrman et al. 2015; X. Zhao et al. 2015) is in Drosophila species. One study (Bergland et al. 2014) found hundreds of polymorphisms that undergo dramatic seasonal shift in allele frequencies in *Drosophila Melanogaster*, suggesting temporally varying selective pressures. In particular, stress tolerance traits such as chill coma recovery time and starvation tolerance seem to be favored in winter, and disfavored during summer, when they are no longer needed.

Not only there are different modes of natural selection, selective pressures can also occur on many different levels of biological organization. There can be a single preferred amino acid mutation - for example, a single amino acid mutation in melanocortin-1 receptor (Mc1r) turned Florida's Gulf Coast beach mice into light-color, which helped them evade their visual predators (Hoekstra et al. 2006). There can also be mutations in regulatory elements – two different SNPs in the 13^{th} introns of gene MCM6 can enhance the promoter activity of the lactase encoding gene LCT in African and European populations (Tishkoff et al. 2007), thus gave them survival advantages for being able to consume milk products.

Recently it has been increasing appreciated that coding sequences harbor numerous regulatory sites that are independent of their protein-coding function (Plotkin and Kudla 2011), such as RNA localization(Jambhekar and Derisi 2007), translation efficiency(Sharp and Li 1987), mRNA splicing (Fairbrother et al. 2002), mRNA stability (Kudla et al. 2006) and accessibility to the translation machinery (Nackley et al. 2006). As an example, it has been shown that in virtually all free-living (Keller et al. 2012; Gu, Zhou, and Wilke 2010) and many viral species (Zhou and Wilke 2011) the region around the translation start site of each mRNA transcript is under natural selection to be less structured, presumably for the efficient recognition of the start codon by initiator-tRNAs. The role of selection on these non-coding sites remains largely unexplored, and is the central theme in this dissertation.

Methods to detect signatures of natural selection

Methods to detect signatures of natural selection can be broadly divided into two classes: Methods to detect selection on the macro-evolutionary level based on comparisons of different species and their relative rates of genetic change, and population-genetics methods to detect selection occurring within a population, often including comparison to one outgroup species (Vitti, Grossman, and Sabeti 2013).

Methods to detect micro-evolutionary (within-population) selection include site-frequency based methods (Ewens 1972; Watterson 1978; Fu and Li 1993; Fu 1997; Tajima 1989; Tajima 1993; Fay 2011), linkage disequilibrium based methods (Sabeti et al. 2002; C. Zhang et al. 2006; Hanchard et al. 2006; Sabeti et al. 2007; Voight et al. 2006;

E. T. Wang et al. 2006; Cai et al. 2011; Han and Abney 2013) and population differentiation based methods (Bonhomme et al. 2010; Excoffier, Hofer, and Foll 2009; Lewontin and Krakauer 1973; Vitalis, Dawson, and Boursot 2001; Shriver et al. 2004; Fariello et al. 2013). There are also methods that combine the signals from these different methods (Kim and Nielsen 2004; Kim and Stephan 2002; Nielsen et al. 2009; Nielsen et al. 2005; Hua Chen, Patterson, and Reich 2010; Zeng et al. 2006; Zeng, Shi, and Wu 2007; Grossman et al. 2010; Grossman et al. 2013). Site frequency methods consider the distribution of frequencies of a set of SNPs in a population, where a surplus of rare alleles would be indicative of recent positive selection or population expansion. Linkage disequilibrium methods search for genomic regions with an unexpected low degree of genetic diversity, called high linkage-disequilibrium, presumably because a newly emerged advantageous mutation has swept through the entire population. Population differentiation methods use measures such as the fixation index (F_{ST}) to measure genetic differences within a population vs. between populations, and a high level of F_{ST} would imply that all genetic variation could be explained by population structure, and the two populations do not share much gene flow.

The methods for macro-evolutionary selection compare the orthologous genes from multiple species to see if there is a signal of natural selection over long timescales. There are two well-known methods in this category: the first is the McDonald-Kreitman test (Hudson, Kreitman, and Aguade 1987; McDonald and Kreitman 1991; Egea, Casillas, and Barbadilla 2008), which compares the ratio of synonymous vs. nonsynonymous mutation rates in individuals within single species vs. from multiple species. The idea is

that an advantageous mutation will quickly fix within a species and lead to the fixed differences among species, thus adaptive mutations should contribute more to between-species substitutions than within-species polymorphisms. The second method, which uses no within-population data at all, is the dN/dS test (Goldman and Yang 1994; Yang 2000; Hurst 2002), which compares the rates of synonymous and non-synonymous substitution rates across evolutionarily divergent lineages. It is assumed that synonymous substitutions among species are strictly neutral. If there is no selective pressure operating, then the rates of synonymous and non-synonymous substitutions are expected to be equal. If the rate of synonymous substitution is higher, then it means the protein is under negative selection to keep the amino acid sequence intact, while a higher non-synonymous substitution rate would imply the protein is under pressure to change its content, probably for the need to adapt to new environments.

Several variations of dN/dS test have been introduced. For examples, (Hoffman and Birney 2007) proposed to use nucleotide substitution rate in pairs of orthologous introns, and use this measure dI as an alternative of dS. In another study, (Hoffman and Birney 2010) proposed a dT/dS test to study the natural selection on promoter sequences, in this test dT denotes the TF binding affinity changes of the promoter. Also inspired by the dN/dS test, (Han Chen et al. 2015; Han Chen, Xing, and He 2015) proposed dJ/dS and dT/dS to study cancer driver genes, where dJ denotes the mutation rate at exon/intron junction and dT denotes the rate of truncating mutations.

Most of the methods mentioned above assume a scenario of hard sweep, i.e. a single beneficial mutation arises from the population and sweeps through the entire

population. However there are many more complicated scenarios of natural selection. For example, there may be soft selective sweeps (Messer and Petrov 2013), in which multiple adaptive alleles sweep through the population at the same time. This happened when brown rats rapidly developed several different allele variants of the gene encoding vitamin K epoxide reductase complex subunit 1 (VKORC1) in response to the rodenticide warfarin (Pelz et al. 2005). Several methods (Messer and Neher 2012; Garud et al. 2015) have been developed to detect selection in this more complicated scenario.

Protein translation dynamics and ribosome profiling: a role for selection?

One of the more intriguing objects of natural selection – which has not received considerable attention yet – is selection on translational efficiency and accuracy. Protein translation is arguably one of the most fundamental biological activities that occur in a living cell. Although individual steps in translation such as the formation of the 43S preinitiation complex are known intricate molecular detail, a global understanding of how these steps combine to set the pace of protein production for individual genes remains elusive (Jackson, Hellen, and Pestova 2010; Plotkin and Kudla 2011). Various factors such as codon usage bias, gene length, transcript abundance, and translation initiation rate are all known to modulate protein synthesis (Bulmer 1991; Chamary and Hurst 2005; Cannarozzi et al. 2010; Tuller et al. 2010; Shah and Gilchrist 2011; Plotkin and Kudla 2011; Gingold and Pilpel 2011; Chu, Barnes, and Haar 2011; Chu and Haar 2012), but how they interact with each other to collectively determine translation rates of all genes in a cell was poorly understood. It remains difficult to make systematic measurements for

some of the critical parameters in a cellular process, such as gene-specific rates of 5'UTR scanning and start codon recognition. As a result, fundamental questions such as the relative role of translation initiation vs. elongation in setting the pace of protein production were actively debated (Kudla et al. 2009; Tuller et al. 2010; Gingold and Pilpel 2011; Chu, Barnes, and Haar 2011; Chu and Haar 2012; Y. Ding, Shah, and Plotkin 2012).

Thanks to the development of expression profiling technologies such as microarrays (Brown and Botstein 1999) and mRNA sequencing (RNA-seq) (Z. Wang, Gerstein, and Snyder 2009) in the past two decades, we now have the ability to simultaneously monitor the mRNA levels of tens of thousands of genes and their changes under various physiological conditions. However, it has been repeatedly shown that transcriptional regulation is only half the story (Plotkin 2010) – the correlation between protein levels and mRNA levels is often weak, and this is at least partially due to the effect of translational regulation. Thus direct analyses of the translation processes can provide a more complete and accurate picture of gene expression in cells than mRNA levels alone. Ribosome profiling, which was first introduced in 2009 (Ingolia et al. 2009), is a deep-sequencing based technology to measure the global cellular translational activity *in vivo* (Ingolia 2014; Brar et al. 2012). It leverages the observation that a translating ribosome can protect about 30 nucleotides of an mRNA from nuclease activity, and by sequencing these remaining 30 nucleotides one can see the "ribosome footprints" left on each mRNA transcript with a nucleotide-level resolution.

Ribosome profiling has provided us with a much more detailed view of protein translation dynamics, and helped us resolve many of the outstanding questions and debate in the field of protein translation dynamics (Ingolia 2014; Brar and Weissman 2015). For example, (Shah et al. 2013) developed a whole-cell stochastic model of yeast translation process and used the ribosome-profiling data from (Ingolia et al. 2009) to parameterize the model. The model showed translation initiation, rather than elongation, is the rate-limiting step in yeast endogenous protein translation. (Weinberg et al. 2015) compared multiple ribosome profiling datasets from yeast (Ingolia et al. 2009; Gerashchenko, Lobanov, and Gladyshev 2012; Zinshteyn and Gilbert 2013; Artieri and Fraser 2014; Guydosh and Green 2014; McManus et al. 2014; Weinberg et al. 2015), and showed that a simple multiple linear regression using six features, including mRNA abundance, upstream open reading frames, cap-proximal RNA structure and GC content, length of coding and 5'UTR regions, can explain most of the observed variation in yeast translation efficiency.

The extent to which selection on non-coding sequence variation is mediated by requirements for gene translation – and variation in these requirements across genes – remains largely unexplored and is one of the central topics of this dissertation.

Prediction and measurement of RNA secondary structures

RNAs play vital roles in myriad cellular functions, including transcription, RNA processing, and translation. They adopt complex structures to perform their functional roles in living cells. One interesting example is riboswitches. Riboswitches are a class of

RNA molecules that can change their structural conformations (Breaker 2011; Breaker 2012) upon the binding of certain metabolites, and switch a gene “on” or “off”. Here I review some basic biology and computational work on RNA structure, with an eye towards eventual analysis of natural election on RNA structures.

RNA secondary structures refer to all the base-pairing inside the RNA molecule. Pseudoknots refer to a special class of secondary structures where a nucleotide inside a loop forms base pair with a nucleotide outside this region. The prediction of pseudoknots is NP-complete in general (Lyngsø and Pedersen 2000), but some restricted classes of pseudoknots are still computational tractable (Rivas and Eddy 1999; Dirks and Pierce 2004; Ren et al. 2005; Cao and Chen 2006), although at a much higher computational complexity than predicting the plenary secondary structures. RNA tertiary structures refer to the 3D structure of an RNA molecule. Although there have been several recent attempts in prediction RNA tertiary structures (Das and Baker 2007; F. Ding et al. 2008; Parisien and Major 2008; Frellsen et al. 2009; Jonikas et al. 2009; Popenda et al. 2012; Y. Zhao et al. 2012; Kerpedjiev, Höner Zu Siederdissen, and Hofacker 2015), the field of RNA tertiary structure prediction is still in its infancy, and there are currently no algorithms that can reliably predict tertiary structures from RNA sequences alone. Because of the above reasons, in the remainder of this dissertation we will focus on pseudo-knot free RNA secondary structures.

Broadly speaking, there are three strategies in predicting RNA secondary structures: There are two approaches to predict RNA secondary structures: One is comparative genomics, in which structures are inferred by the base-pair covariation of

RNA sequences from multiple species (Knudsen and Hein 2003; J. S. Pedersen et al. 2006; Nawrocki, Kolbe, and Eddy 2009). The idea is that although the primary sequences may change, RNA base pairs will co-vary so as to maintain the secondary structures. Another approach is to use thermodynamics-based algorithms (Zuker 1989; Hofacker et al. 1994; Reuter and Mathews 2010), in which the RNA secondary structures are classified as structural motifs such as hairpins, bulges, internal loops and multiloops, each of which are assigned an experimentally-derived energy score, and the secondary structure is predicted as the one with the minimum free energy. Also there are algorithms that combine these two signals(Havgaard, Torarinsson, and Gorodkin 2007; Reuter and Mathews 2010), but they are also computationally more expensive. Although these approaches have been widely used by experimental and computational biologists, the accuracy of the structure prediction algorithms is still pretty limited. The RNA community has long suffered from the lack of high-throughput, accurate measurements of RNA secondary structures. The situation has recently changed due to the development of several sequencing-based RNA structure probing technologies(Kwok et al. 2015; Foley et al. 2015). These techniques have allowed the experimental measurements of RNA base pairings on a whole transcriptome level and will greatly facilitate our understanding of the roles that RNA structures play in various cellular processes.

As with translation, the extent to which non-coding variation in genomes is subject to selection pressures mediated by requirements for proper RNA structure remains largely unexplored, and is a central question in this dissertation.

Evolution of influenza A viral genomes

Influenza A viruses (Nelson and Holmes 2007; Bouvier and Palese 2008) are single-stranded, negative-sense RNA viruses that can infect and cause seasonal epidemics in humans, birds, and other animal species. The genome of influenza A viruses comprise eight viral RNA segments, and each segment encodes 1-2 viral proteins. Influenza A viruses are further characterized by the subtype of their surface glycoproteins, the hemagglutinin (HA) and the neuraminidase (NA). While many genetically distinct subtypes (16 for HA, 9 for NA) have been found in circulating influenza A viruses, only three HA (H1, H2 and H3) and two NA (N1 and N2) have caused human epidemics (Bouvier and Palese 2008). These surface proteins are the targets of human immune system and possibly antiviral drugs (Nelson and Holmes 2007; Bloom, Gong, and Baltimore 2010), so they are under strong selective pressure to evolve resistance. The influenza A genome can achieve this through two processes: one is antigenic drift, characterized by the gradual accumulation of mutations on the antibody-binding sites of the surface proteins so that it can evade the surveillance of the immune system, and this sometimes will cause seasonal epidemics. The other mechanism is called antigenic shift, a much more rare incidence where two or more strains of influenza viruses combine to form a new subtype that has a mixture of surface antigens from these strains. This kind of segment reassortment can happen when the same cells were simultaneously infected by different strains of human and animal viruses, and the resulting viruses can potentially encode novel surface antigens that human populations have no preexisting immunity. Influenza viruses which have undergone antigenic shift has caused many recent flu

pandemics, including the most recent 2009 H1N1 outbreak, where viral reassortment happened between human, avian and swine viruses (Smith et al. 2009).

Since influenza A viruses are such global public health threat, there have been many efforts trying to understand the evolutionary constraints that are imposed upon the influenza A genome, and potentially use these information to predict the flu strains that may become prevalent in the following year. Numerous methods have been proposed to search for amino acid residues or patches that are under selective pressures from influenza sequences.(Bush et al. 1999; Yang 2000; Suzuki 2006; Kosakovsky Pond et al. 2008; X. Ding et al. 2010; Tusche, Steinbruck, and McHardy 2012). Some recent studies (Kryazhimskiy et al. 2011; Neverov et al. 2015) also attempt to identify the pairs of amino acid residues within or among segments that co-evolve each other, so the knowledge of one residue mutated in the "epistatic pair" may help one make the prediction that the other residue may also mutate soon. Another cellular process that is vital to the integrity of influenza viral life cycle is the viral packaging. One challenge the nascent virion needs to face is to assemble its complete genome from a pool of RNA segments. It is known that the presence of conserved terminal promoter sequences at the 5' and 3' end of each viral RNA is necessary to distinguish itself from cellular RNAs (Hutchinson et al. 2010). The 5' and 3' sequences are partially base-paired to form a characteristic panhandle or corkscrew structure. However, to correctly assemble its 8 distinct viral segment, influenza A viruses also need segment-specific packaging signals. Various methods, 1) including studying defective-interfering RNAs, 2) finding the sequence required to efficient package reporter genes, 3) sequence conservation and 4)

analyzing the effect of point mutations on packaging, have been used to probe segment-specific packaging signals.

Besides identifying evolutionary constraints on the influenza A genome, there have also been several studies that try to predict the prevalent strains of influenza viruses in the near future by building a sequence-based influenza viral fitness model (Luksza and Lässig 2014) or by extracting the information from the influenza viral phylogeny (Neher, Russell, and Shraiman 2014). These represent the new frontiers in understanding and predicting influenza viral evolution.

Overview of the dissertation

There has been a long tradition in molecular evolution to study selective pressures operating at the protein level. But protein-coding variation is not the only level on which molecular adaptations occur, and it is not clear what roles non-coding variation has played in evolutionary history, since they have not yet been systematically explored. The absence of technical tools to detect positive selection on non-coding variation is one of the major obstacles along this road.

In this dissertation I systematically explore several aspects of the selective pressures of noncoding nucleotide variation:

Chapter 1 is the General Introduction. It provides the necessary background for the entire dissertation, and it sets the stage for the following discussions of selection on noncoding nucleotide variations.

Chapter 2 describes a research project on the determinants of eukaryotic translation dynamics, which includes selection on non-coding aspects of DNA variation. Deep sequencing of ribosome-protected mRNA fragments (Ingolia et al. 2009) and polysome gradients in budding yeast (Arava et al. 2003) have revealed an intriguing pattern: shorter mRNAs tend to have a greater overall density of ribosomes than longer mRNAs. The same trend has been found in mouse, human, fruit fly, Arabidopsis, malaria, and fission yeast: shorter Open Reading Frames (ORFs) tend do exhibit more densely packed ribosomes (Ingolia et al. 2009; Cataldo, Mastrangelo, and Kleene 1999; Branco-Price et al. 2005; Qin et al. 2007; Hendrickson et al. 2009; Lacsina et al. 2011). There is debate about the cause of this trend. To resolve this open question, I used 5'mRNA secondary structure as a proxy for translation initiation rate, Codon Adaptation Index (CAI, a measure of biased synonymous codon usage) as a proxy for translation elongate rate, and systematically analysed 5'mRNA and CAI patterns in short versus long genes, within each of about 100 sequenced eukaryotic genomes. My results showed that compared with longer ones, short genes initiate faster, and also elongate faster. Thus the higher ribosome density in short eukaryote genes cannot be explained by translation elongation. Rather it is the translation initiation rate that sets the pace for eukaryotic protein translation.

The published research paper arising from Chapter 2, describing my studies on ORF length and 5' mRNA structure (Y. Ding, Shah, and Plotkin 2012), provides a statistical analysis and gives us a global view of the relative roles of translation initiation vs. elongation. To get a mechanistic understanding of the various aspects of protein

translation dynamics, I also helped to develop and parameterize a whole-cell stochastic model of the protein translation process that keeps track of every mRNA, tRNA, and ribosome in a cell (Shah et al. 2013). Using this mechanistic model, we showed that indeed translation initiation is the rate-limiting step in yeast endogenous protein translation, at least in healthy growing cells. Even though I am an author on that paper (Shah et al. 2013), I do not describe my work on the mechanistic modeling of translation in this dissertation, because the project was highly collaborative involving other members of the Plotkin lab and it is not directly related to selection pressures on non-coding variation.

Chapter 3 concerns detecting selective pressures on the influenza A viral RNA structures. Influenza A viruses are negative-sense RNA viruses that cause significant human morbidity and mortality each year. Rapid evolution of antigenic surface proteins allows the virus to re-infect hosts who have recovered from prior strains. It is therefore important to understand the selective pressures that shape the evolutionary trajectories of influenza viral genomes. Most previous research has focused on identifying amino acid residues experiencing positive or purifying selection, whereas selection on RNA structures has received less attention. Here we develop algorithms to scan along the viral genome and identify regions that exhibit signals of purifying or diversifying selection on RNA structure, by comparing the structural distances between actual viral RNA sequences against an appropriate null distribution. Unlike other algorithms that identify structural constraints, our approach accounts for the phylogenetic relationships among viral sequences, as well the observed variation in amino-acid sequences. Our approach

can also detect recent selective pressures, which are of considerable practical interest, including recent positive selection. Our results indicate that a significant portion of influenza A viral genomes have experienced purifying selection for RNA structure, in both the positive- and negative-sense RNA forms, over the past few decades; and we provide the first evidence of recent positive selection on RNA structure in specific regions of these viral genomes. We also identify genomic regions where viral RNA structures may have played a role during shifts from avian to human hosts.

Chapter 4 summarizes the results from previous chapters and provides some perspective on the possible future developments in areas related to the research presented in this dissertation.

Overall, the projects presented in these chapters represent a systematic look at several novel aspects of selection on noncoding nucleotide variation. These projects should open up new directions in studying the molecular signatures of natural selection, including studies on interactions between different layers at which selection may operate (e.g. RNA structure, protein sequence, etc).

Besides the papers discussed above, from 2010-2015 I have also been involved in the following additional publications which are not described explicitly in this dissertation document: (Y. Ding, Grünewald, and Humphries 2011) is a theoretical paper in phylogenetic analysis, where we improved the upper and lower bounds between maximal possible distance between two trees of n leaves. This improves our understanding of the mathematical properties of several tree-editing distance measures. (Y. Ding, Lorenz, and Chuang 2012) presents a motif discovery algorithm to search for

over-representative motifs in protein-coding sequences, correcting for the amino acid background. This, together with a conservation-based algorithm I helped develop earlier (Kural et al. 2009), gives us a set of computational tools to search for functional sequence motifs in protein coding sequences. (Y. Ding et al. 2014) presents algorithms to calculate the partition functions and probabilities of an RNA molecule adopting a secondary structure with k hairpins or multiloops, where k is a positive integer. This gives us a tool to calculate the probability that an RNA molecule can adopt a certain shape. (McCandlish et al. 2013) examines the analysis of (Breen et al. 2012) and showed that their analysis didn't prove epistasis is the primary factor in molecular evolution, as they initially suggested in the paper.

Chapter Two

Systematically Weaker 5'-mRNA Secondary Structures in Short Eukaryotic Genes

Abstract

Experimental studies of translation have found that short genes tend to exhibit greater densities of ribosomes than long genes in eukaryotic species. It remains an open question whether the elevated ribosome density on short genes is due to faster initiation or slower elongation dynamics. Here we address this question computationally using 5' mRNA folding energy as a proxy for translation initiation rates, and codon bias as a proxy for elongation rates. We report a significant trend towards reduced 5' secondary structure in shorter coding sequences, suggesting that short genes initiate faster during translation. We also find a trend towards higher 5' codon bias in short genes, suggesting that short genes elongate faster than long genes. Both of these trends hold across a diverse set of eukaryotic taxa. Thus, the elevated ribosome density on short eukaryotic genes is likely caused by differential rates of initiation, rather than differential rates of elongation.

Introduction

Synonymous sites in coding sequences have long been used as a neutral yardstick against which to compare amino-acid changing substitutions, in the hope of detecting either purifying or positive selection on proteins (Goldman and Yang 1994; Kimura 1977;

McDonald and Kreitman 1991; Muse and Gaut 1994). Nonetheless, synonymous mutations are known to experience selection in many cases (Andersson and Kurland 1990; Chamary and Hurst 2005; Duret 2002; Hershberg and Petrov 2008; Sawyer and Hartl 1992; Sharp et al. 1995; Sharp, Emery, and Zeng 2010) for a variety of mechanisms, including the efficiency of gene translation, the stability of mRNAs (Capon et al. 2004; Chamary and Hurst 2005; Chamary, Parmley, and Hurst 2006; Duan et al. 2003; Shah and Gilchrist 2011; Shen, Basilion, and Stanton 1999) especially near the translation initiation site (Gu, Zhou, and Wilke 2010; Keller et al. 2012; Kudla et al. 2009), the regulation of splicing, among others (Plotkin and Kudla 2011). The fact that synonymous mutations have phenotypic and fitness consequences complicate the interpretation of measures of selection, such as the ratio of substitution rates at synonymous and non-synonymous sites, dN/dS [(Goldman and Yang 1994; Kimura 1977; Muse and Gaut 1994) but see (Hirsh, Fraser, and Wall 2005)].

Selection for translational efficiency remains the dominant explanation for systematic variation in codon usage among the genes in a genome, in diverse taxa (Plotkin and Kudla 2011). In accordance with this explanation, codon bias towards the most abundant iso-accepting tRNA species is generally strongest in those genes expressed at high levels, where efficiency would confer the greatest selective benefit to the cell. Nonetheless, the specific mechanisms by which codon bias confers relative fitness gains are actively debated (Plotkin and Kudla 2011; Shah and Gilchrist 2010).

Our understanding of the dynamics of gene translation, and the role of codon bias in translation, will benefit from new experimental techniques that parse the detailed kinetics

of translation across the entire transcriptome. Especially promising are techniques that use high-throughput sequencing of ribosome-protected RNA to determine a "ribosomal footprint" on each mRNA (Bazzini, Lee, and Giraldez 2012; Brar et al. 2012; Guo et al. 2010; Ingolia et al. 2009; Ingolia, Lareau, and Weissman 2011; G. W. Li, Oh, and Weissman 2012; Oh et al. 2011; Reid and Nicchitta 2012) with greater accuracy than earlier, polysome-based techniques (Arava et al. 2003). Among many other intriguing findings, these experiments have shown that the cell-wide average profile of ribosome densities in yeast exhibits a trend of decreasing ribosome density with codon position, from 5' to 3' – an observation that has been explained, in part, by a trend towards less biased codon usage in the 5' ends of genes, associated presumably with slower elongation and thus higher ribosome density (Tuller et al. 2010).

Aside from the 5' ramp of elevated ribosome densities, sequencing (Ingolia et al. 2009) and polysome gradients in budding yeast (Arava et al. 2003) have also revealed another, possibly independent finding: shorter mRNAs tend to have a greater overall density of ribosomes than longer mRNAs. The same trend has been found in mouse, human, fruit fly, Arabidopsis, malaria, and fission yeast: shorter ORFs tend to exhibit more densely packed ribosomes (Branco-Price et al. 2005; Cataldo, Mastrangelo, and Kleene 1999; Hendrickson et al. 2009; Ingolia et al. 2009; Lackner et al. 2007; Lacsina et al. 2011; Qin et al. 2007).There is debate about the cause of this trend. Some authors have attributed this relationship to a constant-length ramp of elevated 5' density on all transcripts due to elongation dynamics (Ingolia et al. 2009) (so that shorter transcripts would be observed to have larger overall ribosome density); and others have attributed

this trend to an increased rate of initiation in short yeast genes causing an increased density of ribosomes (Arava et al. 2003; Arava et al. 2005; Lackner et al. 2007). As a result, at present it is unclear whether the greater overall density of ribosomes on short yeast genes is caused by a greater rate of initiation for such genes, or a slower rate of early elongation in those genes.

Against this backdrop of open questions, here we analyze the relationship between ORF length and measures of initiation and early elongation rates, across a diverse set of eukaryotic species. As a proxy for the initiation rate of a gene we use the computationally predicted energy of its 5' mRNA structure – a quantity that has been shown experimentally to correlate strongly with protein levels (Kudla et al. 2009) and which has been subject to natural selection in virtually all free-living (Gu, Zhou, and Wilke 2010; Keller et al. 2012; Tuller et al. 2010) and many viral species (Zhou and Wilke 2011). As a proxy for the early elongation rate of a gene we use the codon adaptation index (CAI) (Sharp and Li 1987) of its early codons (Tuller et al. 2010). In general, by performing these analyses we seek to understand whether the trend towards elevated ribosome densities in short genes (Arava et al. 2003; Arava et al. 2005; Branco-Price et al. 2005; Cataldo, Mastrangelo, and Kleene 1999; Hendrickson et al. 2009; Ingolia et al. 2009; Lackner et al. 2007; Lacsina et al. 2011; Qin et al. 2007) is caused by faster initiation in those genes, slower early elongation in those genes, or both.

Results

Codon bias, mRNA structure, and ORF length in *C. elegans*

We first investigated the relationship between ORF length and 5' mRNA folding in the model species *C. elegans*, as well as the relationship between ORF length and 5' codon bias. As described above, we use these two measures as proxies for the initiation rates and early elongation rates of genes. In particular, for each *C. elegans* transcript, we computed its predicted folding energy from nucleotide -4 to +37 (Kudla et al. 2009) relative to start, using RNAfold (Hofacker et al. 1994), and we computed the CAI of its first 50 codons. (We systematically explore alternative definitions of 5' CAI below.)

We performed a Spearman rank correlation test between 5' mRNA folding energy and ORF length, among the 29857 transcripts in *C. elegans* (Assembly WS220). We likewise performed a rank correlation test between 5' CAI values and ORF lengths. Our expectation was that compared with long genes, short genes should tend to have faster initiation rates and/or slower early elongation rates – in order to explain the tendency towards elevated ribosome densities on short genes (Arava et al. 2003; Arava et al. 2005; Branco-Price et al. 2005; Cataldo, Mastrangelo, and Kleene 1999; Hendrickson et al. 2009; Ingolia et al. 2009; Lackner et al. 2007; Lacsina et al. 2011; Qin et al. 2007). Of these two alternative mechanisms we might in principal expect the initiation-driven mechanism to be a stronger determinant of ribosome densities(Andersson and Kurland 1990; Bulmer 1991; Lackner et al. 2007).

In accordance with these expectations, we found a significant negative rank correlation (Spearman rho = -0.12, $p < 7e\text{-}90$) between 5' mRNA folding energy and ORF length, indicating a tendency towards weaker mRNA structure and presumably faster initiation in short *C. elegans* genes (Fig. 1). On the other hand, we also found a

significant negative rank correlation (Spearman rho = -0.16, p < 5e-179) between 5' CAI and length, suggesting shorter genes tend to have *faster* early elongate rates (Fig. 2). Given that shorter genes have higher CAI and hence faster elongation rates, we would expect a lower ribosomal density for shorter genes contrary to the observed patterns. As a result, we conclude that higher ribosomal densities of shorter genes is most likely explained by faster initiation rates as shown by weaker 5' mRNA secondary structures.

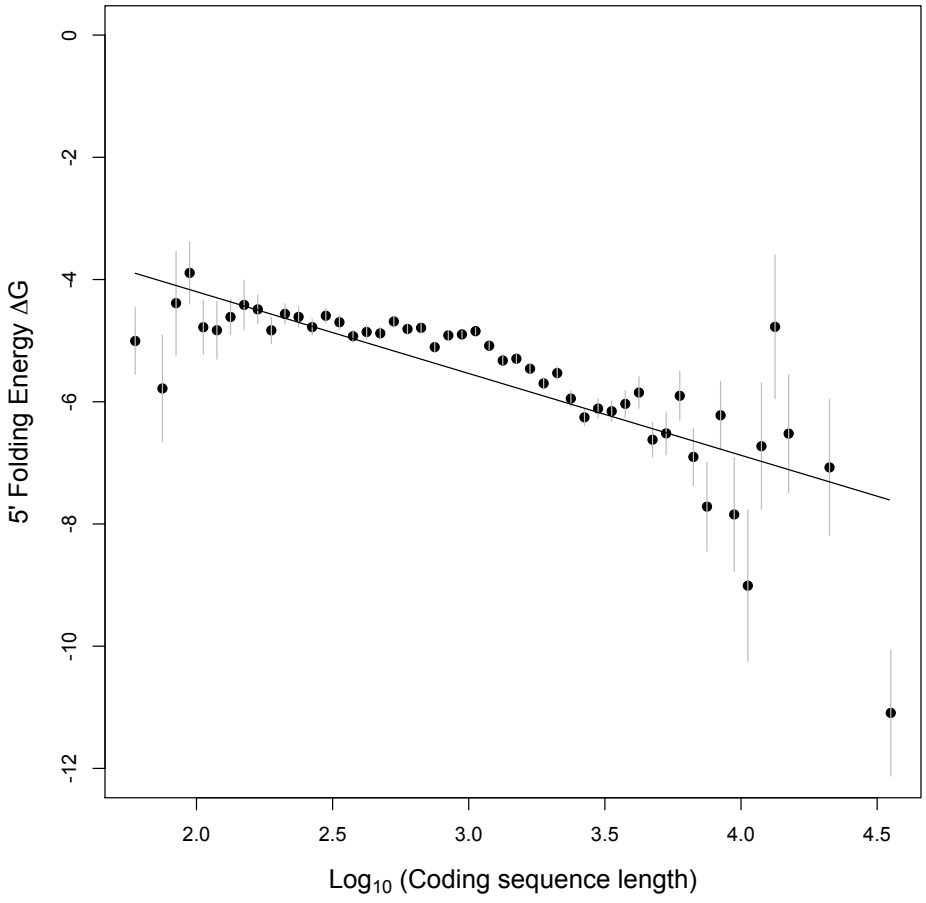


Figure 1. Short *C.elegans* genes have higher 5' mRNA folding energies than long *C. elegans* genes, suggesting faster initiation in short genes. Genes have been binned according to their log (ORF length), with dots showing the mean computed 5' mRNA folding energy in each bin, and lines showing ± 1 standard deviation. The solid line shows best-fit regression (Spearman rho = -0.12, $p < 7e\text{-}90$).

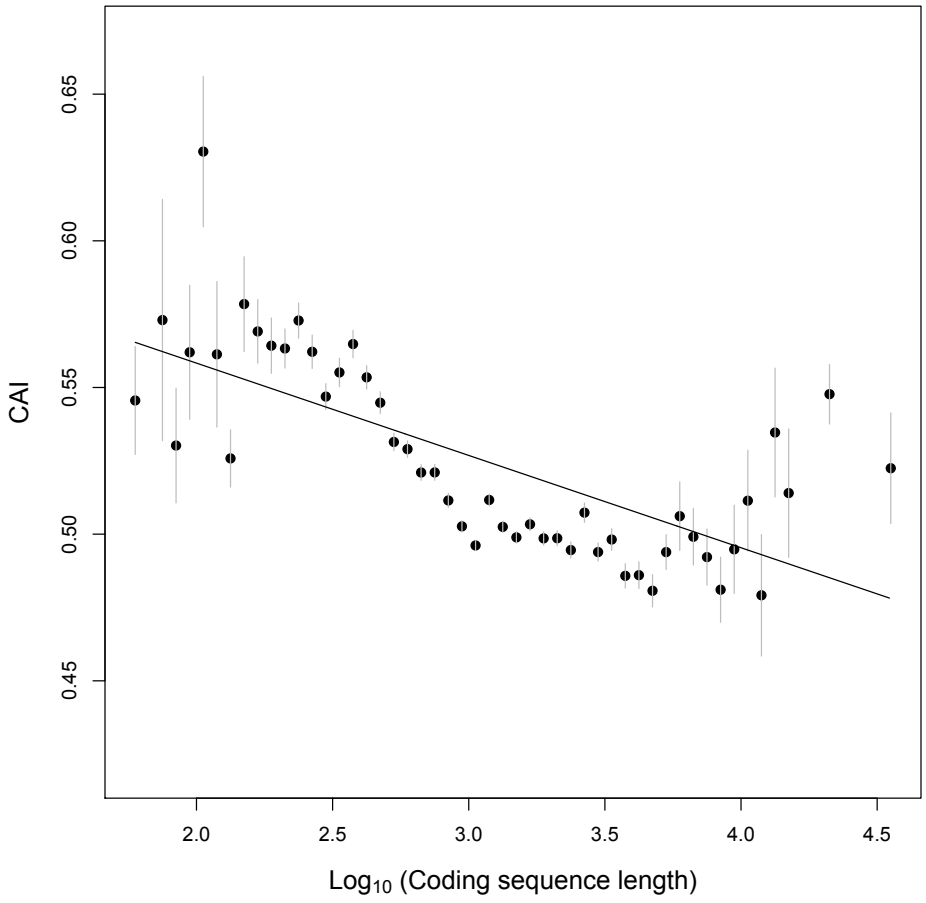


Figure 2. Short *C.elegans* genes have higher 5' CAIs than long *C. elegans* genes, suggesting faster elongation in short genes. Genes have been binned according to

their log (ORF length), with dots showing the mean computed 5' CAI in each bin, and lines showing ± 1 standard deviation. The solid line shows best-fit regression (Spearman rho = -0.16, p < 5e-179).

Codon bias, mRNA structure, and ORF length in 120 Eukaryotic Species

Given our results in *C. elegans* we then asked how broadly these trends in gene length and 5' mRNA structure hold across eukaryotes. We repeated the 5' mRNA folding energy calculations in 120 eukaryote species, and the 5' CAI calculations in 89 of those species for which a reliable reference set of genes was available for computing CAI. (The sets of species used in 5' mRNA folding energy and 5' CAI calculations are listed in supplementary table S1). The results of these calculations and their correlations with ORF length are summarized in table 1.

Table 1 summarizes the proportion of species tested that exhibit a negative rank correlation between 5' mRNA folding energy and ORF length, or between 5' CAI and ORF length. In addition we report the proportion of species that feature a significant negative correlation, at the 5% significance level. As the table shows, the results found in *C. elegans* hold very broadly across eukaryotes: about 80% of tested eukaryotes exhibit negative correlations between mRNA folding and length, and between 5' CAI and length. The preponderance of significant negative correlations with ORF length among eukaryotes is itself highly significant, for both 5' mRNA folding energy (binomial p<1e-11) and 5' CAI (binomial p<1e-9) – suggesting a systematic eukaryotic trend towards faster translation initiation and faster early elongation in short versus long genes. Thus

our results suggest that the higher ribosome density observed in shorter eukaryotes genes is likely due to faster initiation rates in shorter genes.

correlations with ORF length	5' free energy (120 species)	5' CAI (89 species)
% species with negative correlation	82%	83%
% species with significant negative correlation	73%	67%
% species with positive correlation	18%	17%
% species with significant positive correlation	11%	15%
two-sided Binomial P value	1.2 e-12	1.5 e-10

Table 1. Most eukaryotic species show a tendency towards weak 5' mRNA structure and high 5' codon bias in shorter genes. In particular, there is a negative rank correlation between 5'mRNA folding energy and ORF length 82% of the 120 eukaryotic species tested, and likewise a negative rank correlation between 5'

CAI and ORF length in 83% of the 89 species tested. The overall tendency towards negative correlations is highly significant, in both cases.

The distribution of correlations for energy and CAI are presented in Fig. 3 and Fig. 4, and the complete results for each species used in the energy and CAI calculations are presented in supplementary tables S2 and S3, respectively.

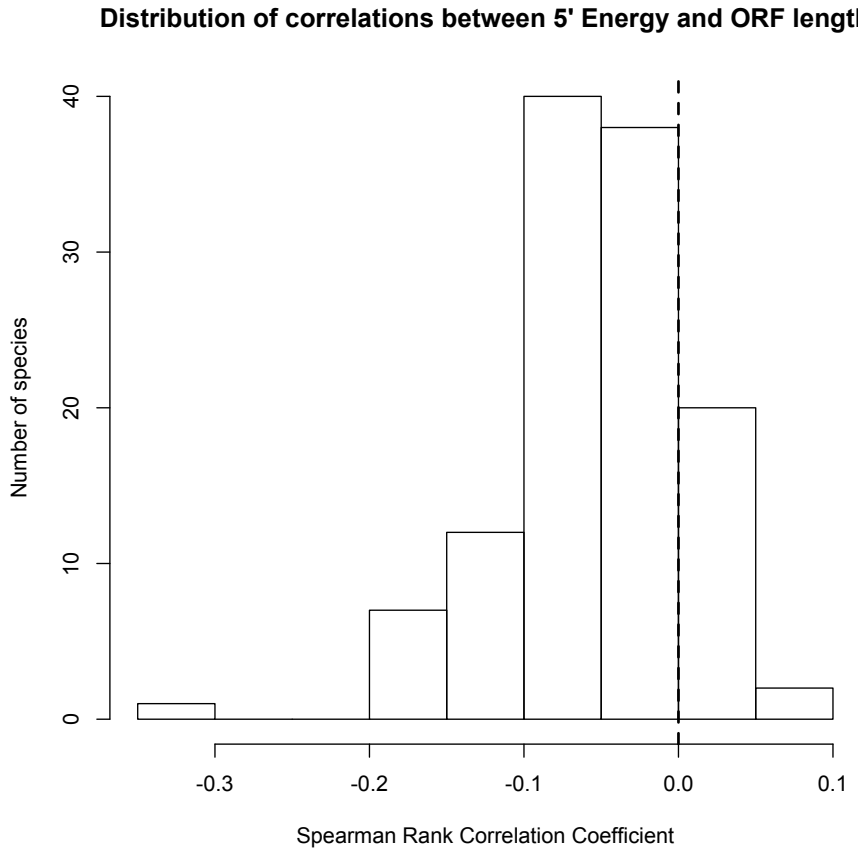


Figure 3. The distribution of Spearman Rank Correlation Coefficients between 5' Energy and ORF length in 120 eukaryotic species.

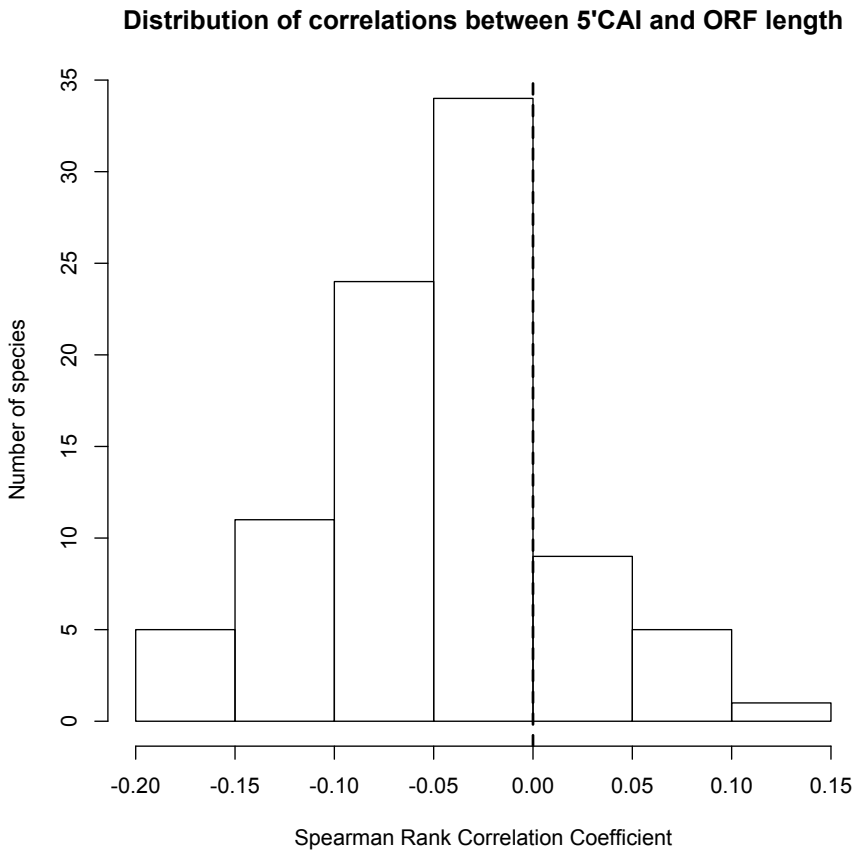


Figure 4. The distribution of Spearman Rank Correlation Coefficients between 5' CAI and ORF length in 89 eukaryotic species.

Weak 5' mRNA folding in short genes, controlling for 5' CAI

In the previous sections we have established a systematic trend towards weaker 5' mRNA structure in short genes, as opposed to long genes; and we argued that the resulting increase in initiation rates is responsible for the greater density of ribosomes

typically found in short eukaryotic genes. Nonetheless, we have also found a trend towards increased CAI in the same region, in short genes – and so the possibility remains that some subtle patterns of 5' CAI might be responsible for the trend observed in mRNA structure. To resolve this issue, we have performed a randomization procedure that isolates the effects of synonymous codons on 5' mRNA structure, controlling for 5' CAI.

For each species, we randomly shuffled the first 50 codons of each coding sequence, and we repeated this process 100 times for each gene. In each such permutation the 5' CAI of the gene is preserved, whereas the mRNA structure is possibly perturbed. We then computed the quantile of the 5' mRNA folding energy for the true gene sequence with respect to this null distribution of permuted sequences. Since our hypothesis is that shorter genes are under selection for weaker 5' mRNA folding (i.e. higher energy) regardless of 5' CAI, we expect a higher quantile for shorter genes. We tested this expectation by computing the Spearman rank correlation between the length of each ORF in the genome and the quantile of its true mRNA folding energy compared to the null distribution.

As shown in table 2, we observed a negative rank correlation between the energy quantile and the ORF length in the great majority species (binomial p-value < 6e-15) – indicating that the trend towards weak mRNA structure in short genes holds even after controlling for 5' CAI. These analyses substantiate our hypothesis that shorter eukaryotic genes are under selection to have faster translation initiation rates, achieved through weaker 5'mRNA folding.

Correlation between ORF length and quantile of observed 5’ free energy	% species (of 120 tested)
negative correlation	84%
significant negative correlation	65%
positive correlation	16%
significant positive correlation	2.5%
one-sided Binomial P value	5.38e-015

Table 2. Most species exhibit a tendency towards weak 5’ free energy in short genes, even after controlling for 5’ CAI. In the majority of species tested we find a negative rank correlation between ORF length and the quantile of the observed 5’ mRNA free energy among the free energies of permuted sequences that retain the same 5’ CAI value. The tendency towards negative correlations across species is highly significant.

Robustness of Results

In the preceding analyses we calculated 5’ CAI using the first 50 codons of each ORF. We chose this region to coincide as much as possible with the ramp of slow codons reported by Tuller et al. (Tuller et al. 2010). We repeated the 5’ CAI calculations using the first 13, 15, 20, 30, 40, and 60 codons, and obtained similar qualitative results in each

case (See supplementary table S4). The ribosomal density on a gene might be affected by codons beyond the 5' region of gene as well. For instance, slow codons in the middle or end of a gene might cause a bottleneck for ribosomes, leading to higher ribosomal densities irrespective of the codon composition in the 5' region. As a result, we also verified the robustness of our results by considering the CAI of entire ORF, producing the same qualitative, but slightly weaker, result (36% positive correlations, 64% negative correlations, two-sided Binomial P value < 0.011. For the complete tabulation of these results see supplementary table S8).

Another potential concern that may arise from our 5' CAI calculation is that we excluded sequences shorter than 51 codons. Is it possible that the sequences shorter than 51 codons could have a different CAI pattern and somehow diluted the observed CAI pattern? To answer this question we modified the definition of 5' CAI to include coding sequences shorter than 51 codons long, by computing the geometric mean of the relative adaptiveness of all the non-stop codons in the sequence. Again, this did not change our qualitative results. (See supplementary table S5).

Discussion

We have reported a strong trend towards weaker 5' mRNA structure in short genes, as compared to long genes, among eukaryotic species. Moreover, we also observed a trend towards higher 5' codon bias in short versus long genes – indicating that elongation dynamics driven by codon bias are unlikely to be the cause of higher ribosomal densities on short genes. For each individual species, the correlation between ORF length and

5’mRNA folding energy/ 5’CAI is usually statistically significant but not strong. Nonetheless, the trend of reduced 5’ secondary structure in short coding sequences was observed in the vast majority of eukaryotic species (82%) tested. The statistical significance of this trend is extraordinarily strong, and so too is the biological significance: more than three-quarters of eukaryotic species exhibit reduced 5’ mRNA structure in short genes.

To the extent that 5’ mRNA structure modulates initiation (Bettany et al. 1989; de Smit and van Duin 1990; Eyre-Walker and Bulmer 1993; Gu, Zhou, and Wilke 2010; Keller et al. 2012; Kudla et al. 2009), our results suggest that faster initiation is responsible for the empirical observation in diverse eukaryotes (Arava et al. 2003; Branco-Price et al. 2005; Cataldo, Mastrangelo, and Kleene 1999; Hendrickson et al. 2009; Lackner et al. 2007; Lacsina et al. 2011; Qin et al. 2007) that short mRNAs are more densely packed with ribosomes than long mRNAs.

Our analyses across a diverse set of eukaryotic species substantiates several authors’ interpretation of patterns of ribosomal densities and ORF length, which have been attributed to initiation-driven mechanisms as opposed to elongation effects (Arava et al. 2003; Arava et al. 2005; Lackner et al. 2007). Our results confirm that the effects of initiation, modulated by ribosomal binding to the 5’ end of mRNA and scanning to start codon, strongly outweigh those of elongation dynamics, modulated by codon bias. This view is in contrast, however, with other studies that propose a dominant role of codon usage in shaping ribosomal occupancies(Tuller et al. 2010). Our results do not directly contradict those of Tuller et al (Tuller et al. 2010), however, because those authors

considered relative codon usage within each ORF, whereas we have studied absolute codon usage across different ORFs.

Other factors such as protein folding(Kimchi-Sarfaty et al. 2007) and sequence similarity to ribosome binding sites (G. W. Li, Oh, and Weissman 2012) may also influence ribosome density. However, such effects are generally not considered as major determinants in shaping overall ribosome density(G. W. Li, Oh, and Weissman 2012; Plotkin and Kudla 2011). These factors, which are difficult to quantify systematically, are probably less likely to show systematic trends with respect to ORF length, such as those we have observed for 5'CAI and 5' mRNA secondary structure.

It is interesting to ask whether there are any commonalities among the 22 "counterexample" species in which we observed a positive rank correlation between 5'energy and ORF length. What differentiates these organisms from the other eukaryotes we have studied? To answer this question, we examined the phylogenetic relationship of all the studied species, and the distribution along this phylogeny of those 22 species exhibiting a positive rank correlation between ORF length and 5' free energy (see supplementary fig. S1). Although a few of these counter-examples are clearly closely related sister species, overall these 22 species are distributed relatively uniformly among eukaryotes, as opposed to being mostly monophyletic. And so we do not find any obvious commonality among these species with respect to their evolutionary history and, likely, ecological contexts.

Our results on systematically weaker 5' mRNA structure in short genes beg the question: why should short genes experience selection for fast translation initiation? It

has been suggested that highly expressed genes are shorter in many eukaryotes (Duret and Mouchiroud 1999; Eisenberg and Levanon 2003; Eyre-Walker 1996; Rao et al. 2010), also short genes are enriched for constitutively expressed housekeeping and ribosomal genes (Hurowitz and Brown 2003), which must produce protein as rapidly as possible. This alone might explain why short genes experience selection for faster initiation(Reuveni et al. 2011). In addition, housekeeping genes tend to have shorter 5' untranslated regions (UTRs) and are under weaker post-transcriptional regulation (David et al. 2006; Hurowitz and Brown 2003; Lin and Li 2012). The probability of successful ribosomal binding and scanning on an mRNA may depend on the length of its 5' UTRs. As a result, genes that require post-transcriptional regulation tend to have longer 5' UTRs, leading to lower initiation probabilities(Lin and Li 2012).

In summary, we find that shorter genes have higher 5’ mRNA folding energies and codon bias, suggesting that shorter genes both initiate and elongate faster than longer genes. Both of these trends hold across a diverse set of eukaryotic taxa. Since faster elongation leads to lower ribosome densities, the elevated ribosome densities of short eukaryotic genes is a result of initiation rates, rather than elongation rates.

Methods

Datasets

Coding sequences with 4bp upstream data for most species were downloaded from ensembl genomes servers (http://www.ensemblgenomes.org). The coding sequences of Y.lipolytica with 1000bp upstream sequences and 300 bp downstream sequences were

downloaded from Génolevures (Sherman et al. 2009) (www.genolevures.org/yali.html). All the coding sequences were preprocessed so that sequences whose length are not a multiple of 3, those with premature stop codons, or a continuous string of more than 3 ambiguous "N" symbols are discarded. We only considered coding sequences at least 42 nucleotides long. The complete list of species used in this study is listed in supplementary table S1.

We identified ribosomal genes for the purpose of computing CAI from one of three sources: 1. The ribosomal gene sequences for 24 species were downloaded from the HOGENOMDNA (Penel et al. 2009) database (http://pbil.univ-lyon1.fr/databases/hogenom/acceuil.php).
Orthologous groups of ribosomal genes from the HOGENOM database are listed in supplementary table S6. 2. The ribosomal genes for 64 species were obtained from Orthologous MAtrix Project (Altenhoff et al. 2011) (http://omabrowser.org). We used S. cerevisiae as our genome of reference and obtained orthologues of its ribosomal genes. The OMA orthologous groups and organism-specific ribosomal genes are listed in supplementary table S7. 3.The ribosomal genes for Y. lipolytica were obtained by performing a protein blast search against the ribosomal gene coding sequences for S. cerevisiae, and taking the top hit for each gene provided it has an E-value less than 1e-05. The number of identified ribosomal genes per species in our dataset ranged from 19 to 184 genes with a median value of 44.

Calculating 5' mRNA folding free energy

To get an estimate of the translation initiation rates, we used the program RNAfold from Vienna RNA package (Hofacker et al. 1994) to calculate the mRNA folding energy from base -4 to 37 for each gene. For each species we calculated the 5' folding energy and length of every gene, and then obtained the Spearman's rank correlation coefficient and a two-tailed p-value using the function spearmanr in the SciPy (Jones, Oliphant, and Pearu 2001) package of Python (Van Rossum and Drake 2001). We chose 0.05 as the significance level.

We then counted the number of species in which the 5' free energy has a negative Spearman's rank correlation with sequence length, and also the number of species in which the correlations are significant. We calculated a two-tailed P value to assess if there is an overall trend in the direction of rank correlation between 5' mRNA folding energy and coding sequence length.

Calculating 5' Codon Adaptation Index

To obtain an estimate of the translation early elongation rates, we calculated the Codon Adaptation Index (CAI) (Sharp and Li 1987) for the first 50 codons of each gene. The 5' CAI of a gene is defined as the geometric mean of the relative adaptiveness values of all the considered codons in a particular gene. The relative adaptiveness values of each codon is defined as ratio of occurrences of the codon to occurrences of the most abundant synonymous codon, using the ribosomal gene sequences from each species. In the above calculations, we removed coding sequences less than 51 codons long. Alternatively, for these short sequences we also calculated 5' CAI using the whole sequence, and obtained the same qualitative results (see supplementary table S5).

Chapter Three

Signatures of Natural Selection on RNA Structures in Influenza A Viruses

Abstract

Influenza A viruses cause significant human morbidity and mortality each year. Rapid evolution of antigenic surface proteins allows the virus to re-infect hosts who have recovered from prior strains. It is important therefore to understand the selective pressures that shape the evolutionary trajectories of influenza viral genomes. Most previous research has focused on identifying amino-acid residues experiencing positive or purifying selection, whereas selection on RNA structures in the negative-sense viral genome or in the positive-sense viral mRNA has received less attention. Here we develop algorithms to scan along a viral genome and identify regions that exhibit signals of purifying or diversifying selection on RNA structure. The algorithms work by computing predicted secondary RNA structures, and comparing the structural distances observed between actual viral RNA sequences against an appropriate null distribution. Unlike other algorithms that identify structural constraints by permutation of sites, our approach accounts for the phylogenetic relationships among viral sequences, as well the observed variation in amino-acid sequences. Unlike other algorithms, our approach can also detect recent selective pressures, which are of considerable practical interest in the context of viral evolution. Our analysis of viral sequence data indicates that a significant portion of influenza A viral genomes have experienced purifying selection for RNA structure, in

both the positive- and negative-sense RNA forms, especially since the divergence of human and avian strains. And we provide the first evidence of positive selection on RNA structure in specific regions of these viral genomes. We also identify genomic regions where viral RNA structures may have played a role during shifts from avian to human hosts.

Introduction

Influenza A viruses circulate widely in human hosts and cause considerable illnesses and death around the globe every year. Therefore it is of interest and importance to identify the evolutionary constraints operating on influenza A genomes, which consist of eight negative-sense RNA segments. However, most of the previous research efforts in this direction have been concentrated on identifying the amino acid residues that are under purifying or positive selection, whereas the selective pressure on the RNA structures in their negative-sense viral genome has received little attention. In this paper we will focus on this level of potential constraints on the evolutionary trajectories of influenza genomes: RNA structures.

RNA structures have been shown to play important roles during various stages of the influenza life cycle. For example, the panhandle structure (Hutchinson et al. 2010) at the two ends of each viral segment is critical for the virus to distinguish cellular RNA from viral RNAs. There has also been recent evidence suggesting RNA structures may have played a role during the host shift from avian to human (Brower-Sinning et al. 2009). This begs the question: are there other evolutionary constraints on the RNA structures in

influenza A viruses? It has been shown in other prokaryotic (Chursov, Frishman, and Shneider 2013; Gu et al. 2014), eukaryotic (Gu et al. 2014) and viral (Tuplin et al. 2002; Tuplin, Evans, and Simmonds 2004; Zanini and Neher 2013) species that evolution tends to preserve RNA structures. Does this also hold for influenza A viral RNA? And, conversely, are there any regions in influenza virus genomes or mRNA that are under selective pressures to change their RNA structures?

Several prior studies in influenza viruses have explored conserved RNA structures and their functions in influenza genomes (Moss, Priore, and Turner 2011; Priore, Moss, and Turner 2012; Moss et al. 2012; Priore, Moss, and Turner 2013; Priore et al. 2013; Dela-Moss, Moss, and Turner 2014; Jiang et al. 2014; Gultyaev et al. 2014). Our research is different from these studies in two major aspects: First, for one portion of our analysis, we leverage the influenza sequences that have been accumulated during the past half a century, and we base our estimates of selection pressures only on those substitutions that have accrued during this time period. As a result, our analysis quantifies the extent of recent selection pressures on viral RNA structures. Second, in addition of purifying selection on RNA structures, our approach can identify diversifying selective as well. And indeed we find evidence to support the view that some genomic regions have experienced selection to change RNA structures, especially during the transition from avian to human hosts.

Results

Adaptation from avian hosts to human hosts

We first want to investigate whether the influenza A genomes have been under selective pressure for conserved or variable RNA structures during the shift from avian to human hosts. To answer this question, we developed a simple algorithm to detect selection on RNA structures by pairwise sequence analyses, applied to two distinct sets of influenza genomic sequences: those collected from human and avian hosts, respectively. The basic idea behind the method that compares two sequences is to scan along the positive- or negative-sense viral RNA in a moving window, compute predicted secondary structures for an avian and human form of the viral sequence, in a given window, compute a structural distance metric to quantify the structural divergence between the human and avian RNA forms, and finally compare this observed structural distance to an appropriate null distribution based on the randomly permuting the positions at which the two sequences differ synonymously, whilst preserving the total number of synonymous mutations between the two sequences.

This algorithm for detecting selection is described in detail in the Method. But, briefly, the algorithm scans along each viral segment with a sliding window of 60 bases and step size of 9 bases. In order to calculate the actual pairwise RNA structural distances for a particular window, we selected non-redundant datasets of human and avian influenza sequences, and we randomly paired each human influenza viral sequence with an avian influenza viral sequence. We computed the average RNA structural distance in each window among all such pairs, to be compared to a null distribution. The null distribution was generated by starting with the avian influenza sequence of each human-avian pair (avian influenza evolves more slowly, so they were considered “ancestral

sequences", but the result is robust with respect to choice of "ancestral genome"), and introducing the observed number of synonymous mutations in random positions, while preserving all non-synonymous mutations, so that the resulting "simulated" human influenza sequences encode the same amino acid sequences as the "original" human influenza sequences. We then calculated the RNA structural distance between the resulting "simulated" human influenza sequence with the original influenza sequence, and the average from all such pairs contributes value in the null distribution. Finally, by comparing the average RNA structural distance in the true data with this null distribution we derived a quantile score for each window of the genome. If the average pairwise structural distance among the actual pairs of human-avian viral sequences is smaller than most the distances in the null distribution, then this suggests that purifying selection has constrained synonymous mutations along the phylogeny to preserve RNA structures. On the other hand, if the average structural distance among the actual viral sequences is larger than most of the distances in the null distribution, then this suggests there has been diversifying selection to change RNA structures in the recent past, presumably adapting to some novel environmental or genetic context after the host shift from avian to human hosts.

We first studied the overall distribution of quantile scores for all windows in all eight segments of human vs. avian influenza A viral genomes (Figure 1). Our dataset of human and avian influenza A viruses contained 2,699 non-redundant sequences (meaning the pairwise sequence similarity is below a specified threshold; see the description of the pairwise algorithm in Materials and Methods) selected from a total of 81,997 viral

segments that were isolated from infected hosts between the years 1918 and 2013. In the absence of any selective pressure, we would expect the quantile scores to be distributed uniformly between 0 and 1. We found the observed distribution is bi-modal: we observed both an excessive number of windows with low quantiles, suggesting purifying selection, as well as an excessive number of windows with high quantiles, suggesting diversifying selection. There have been reports in some prokaryotic, eukaryotic and other viral species (Chursov, Frishman, and Shneider 2013; Zanini and Neher 2013; Gu et al. 2014) that a significant proportion of the genome is under purifying selection to preserve RNA structures, and our analysis confirms these findings. Moreover, our results provide evidence that some portions of influenza genomes are under diversifying selective to change their RNA structures. These analyses provide a global view of recent diversifying and purifying selective pressures on RNA structures that influenced the synonymous substitution accrued in influenza A genomes.

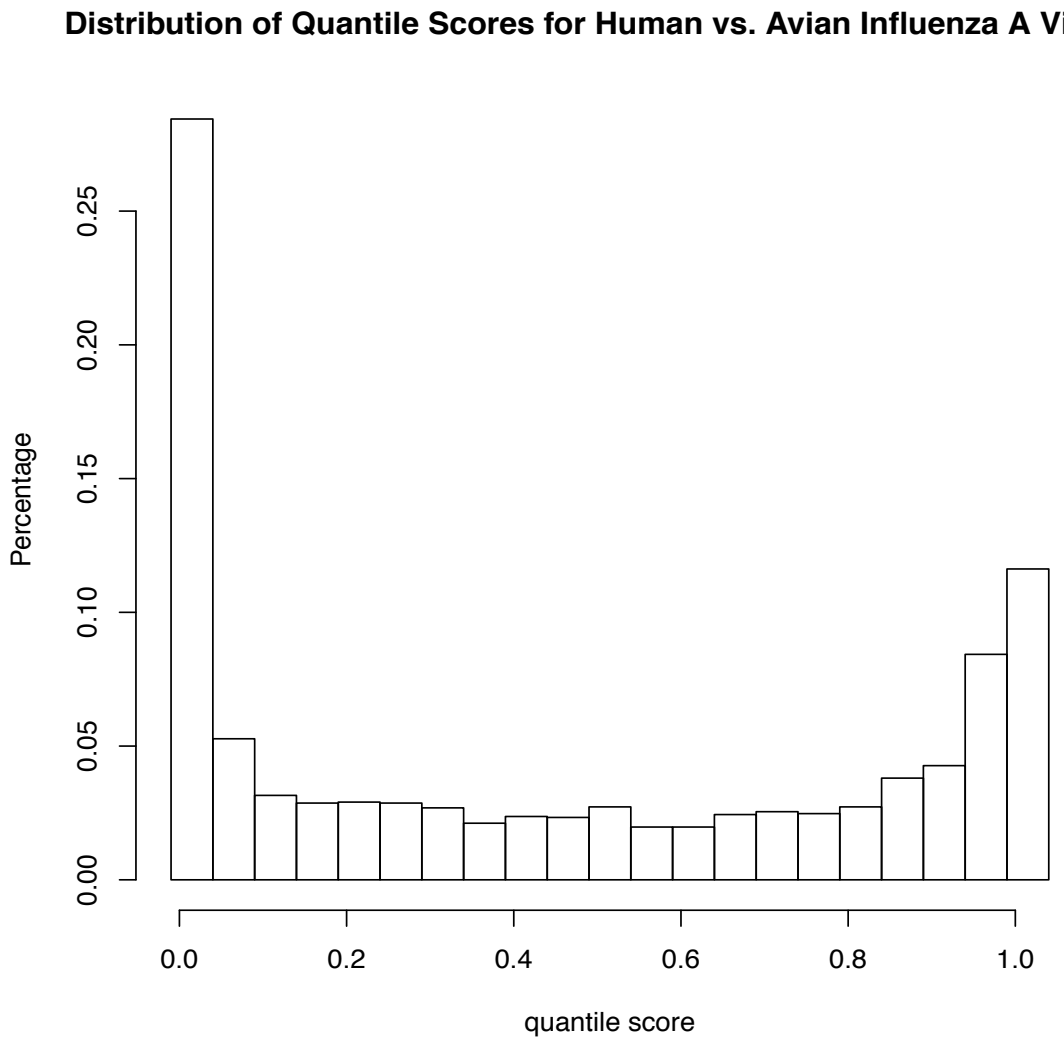


Figure 1: Distribution of quantile scores from all the 8 segments of Human vs. Avian influenza A viruses, under pairwise analysis. Both purifying and diversifying selective pressure are observed in avian- vs. human-derived influenza viral samples, as evidenced by enrichment for low and high quantiles of the observed test statistic compared to the null distribution (two-sided p-value of Kolmogorov-Smirnov test against uniform distribution<2.2 e-16).

Next we examined the results from each viral segment in more detail. As an example, we looked at the predictions for a particular Influenza A segment: the PA segment (Figure 2). Our algorithm predicts regions 244-313, 883-952, 1738-1798, 1963-2041 on the (+) strand, and regions 1-70, 226-286, 334-394, 496-565, 892-961,1108-1177, 1639-1717 on the (-) strand show signs of purifying selection on RNA secondary structures (quantile scores less than 0.05). In a previous study (Moss, Priore, and Turner 2011) that investigated the conserved RNA secondary structures in influenza A coding regions, authors compared six genome sequences from H5N1, H1N1 subtypes, isolated from human, swine and avian hosts. They predicted positions 1611-1860, 1941-2120 on the (+) strand, and positions 41-290,1161-1280 with ambiguous strand bias, as conserved regions for RNA structures. Our results for segment 3 are thus in qualitative agreement with those of Moss et al (2011). In addition to this, our algorithm has the unique ability to detect genomic regions that exhibit signs of positive selection for RNA secondary structural change, and indeed the algorithm predicts positions 100-205, 523-583, 1036-1114 on the (+) strand, as well as positions 118-187, 568-628, 1009-1078, 1585-1645, 2026-2086 on the (-) strand of the PA segment that may be under selective pressure to change RNA secondary structures (quantiles exceeding 0.95).

We summarized the quantile scores for each segment of Human vs. Avian influenza A viruses in Figure 3. We found the same general conclusions hold for other 7 segments as well – we observed many regions that are likely to be under purifying and diversifying selective pressure in other segments, especially in segments 3 (PA), 5 (NP), 7 (M1/M2), 8 (NS1/NS2). This largely agrees with several previous reports that found

segment 5 (Gultyaev et al. 2014), segment 7(Moss, Priore, and Turner 2011; Moss et al. 2012) and segment 8 (Plotch and Krug 1986; Nemeroff et al. 1992; Moss, Priore, and Turner 2011) to be potentially enriched for conserved RNA structures, although the exact location of the preserved secondary structures sometimes differ. Unique to this study, we also observed that segments 3,5,7,8 also harbor regions under diversifying selective pressures on RNA structure. This not only suggests that indeed RNA secondary structures in these segments are functionally important, it also suggests that these four viral segments may have played important roles in influenza A viral host shift from avian to human hosts, such as structural changes to accommodate different host body temperatures.

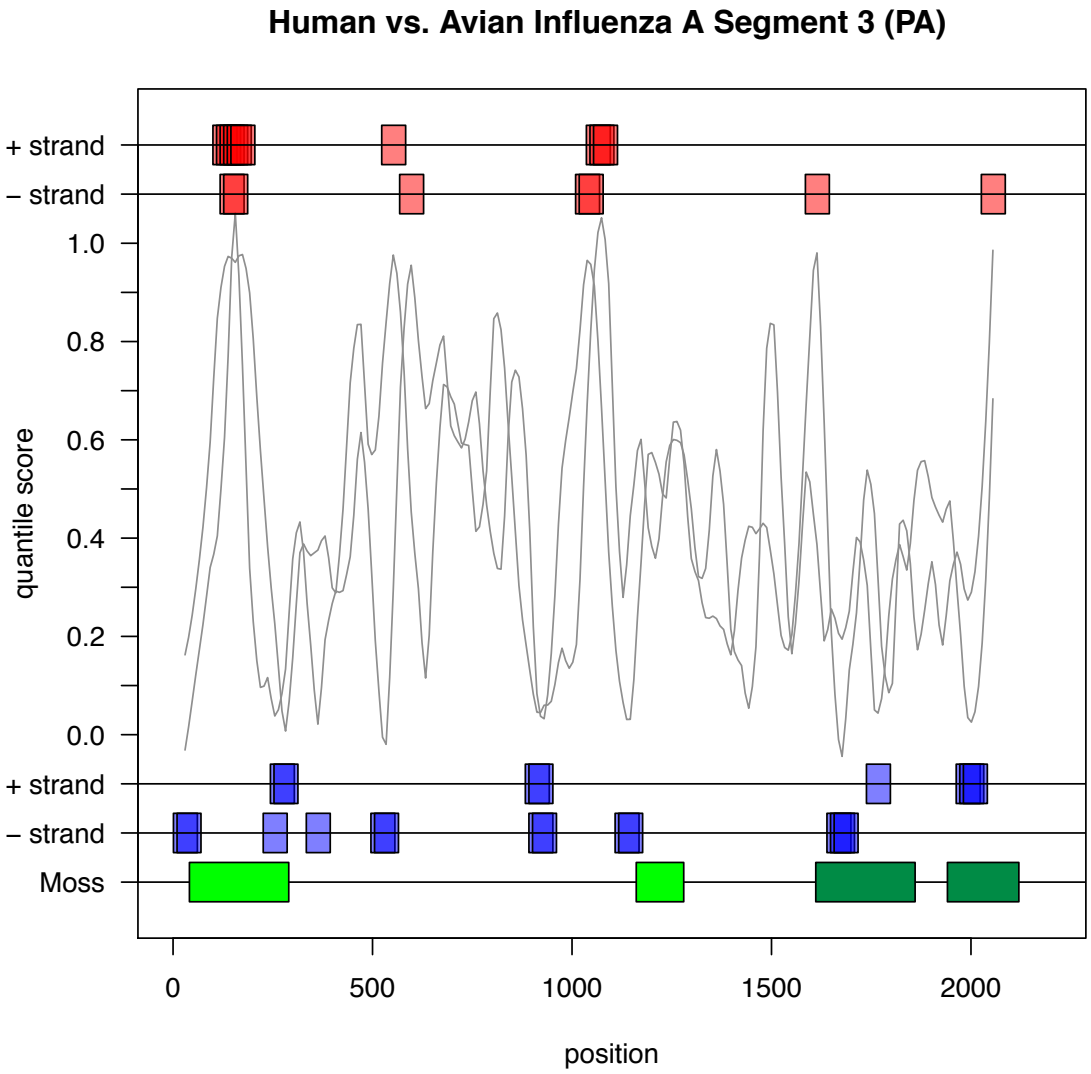


Figure 2: Quantile scores from Pairwise Selection Detection Algorithm for influenza A segment 3 (PA) using avian- and human-derived influenza viral samples. The two grey curves are the predicted position-specific structural quantile scores for two strands, smoothed by Local Polynomial Regression Fitting. The blue bars represent regions where there is evidence for purifying selection for RNA secondary structures; whereas red bars represent regions there is evidence for diversifying selection. We chose a quantile cut-off

score of 0.05 as the threshold for purifying selection, and a quantile cut-off of 0.95 as the threshold for diversifying selection. For comparison, the green bars represent the regions predicted to have conserved RNA structures by (Moss, Priore, and Turner 2011), where dark green indicates the predicted structure is on the (+) strand, while light green indicates regions with ambiguous strand biases.

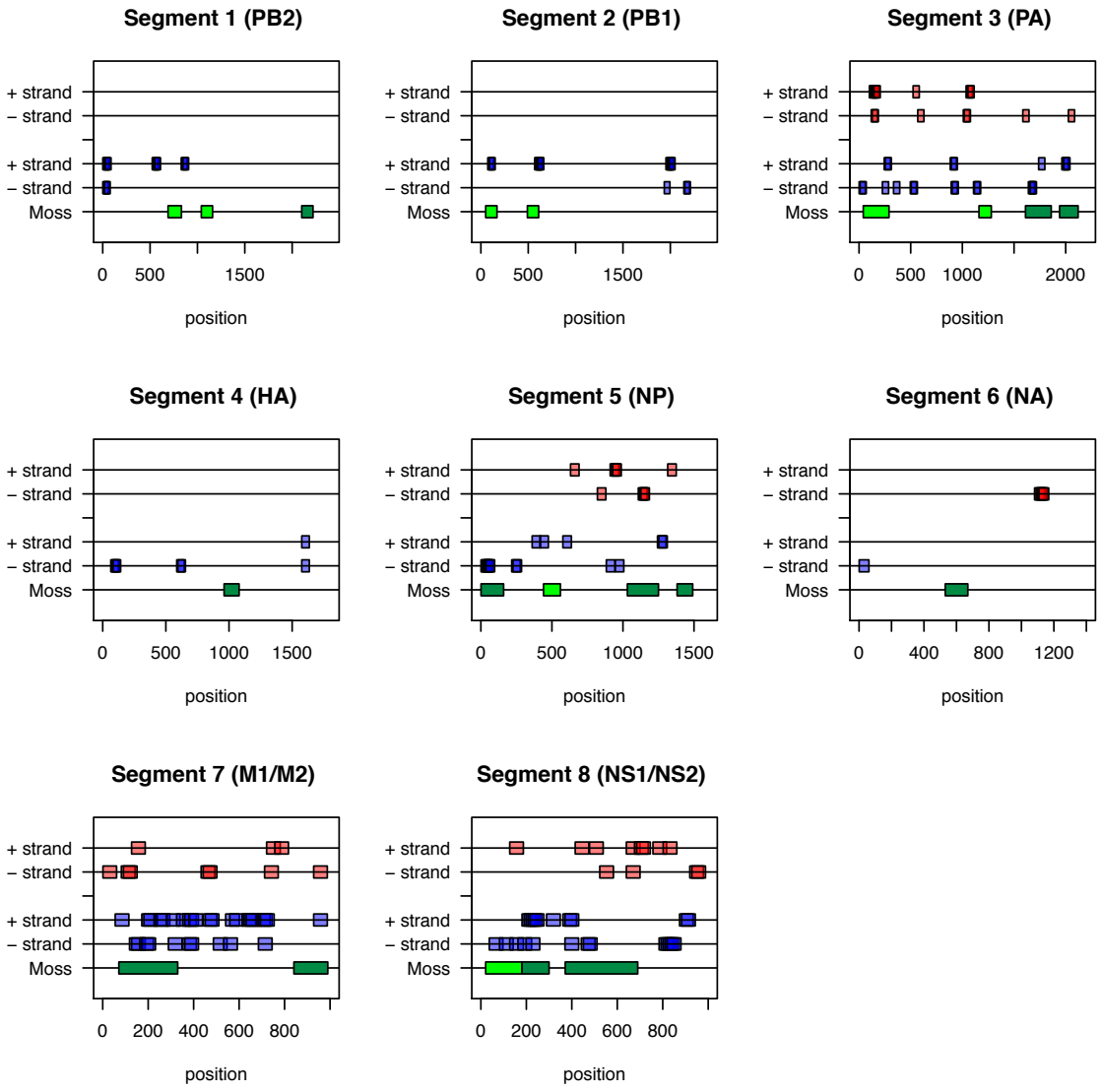


Figure 3: Quantile scores from Pairwise Selection Detection Algorithm for all

segments using avian- and human-derived influenza A viral samples. The blue bars represent regions where there is evidence for purifying selection for RNA secondary structures; whereas red bars represent regions there is evidence for diversifying selection. We chose a cut-off score of 0.05 as the threshold for purifying selection, and a cut-off of 0.95 as the threshold for diversifying selection. For comparison, the green bars represent the regions predicted to have conserved RNA structures by (Moss, Priore, and Turner 2011), where dark green indicates the predicted structure is on the (+) strand, while light green indicates regions with ambiguous strand biases.

A previous report (Brower-Sinning et al. 2009) found a clear separation in the RNA folding energies between human and avian influenza polymerase genes, and further suggested that the body temperature differences between human and avian hosts may cause RNA structures to fold differently in the two hosts, thus creating selective pressures for the proper folding of RNAs. In our reported results so far, all the viral RNA sequences were folded at human body temperature (37 °C). We repeated the above analysis for three polymerase genes (PB1, PB2, PA) except this time folding the avian influenza viruses at avian body temperature (40 °C), and the results are qualitatively unchanged.

Detecting recent selection on Influenza A viral RNA structures within a single host species

In the previous section we developed a pairwise selection detection algorithm to

compare human vs. avian influenza viruses and detected genomic regions exhibit signs of purifying or diversifying selection for RNA structures. In this section, we develop a phylogeny-controlled selection detection algorithm and use it to detect RNA structural selective pressures that have operated in the recent past – that is over the timescale of decades, while viruses have remained within a single (human) host species. Similar to the pairwise selection algorithm, the algorithm scans along each viral segment with a sliding window of 60 bases and step size of 9 bases. For each window we computed a quantile score by comparing the average pairwise RNA structural distances between actual influenza isolates against a null distribution. The null distribution preserves (i) the amino acid sequence of each viral isolate (ii) the phylogenetic topology relating all the viral isolates and (iii) the number of substitutions along each branch in the phylogeny, but otherwise randomizes the locations of the synonymous mutations (see Materials and Methods for details).

We ran the phylogeny-controlled algorithm for all the eight segments of human H1N1 influenza viruses. Our dataset of human H1N1 viruses contained sequences from 54,465 viral segments, sampled between the years 1918 and 2013. The distribution of the quantile scores from all the windows of the 8 segments is shown in Figure 4. Our results are qualitatively different from the results we found when comparing human vs. avian influenza A viruses: We only see an enrichment of genomic regions on the side of purifying selection; while the side indicating diversifying selection is conspicuously flat.

Since the approach used in this section leverages the information from thousands of sequence polymorphisms that existed in the past few decades, our analysis confirms

the purifying selective pressures on RNA structures have persisted in the recent past (that is, on substitutions accrued over the past 50 years). On the other hand, it is not surprising that we do not find as many regions that are under diversifying selection as in the human vs. avian comparison, since presumably the diversifying selective pressures on RNA structures to adapt to novel environment should be much milder, if exists, during the viral evolution inside a single host species. We constructed the same histogram of quantile scores for human H3N2 viruses (34,393 viral segments, sampled between the years 1968 and 2013) and observed qualitatively similar results as for H1N1 human viruses (see Supplementary Figure 1).

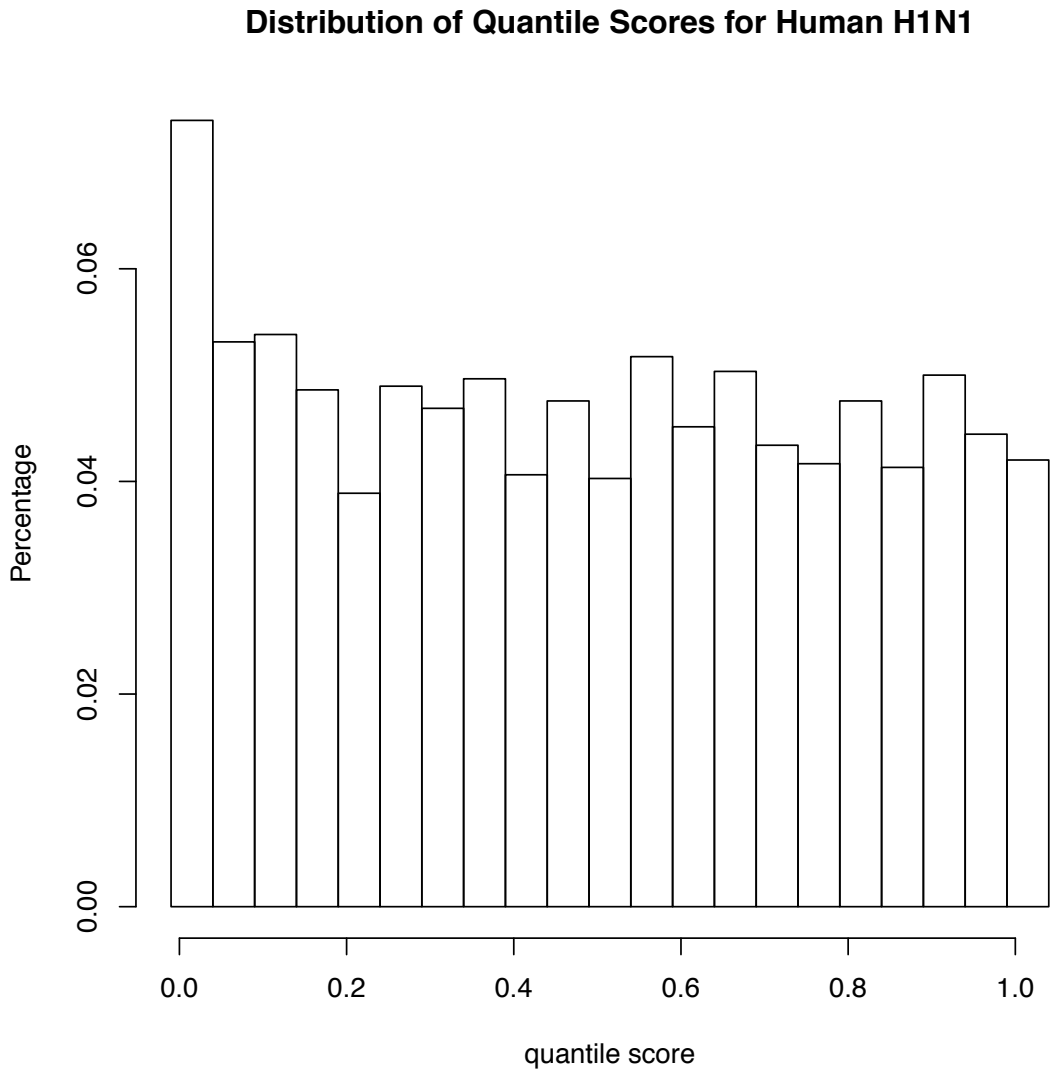
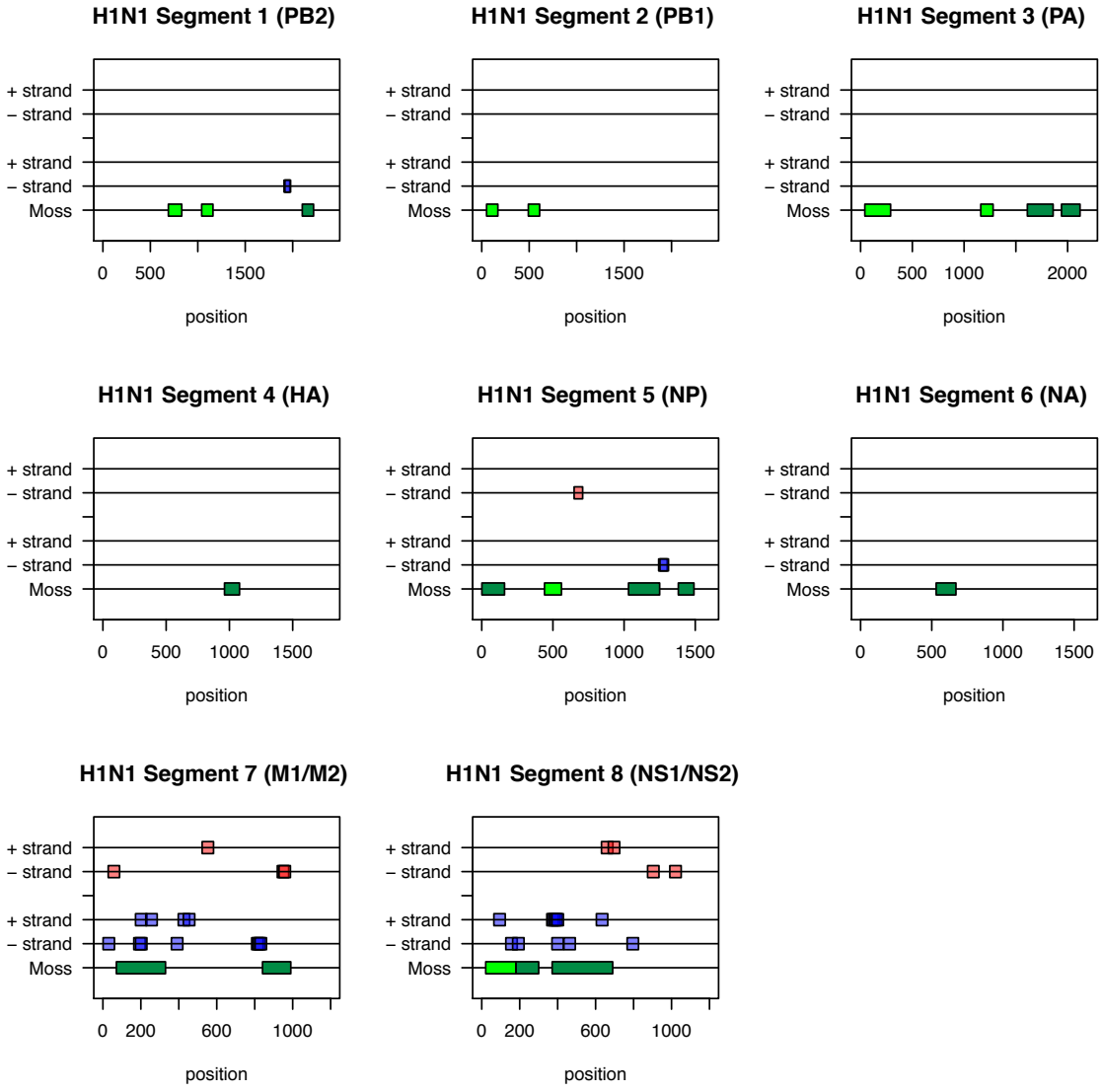


Figure 4: Distribution of quantile scores from all the 8 segments of Human H1N1 influenza A viruses. Only purifying selective pressure are observed within Human H1N1 influenza viral samples, as evidenced by enrichment for low quantiles of the observed test statistic compared to the null distribution (two-sided p-value of Kolmogorov-Smirnov test against uniform distribution<2.2 e-16).

In Figure 5 we gave a survey of analysis of selection on RNA structures for each human H1N1 and H3N2 viral segment. As the figure indicates, we identify many fewer regions that show signs of selection on RNA structures, except for segment 7(M1/M2) and 8(S1/S2). This is consistent with previous reports (Moss, Priore, and Turner 2011) suggesting that segment 7 and 8 are most enriched for functional RNA structures. Our results suggest purifying selective pressures on RNA secondary structures likely had some effect on the shape of the evolutionary trajectory of influenza A genome in the past half a century, within human hosts alone. This highlights the importance of detecting the selective pressure "in action", which is what our phylogeny-controlled algorithm is specifically designed for. The fact that we identify many fewer genomic regions under selection is likely because we only used influenza sequences from a single subtype within a single host species collected in the past few decades, there is less sequence variation to provide power for detection.


H1N1 Segment 1 (PB2)
+ strand
– strand
+ strand
– strand
Moss
0 500 1500
position
H1N1 Segment 2 (PB1)
+ strand
– strand
+ strand
– strand
Moss
0 500 1500
position
H1N1 Segment 3 (PA)
+ strand
– strand
+ strand
– strand
Moss
0 500 1000 2000
position
H1N1 Segment 4 (HA)
+ strand
– strand
+ strand
– strand
Moss
0 500 1000 1500
position
H1N1 Segment 5 (NP)
+ strand
– strand
+ strand
– strand
Moss
0 500 1000 1500
position
H1N1 Segment 6 (NA)
+ strand
– strand
+ strand
– strand
Moss
0 500 1000 1500
position
H1N1 Segment 7 (M1/M2)
+ strand
– strand
+ strand
– strand
Moss
0 200 600 1000
position
H1N1 Segment 8 (NS1/NS2)
+ strand
– strand
+ strand
– strand
Moss
0 200 600 1000
position

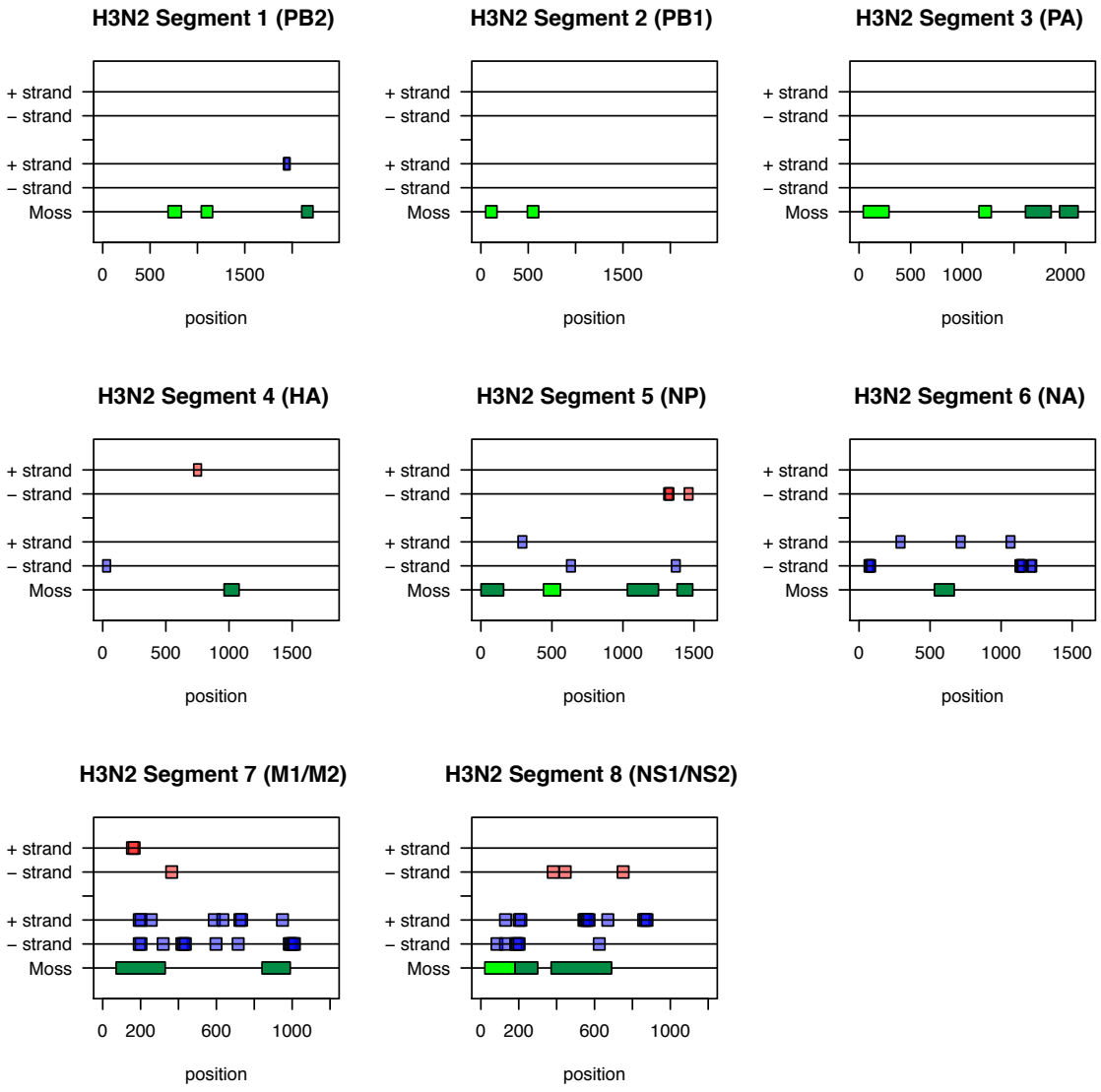


Figure 5: Quantile scores for all segments of influenza A H1N1 and H3N2 subtypes. The blue bars represent regions where there is evidence for purifying selection for RNA secondary structures; whereas red bars represent regions there is evidence for diversifying selection. We chose a cut-off score of 0.05 as the threshold for purifying selection, and a cut-off of 0.95 as the threshold for diversifying selection. For comparison, the green bars represent the regions predicted to have conserved RNA structures by

(Moss, Priore, and Turner 2011), where dark green indicates the predicted structure is on the (+) strand, while light green indicates regions with ambiguous strand biases.

No coordination in selection for RNA structures and for amino acid residues

Numerous studies on Influenza viruses (Bush et al. 1999; Yang 2000; Suzuki 2006; Kosakovsky Pond et al. 2008; X. Ding et al. 2010; Tusche, Steinbruck, and McHardy 2012) have analyzed amino-acid residues or collections of residues that may experience selective pressures. It is thus of interest to ask whether selection on the level of RNA structure which we identified in the previous sections are coordinated with selection on the level of amino acid, i.e. if they both act as diversifying or purifying selective forces on the same genomic regions. To answer this question, we calculated the dN/dS ratio at each individual site within each influenza A segment using the standard counting approach (Nei and Gojobori 1986), adapted for an entire phylogenetic tree instead of two sequences (See Materials and Methods for more details). Figure 5 displays the result for all 8 segments of H1N1 influenza viral sequences from all hosts, where the dN/dS ratio for each amino acid residue is shown in conjunction with the quantile score of RNA structural selection as we calculated previously. As the figure shows, there is generally a lack of coordination between the amino acid and RNA structural levels of selective pressures, suggesting the two levels are under selection for independent selective forces. This is perhaps to be expected as we hypothesize that the selection on RNA structures is likely due to the requirements for genome packaging (Hutchinson et al. 2010) and potentially other regulatory functions, whereas the amino acid level selection

may act in various stages of the virus life cycle, including for efficient replication, transmission, virulence, host adaptation etc. In any case, this result helps to confirm that our RNA structural selection detection algorithm has controlled for potential amino acid selection effects.

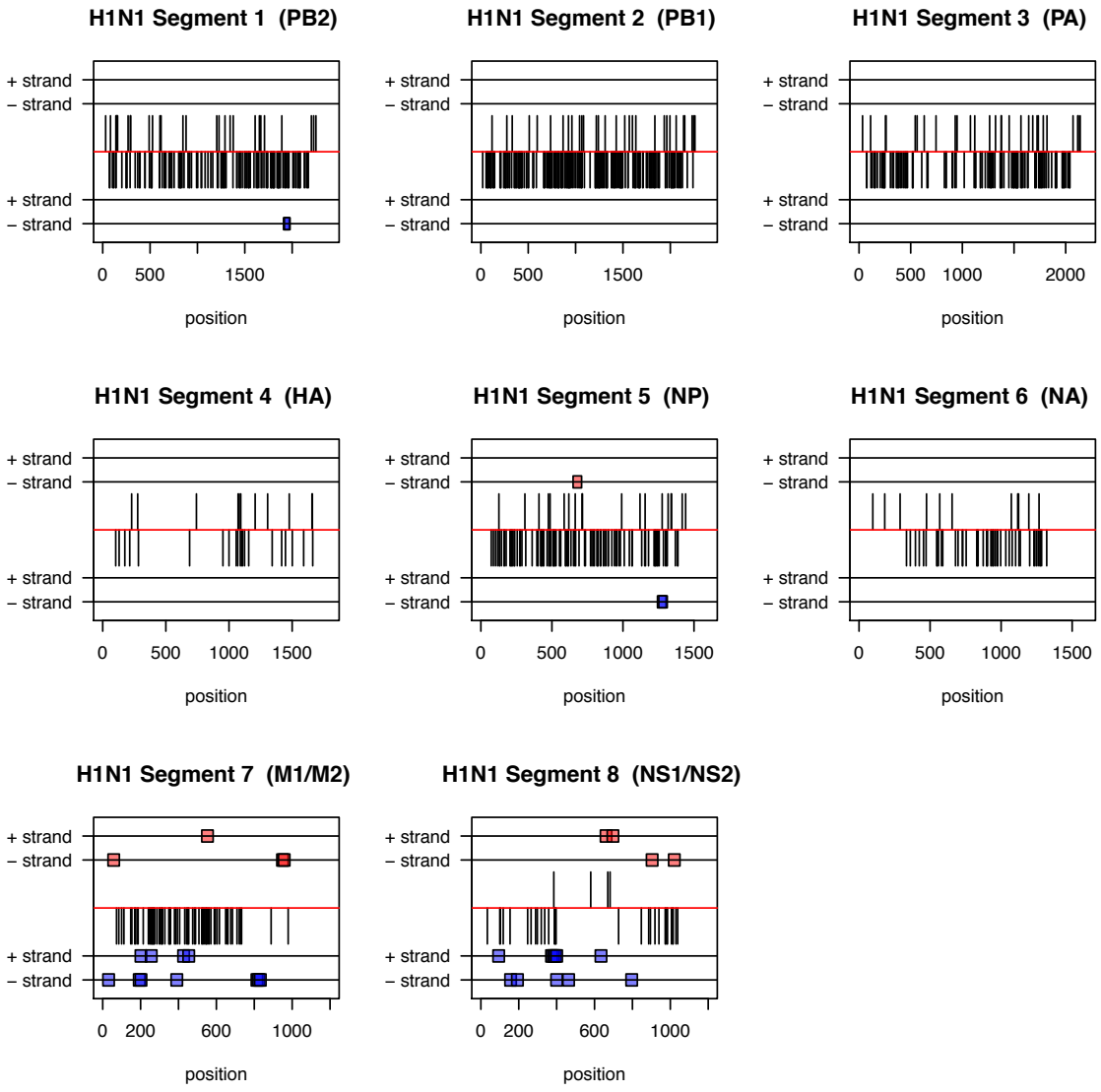


Figure 6: Quantile scores vs. dN/dS values for all segments of the human H1N1

subtype. The blue bars represent regions where there is evidence for purifying selection for RNA secondary structures; whereas red bars represent regions there is evidence for diversifying selection. We chose a cut-off score of 0.05 as the threshold for purifying selection, and a cut-off of 0.95 as the threshold for diversifying selection. Positions with extreme high dN/dS values (>4) are represented with vertical bars above the red line, while positions with extreme low dN/dS values (<0.00001) are represented with vertical bars below the red line.

Robustness of results

In the last four sections we reported regions of influenza A genomes are selective pressures for RNA structures, at two different timescales using two different, but closely related, algorithms. These algorithms required several arbitrary choices, and here we query: how robust are our results with respect to different parameters and choices made. To address this question, we investigated the robustness of our results with respect to two most critical aspects in our algorithms: the choice of RNA secondary structural distance metric, and window size for the moving-window analysis.

There are two widely used RNA distance metrics: base-pair distance and tree-edit distance (Shapiro 1988; Shapiro and Zhang 1990; Fontana et al. 1993; Hofacker et al. 1994). The base-pair distance reflects the naïve notion of Hamming distance between two dot-bracket representations of RNA structures. This metric suffers from the defect of not considering the interactions between base pairs, as well as any shift in the RNA structures. Tree-edit distance, on the other hand, is a metric of RNA structural distances based on

graph theory, in which the secondary structures are modeled as tree graphs, and the distance between two secondary structures is defined as the number of operations needed to transform one tree into another. In addition to the full-structure tree-edit distance we used in the previous sections, we repeated our analysis to compare human vs. avian influenza A viruses using base-pair distance and a coarse-grain (HIT) tree-edit distance. We calculated the Spearman rank correlation coefficients of the per-window quantile scores between full-structure tree-edit distance and each of these two alternatives for the NP segment. Both correlations are highly significant (full-structure vs. HIT: rho=0.81, p-value<2.2e-16, full-structure vs. base pair distance: rho=0.54, p-value<2.2e-16), which indicate our results are generally robust with respect to different RNA structural distance metrics.

Next we investigated the robustness of our results with respect to different window sizes. For the considerations of genome scan resolution and computational time, our default window size was 60 bases. Since limiting the shifting window size can potentially omit base pairs that happen across a long distance, we repeated our experiment to compare human vs. avian influenza A viruses, scanned along the NP segment of influenza genomes with our algorithm using a window size of 120 bases, and computed the Spearman rank correlation coefficient between corresponding quantile scores for two window sizes (there were a few more windows if using 60 bases as the window size and they were discarded for comparison). The high correlation (rho=0.18, p-value=0.00099) between results for two different window sizes confirms the robustness of our results with respect to different window sizes.

Discussion

In this study we developed two algorithms to systematically detect selective pressures on RNA structures in genomic sequences, and applied it to influenza A viral genomes. We identified candidate regions in influenza A genomes that are under purifying or diversifying selection for RNA structures. The results are robust with respect to various parameters including different distance metrics, window size etc. Our analyses suggest a significant proportion of the influenza A genomes are under selection for RNA structures, and the selective pressures for RNA structures are still operating in recent times. And we also found evidence of diversifying selective pressures for RNA structures especially when comparing human and avian influenza A viruses, suggesting they may have played important roles during the viral host shifts.

For the task of predicting consensus secondary structure using a multiple sequence alignment for noncoding RNA sequences, RNA folding algorithms based on a single sequence are generally considered not as accurate as algorithms that leverage the information of all the aligned sequences, such as Dynalign (Mathews and Turner 2002) and RNAalifold (Hofacker, Fekete, and Stadler 2002; Bernhart et al. 2008). However, there are several problems with alignment-based approaches if they were to be applied to influenza viral sequences. First, one of the major novelties of this work is that we discovered candidate regions that are under diversifying selection. This finding cannot be achieved, even in principle, by looking for the consensus base-pairing patterns in the aligned RNA sequences. Another major feature of our algorithms is that compared with alignment-based algorithms we can study relatively recent selective pressures, based on

mutations accrued over the past 50 years alone within a single host. The basic rational behind previous approaches such as (Moss, Priore, and Turner 2011) is the idea that the actual RNA sequence alignment should be structurally more stable than the permuted sequences, and so they mostly detect sequence features that have are fixed in all influenza subtypes. Lastly, in our null distributions, all the simulated sequences respect the same amino acid sequence and phylogenetic relationships as the actual sequences. One of the major signals used in other, alignment-based approaches is to compare the energy profile of the real influenza sequence alignment with the dinucleotide/dicodon shuffled sequence alignment. This signal is very effective in detecting functional noncoding RNA sequences, but is not well justified when applied to coding RNA sequences because such shuffling will disrupt the amino acid sequences, which are arguably is the strongest objects of selective pressure on the influenza genomes. Even disregarding the problems that such algorithms have suffer by not respecting amino-acid sequences, it not clear whether functional RNAs in the protein coding regions should have lower energy than randomly-shuffled RNAs (Clote et al. 2005). Finally, in the alignment-based approach every sequence is treated as of equally distance from each other phylogenetically, which assuredly false for the influenza viruses we and earlier authors have analyzed.

Despite some advantages, our approach is surely not without limitations. First, there are three influenza viral segments: PB1, M1/M2, NS1/NS2 that have a small portion of overlapping open reading frames (ORFs). Since our algorithm compares the average pairwise structural distance of the real influenza sequences with a null distribution where the simulated sequences have the same synonymous and non-

synonymous distance, our algorithm does not produce results for these regions, as there is typically no distinction between synonymous and non-synonymous mutations in those regions. Second, since our algorithm use the commonly used software RNAfold to predict RNA secondary structures, and then uses RNAdistance to calculate pairwise RNA structural distances, our algorithm shares the same limitations and these structural algorithms: we cannot predict or use information about pseudo-knots. There are some structure-predictions algorithms that can potentially predict the existence of pseudo-knotted RNA structures, but there are no well-accepted distance metrics to compare two pseudo-knotted structures, and so we have used the more traditional and widely used RNAfold approach.

Materials and Methods

Datasets and Preprocessing

The influenza genome datasets were downloaded from NCBI influenza virus resource (cite, http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html, last accessed: 08/21/2013). Because the noncoding regions for each influenza segment consists of a tiny proportion of the genome in length, and they were only occasionally sequenced and collected, in this study we focus on protein-coding regions.

We first translate all the influenza nucleotide sequences into amino acid sequences, then align the amino acid sequences using multiple sequence alignment package MUSCLE version 3.8 (Edgar 2004), the aligned amino acid sequences are then

back-translated into nucleotide sequences. For all the gaps in the alignment, we filled the gapped position with the 50% consensus nucleotide (for segment 4 and 6 the threshold is 30%) in that position, and if there is no single nucleotide that appears more than 50% of the time in the particular position, we remove the sequences with this gap from the collection. For segment 7 and 8, where the segments have two overlapping open reading frames, we concatenate the two pieces into a single sequence. After this step, we have the necessary data for the pairwise selection detection algorithm. For the phylogeny-controlled algorithm, we then reconstructed the phylogeny of the influenza sequences using RAXML 7.3.0 (Stamatakis 2006) with model GTRGAMMA, and inferred the internal and root sequences using PAUP* version 4.0 (Swofford 2003). So for each viral segment, we inferred a phylogenetic tree and all the internal and root sequences. And together with the actual observed influenza sequences as leaves of the phylogenetic tree, for each node of the phylogeny we have their nucleotide sequence. We inferred the phylogeny for each segment separately because of the influenza viral re-assortment events, in which different strains of influenza viruses exchange their viral segments. Now we have all the necessary data for the phylogeny-controlled algorithm.

Description of the pairwise algorithm to detect selection

In this algorithm, we are given two distinct sets of influenza sequences, and want to ask if there are diversifying or purifying selective pressures on RNA structures between the two groups. The pairwise selection detection algorithm (Figure 7 is an illustration of the pipeline of this algorithm.) works by scanning along each viral segment

in a window size of 60 bases and step size 9 bases. Since in this algorithm the phylogenetic relationships between the viral samples were not explicitly accounted, we selected a non-redundant dataset of ~ 150 sequences from each host. Then we randomly pair the same segment from a human influenza viral sequence with an avian influenza viral sequence. For each window, we first computationally fold the RNA sequences using the program RNAfold from Vienna RNA package (Hofacker et al. 1994). We then calculate the average pairwise structural distances among each pair of these sequences using the program RNAdistance (with full-structure tree-edit distance, also available from Vienna RNA package). We compute a quantile score for this observed average distance by comparing it to a null distribution of average structural distances. The null distribution is generated in the following manner: we start from the avian influenza sequence of each human-avian flu pair (Avian influenza evolve more slowly, so they are considered "ancestral sequences"), and introduce the same number of synonymous mutations, as well as preserve all the non-synonymous mutations, so that the "simulated" human flu sequence has the same number of synonymous and non-synonymous mutations as the actual pair. We also preserved the 3rd base codon usage when introducing mutations. We calculated the RNA structural distance between the resulting "simulated" human influenza sequence with the original influenza sequence, and the average from all such pairs contributes one data point to the null distribution. We repeated this process 200 times, so the null distribution consists of the distances from 200 simulations. We also explored a couple of different RNA structural metrics (Shapiro 1988; Shapiro and Zhang 1990; Fontana et al. 1993; Hofacker, Fekete, and Stadler

2002): tree-edit distance with full structure, tree-edit distance with coarse grained HIT structure and base pair distance (all available from RNAdistance program in the Vienna RNA package (Hofacker et al. 1994)), and they all yielded very similar results.

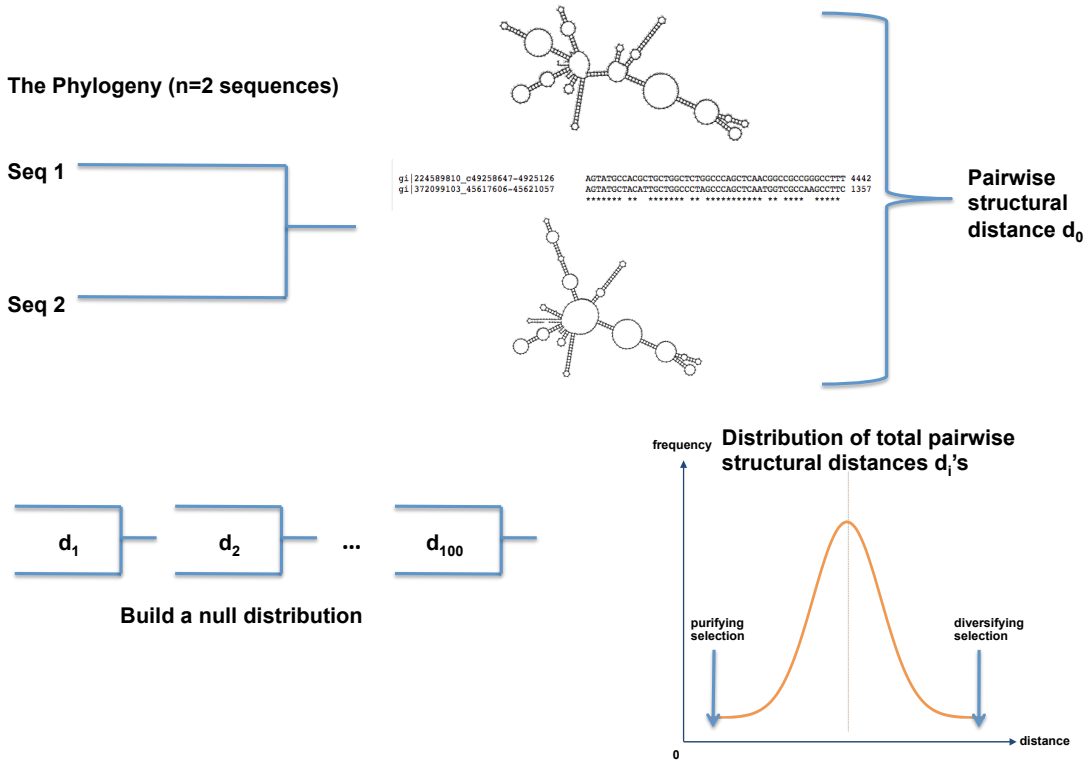


Figure 7: An illustration of the pairwise algorithm to detect selection.

Description of the phylogeny-controlled algorithm to detect selection

The algorithm (Figure 8 is an illustration of the pipeline of this algorithm.) scans along each viral segment with a sliding window of 60 bases and step size of 9 bases. For

each window we computed a quantile score by comparing the average pairwise RNA structural distances between actual influenza samples with a null distribution where all the average RNA pairwise structural distances come from simulated sets of influenza viral sequences that have the same amino acid sequences and phylogenetic relationship as the actual influenza sequences. The initial null distribution consists of average pairwise RNA structural distances from 20 simulations. For the windows that have extreme quantile scores (<0.1 or >0.9), we run 80 additional simulations (100 in total, so the minimum quantile score is 0.01) to get a more accurate estimate of the quantile score. The null distribution was generated as follows: starting from the root sequence, we introduced the same number of synonymous mutations along each branch (We kept all the non-synonymous changes along the trees so the resulting pool of simulated sequences will have the exact same amino acid sequences as the actual influenza sequences), until we reach the leaves of the phylogenetic tree. This process ensures that the resulting simulated sequences respect the exact phylogenetic relationships of the actual sequences, so that we can avoid the biased quantile scores produced by a subset of highly correlated sequences. We computed the average pairwise structural distance of these simulated sequences, and add this distance to the null distribution. The average pairwise RNA structural distance is calculated in the following manner: 1. For each window we first collect the set of RNA sequences, fold them using the program RNAfold from Vienna RNA package (Hofacker et al. 1994). 2. We calculated the average pairwise structural distances among each pair of these sequences using the program RNAdistance (with full-structure tree-edit distance, also available from Vienna RNA package).

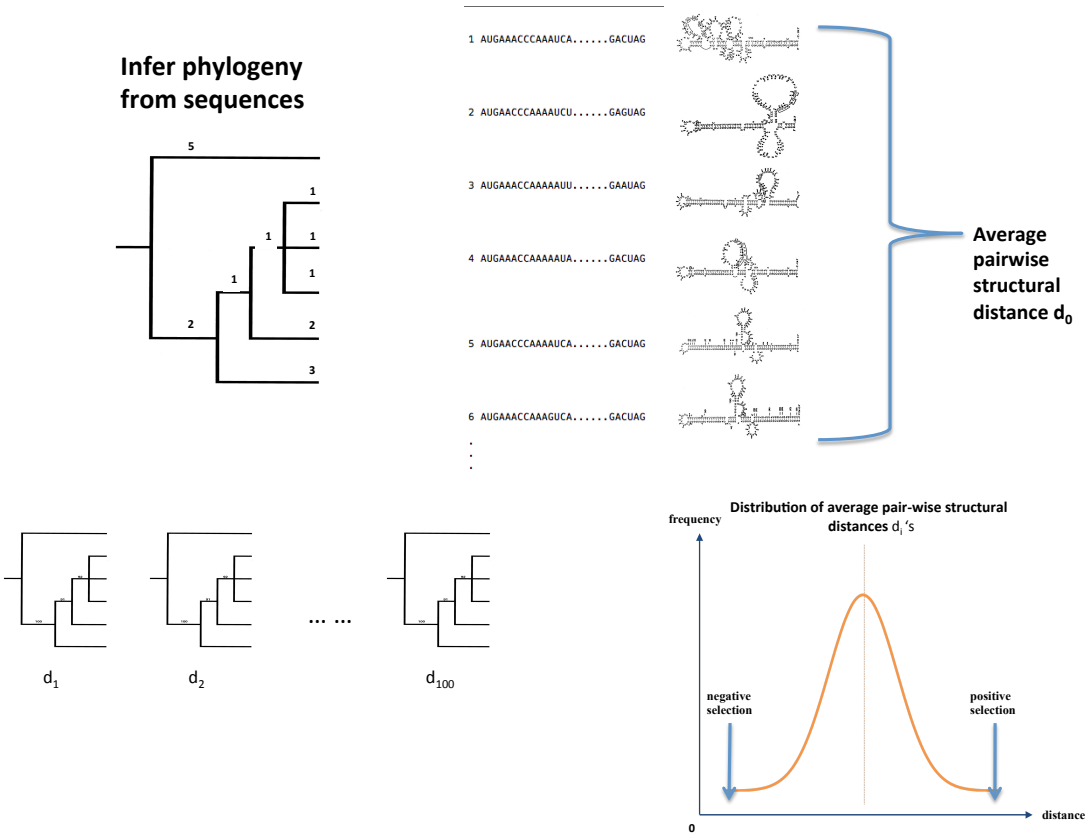


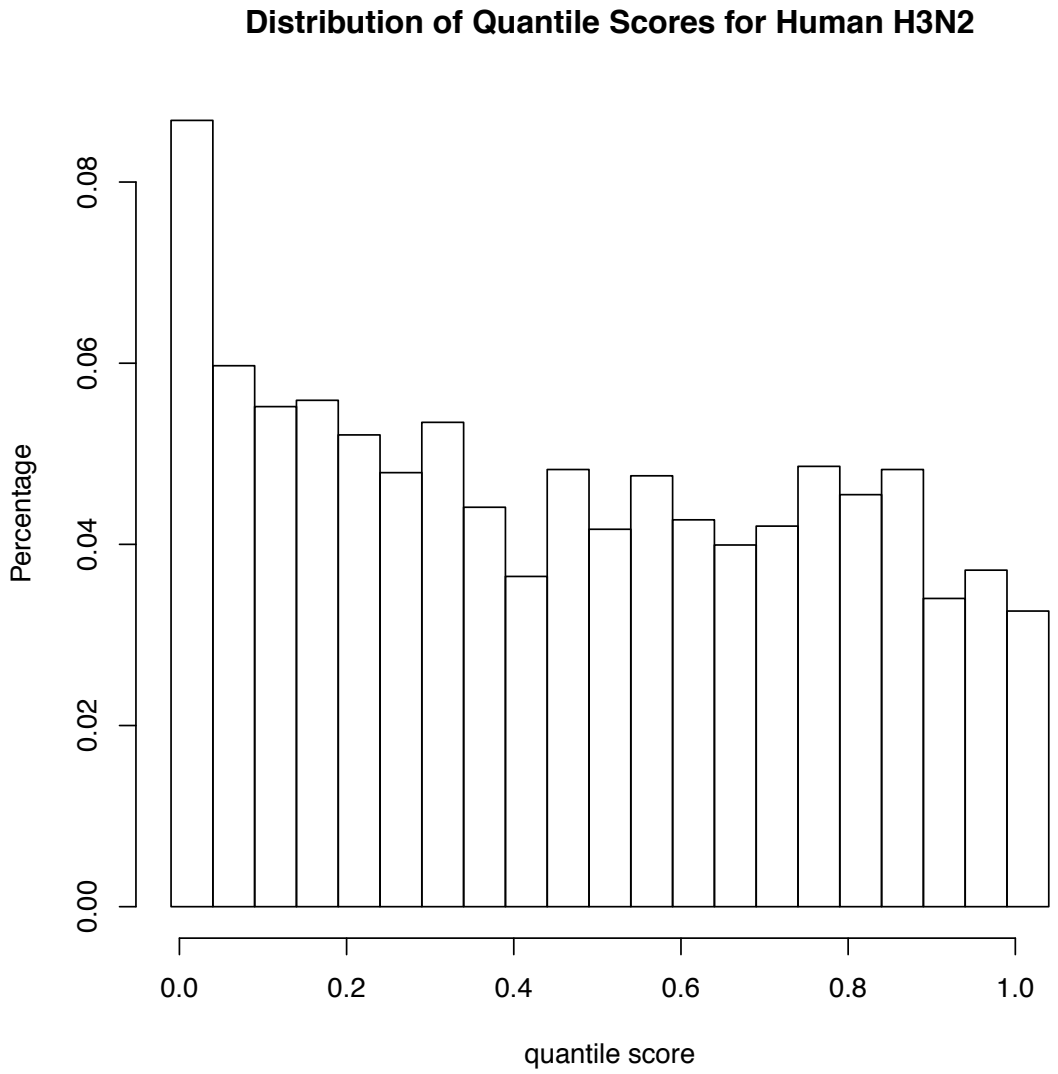
Figure 8: An illustration of the phylogeny-controlled algorithm to detect selection.

Calculating dN/dS values using the counting method

The algorithm we used to detect site-specific amino acid level selective pressure is adapted from a classic "counting method" (Nei and Gojobori 1986), which is appropriate because our dataset is so heavily sampled that parent-child nodes typically differ by a single substitution at most. For each amino acid residue, we start from the root of the reconstructed phylogenetic tree, and sum up the total number of synonymous (S)

and non-synonymous sites (N), as well as the total number of synonymous (S_d) and non-synonymous (N_d) differences, for each parent-child amino-acid pair. We calculate proportion of synonymous (p_S) and non-synonymous (p_N) differences as $p_S=S_d/S$, $p_N = N_d/N$. One then use the Juke-Cantor correction(Jukes Th 1969) to compute dN and dS. Then the dN/dS value is simply the ratio of these count-based estimates for dN and dS.

Supplementary Figures



Supplementary Figure 1: Distribution of quantile scores from all the 8 segments of Human H3N2 influenza A viruses. The two-sided p-value of Kolmogorov-Smirnov test against uniform distribution is <2.2 e-16.

Chapter Four

Conclusion and Future Directions

In my PhD study I mainly pursued two sets of research projects: 1) I analyzed genomes of eukaryotic species, and found evidence that short eukaryotic genes are selected to initiate faster, and also elongate faster. This implies that initiation is likely responsible for the pattern of higher ribosome densities on short eukaryotic genes. I also helped build and parameterize a mechanistic model of yeast protein translation to further identify the source of this selection. 2) In a very different domain of life, I developed two new algorithms to detect selective pressures on RNA structures in genomic sequences, applied to influenza A viruses. My analyses suggest a significant proportion of the influenza A genomes are under negative selection for RNA structures, primarily over long timescales. And I provided the first evidence of diversifying selective for RNA structures between human and avian influenza A viruses, suggesting a role for RNA structural changes in adaptation to host. Overall, this research suggests that we can find signatures of natural selection on noncoding nucleotide variation in completely different systems, across different time scales and for various phenotypic features. Selection on noncoding nucleotide variation seems to be widespread.

The research I described in this dissertation is only the tip of the iceberg in this important and currently underexplored field. There can potentially be many other noncoding nucleotide features that are under selection, for example epigenetic markers such as nucleosome positions(Prendergast and Semple 2011) and methylation sites.

Further research can be done to develop methods to detect selection on these important features.

In chapter 2 we discussed the relative importance of translation initiation vs. elongation in determining the various patterns of ribosome density. However other factors may also influence ribosome density. For example, some mRNA transcripts may be are more "accessible" to ribosomes, while others are membrane-bound, and thus are more difficult for ribosome to bind to them. This property can potentially influence ribosome densities, too, although its importance in determining patterns of observed ribosome densities still needs to be explored.

In this study, we found that short eukaryotic genes are selected for weak 5'mRNA structures, but are there other cases where there may be selection for strong mRNA structures? This will require us to have a sense of "neutral" variation, so that we can turn from relative selective pressure to study absolute selective pressure on RNA structures. Also gene length is likely a proxy for some other functional factors that are directly under selection, which we should strive to find. For example, we can divide genes into different functional categories, such as essential vs. nonessential genes, and see if there is stronger signal for the differential selection for translation initiation and elongation rates.

As to selection on codon bias, there has been a tremendous amount of research on the causes of this phenomenon. Besides selection for translation elongation rate, there have also been suggestions that codons that have more abundant corresponding tRNAs can cause less translational error(Drummond and Wilke 2009; Akashi 2001; Akashi 1994; Arava et al. 2005; Stoletzki and Eyre-Walker 2007), because more abundant tRNAs will

make it much easier for ribosomes to find the tRNAs that are the correct fit. However some recent studies challenge this assumption (Shah and Gilchrist 2010). There is still no consensus on what is the most important source of selection on codon usage. Overall, selection for faster elongation is probably more important, since this is a more universal source of selection. On the other hand, in most cases selection for less translational error may not be too strong, except for a few very important amino acid residues. However there is currently no firm evidence to support this view, more computational studies and clever experiments need to be done to distinguish the relative importance of these sources of selection.

Another important future research direction is to study the interaction between multiple layers of selection, for example selection for specific amino acid residues vs. RNA structure. We now know that controlling for the amino acid background, there is selection for RNA structures, however there may be cases where selection for RNA structures can be so strong that they can influence the choice of amino acids in certain regions, such as the 5' ends of genes. That is a very interesting question to be explored in the future.

For the stochastic simulation model for yeast protein translation, the model only considers the protein translation process. For application to multicellular differentiated eukaryotic organisms, our model needs to include more complicated features such as post-translational regulation and cell type. Another direction is to integrate more cellular processes such as transcription, translation, and metabolic reactions to build whole-cell computational models, and use them to predict cellular phenotypes.

For the algorithms to detect selection on RNA structures described in Chapter 3, there are also a number of future directions that we can pursue. First, in the dissertation we compared the influenza sequences derived from human vs. avian hosts and asked if there is a signature of selection for RNA structures. The same can be asked for other hosts, for example it would be interesting to do human vs. swine and avian vs. swine comparisons and see if there is some consistency in the genomic regions we detect. For the comparisons of human vs. swine influenza viruses, we would expect to see less positive selection for RNA structures, since human and swine hosts are closer, as compared to avian hosts. Second, another interesting question to ask if what are the functions of the RNA structural selection we detected? One hypothesis is that they are the regions that bind to the host RNA-binding-proteins (RBPs). There is a curated database for known human RBPs and their binding motifs. One can use them to search against influenza viral genomes and see if there are potential regions that may serve as the human RBP binding sites. There are also a number of longer-term future directions. For example, our algorithm takes RNA folding algorithms as given, so improved RNA folding algorithm can certainly improve the accuracy of our predictions as well. Several studies (for example(Vandivier et al. 2015)) have shown that chemical modifications on nucleotides can influence RNA base pairings, however these information are generally not taken into account in RNA folding algorithms, nor are they available for influenza viral sequences. Future progress in RNA folding algorithms can improve the predictions of our algorithm as well. Fourth, in the past few years a series of high-throughput methods to experimentally measure RNA structures have been developed, by combining

these measurement with RNA folding algorithms, one can get much more accurate RNA structure predictions. However our current version of the algorithm cannot use these experimentally RNA structural data, since our algorithm is based on comparing the observed RNA structural distances with a null distribution of simulated RNA structural distances. One can potentially measure the RNA structures for the observed RNA sequences, but it will be unrealistic to synthesize all the simulated RNA sequences and measure their RNA structures. How to leverage the available RNA structural measurements to improve our analysis of selection on RNA structure is the next important step in this direction.

References

Akashi, H. 1994. "Synonymous Codon Usage in Drosophila Melanogaster: Natural Selection and Translational Accuracy.." *Genetics* 136 (3). Genetics Society of America: 927–35.

Akashi, H. 2001. "Gene Expression and Molecular Evolution.." *Curr Opin Genet Dev* 11 (6): 660–66.

Altenhoff, A M, A Schneider, G H Gonnet, and C Dessimoz. 2011. "OMA 2011: Orthology Inference Among 1000 Complete Genomes." *Nucleic Acids Res* 39: D289–94.

Andersson, S G, and C G Kurland. 1990. "Codon Preferences in Free-Living Microorganisms." *Microbiol Rev* 54: 198–210.

Arava, Y, F E Boas, P O Brown, and D Herschlag. 2005. "Dissecting Eukaryotic Translation and Its Control by Ribosome Density Mapping." *Nucleic Acids Res* 33: 2421–32.

Arava, Y, Y Wang, J D Storey, C L Liu, P O Brown, and D Herschlag. 2003. "Genome-Wide Analysis of mRNA Translation Profiles in Saccharomyces Cerevisiae." *Proceedings of the National Academy of Sciences of the United States of America* 100: 3889–94.

Artieri, Carlo G, and Hunter B Fraser. 2014. "Evolution at Two Levels of Gene Expression in Yeast.." *Genome Research* 24 (3). Cold Spring Harbor Lab: 411–21. doi:10.1101/gr.165522.113.

Barrett, R D, and H E Hoekstra. 2011. “Molecular Spandrels: Tests of Adaptation at the Genetic Level.” *Nature Reviews.Genetics* 12 (11). England: 767–80. doi:10.1038/nrg3015.

Bazzini, A A, M T Lee, and A J Giraldez. 2012. “Ribosome Profiling Shows That miR-430 Reduces Translation Before Causing mRNA Decay in Zebrafish.” *Science (New York, N.Y.)* 336: 233–37.

Behrman, E L, S S Watson, K R O'Brien, M S Heschel, and P S Schmidt. 2015. “Seasonal Variation in Life History Traits in Two Drosophila Species..” *Journal of Evolutionary Biology* 28 (9): 1691–1704. doi:10.1111/jeb.12690.

Bergland, Alan O, Emily L Behrman, Katherine R O'Brien, Paul S Schmidt, and Dmitri A Petrov. 2014. “Genomic Evidence of Rapid and Stable Adaptive Oscillations Over Seasonal Time Scales in Drosophila..” Edited by Daniel Bolnick. *PLoS Genetics* 10 (11). Public Library of Science: e1004775. doi:10.1371/journal.pgen.1004775.

Bernhart, S H, I L Hofacker, S Will, A R Gruber, and P F Stadler. 2008. “RNAalifold: Improved Consensus Structure Prediction for RNA Alignments.” *BMC Bioinformatics* 9. England: 474–2105–9–474. doi:10.1186/1471-2105-9-474.

Bersaglieri, T, Sabeti, P. C., N Patterson, T Vanderploeg, S F Schaffner, J A Drake, M Rhodes, D E Reich, and J N Hirschhorn. 2004. “Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene.” *American Journal of Human Genetics* 74 (6). United States: 1111–20. doi:10.1086/421051.

Bettany, A J, P A Moore, R Cafferkey, L D Bell, A R Goodey, B L Carter, and A J Brown. 1989. “5'-Secondary Structure Formation, in Contrast to a Short String of Non-Preferred Codons, Inhibits the Translation of the Pyruvate Kinase mRNA in Yeast.” *Yeast* 5: 187–98.

Bloom, Jesse D, Lizhi Ian Gong, and David Baltimore. 2010. “Permissive Secondary Mutations Enable the Evolution of Influenza Oseltamivir Resistance..” *Science* 328 (5983). American Association for the Advancement of Science: 1272–75. doi:10.1126/science.1187816.

Bonhomme, Maxime, Claude Chevalet, Bertrand Servin, Simon Boitard, Jihad Abdallah, Sarah Blott, and Magali Sancristobal. 2010. “Detecting Selection in Population Trees: the Lewontin and Krakauer Test Extended..” *Genetics* 186 (1): 241–62. doi:10.1534/genetics.104.117275.

Bouvier, Nicole M, and Peter Palese. 2008. “The Biology of Influenza Viruses..” *Vaccine* 26 Suppl 4 (Suppl 4). NIH Public Access: D49–D53.

Branco-Price, C, R Kawaguchi, R B Ferreira, and J Bailey-Serres. 2005. “Genome-Wide Analysis of Transcript Abundance and Translation in Arabidopsis Seedlings Subjected to Oxygen Deprivation.” *Ann Bot* 96: 647–60.

Brar, G A, M Yassour, N Friedman, A Regev, N T Ingolia, and J S Weissman. 2012. “High-Resolution View of the Yeast Meiotic Program Revealed by Ribosome Profiling.” *Science (New York, N.Y.)* 335: 552–57.

Brar, Gloria A, and Jonathan S Weissman. 2015. “Ribosome Profiling Reveals the What, When, Where and How of Protein Synthesis..” *Nature Reviews.Molecular Cell Biology* 16 (11). Nature Publishing Group: 651–64. doi:10.1038/nrm4069.

Breaker, R R. 2012. “Riboswitches and the RNA World.” *Cold Spring Harbor Perspectives in Biology* 4 (2). United States: 10.1101–cshperspect.a003566. doi:10.1101/cshperspect.a003566.

Breaker, Ronald R. 2011. “Prospects for Riboswitch Discovery and Analysis.” 43 (6): 867–79. doi:10.1016/j.molcel.2011.08.024.

Breen, Michael S, Carsten Kemena, Peter K Vlasov, Cedric Notredame, and Fyodor A Kondrashov. 2012. “Epistasis as the Primary Factor in Molecular Evolution..” *Nature* 490 (7421): 535–38. doi:10.1038/nature11510.

Brower-Sinning, R, D M Carter, C J Crevar, E Ghedin, T M Ross, and P V Benos. 2009. “The Role of RNA Folding Free Energy in the Evolution of the Polymerase Genes of the Influenza a Virus.” *Genome Biology* 10 (2). England: R18. doi:10.1186/gb-2009-10-2-r18.

Brown, P O, and D Botstein. 1999. “Exploring the New World of the Genome with DNA Microarrays..” *Nature Genetics* 21 (1 Suppl): 33–37. doi:10.1038/4462.

Bulmer, M. 1991. “The Selection-Mutation-Drift Theory of Synonymous Codon Usage.” *Genetics* 129: 897–907.

Bush, R M, W M Fitch, C A Bender, and N J Cox. 1999. “Positive Selection on the H3 Hemagglutinin Gene of Human Influenza Virus A.” *Molecular Biology and Evolution* 16 (11). UNITED STATES: 1457–65.

Byars, S G, D Ewbank, D R Govindaraju, and S C Stearns. 2010. “Colloquium Papers: Natural Selection in a Contemporary Human Population.” *Proceedings of the National Academy of Sciences of the United States of America* 107 Suppl 1. United States: 1787–92. doi:10.1073/pnas.0906199106.

Cai, Zheng, Nicola J Camp, Lisa Cannon-Albright, and Alun Thomas. 2011. “Identification of Regions of Positive Selection Using Shared Genomic Segment Analysis..” *European Journal of Human Genetics : EJHG* 19 (6). Nature Publishing Group: 667–71. doi:10.1038/ejhg.2010.257.

Cannarozzi, Gina, Gina Cannarrozzi, Nicol N Schraudolph, Mahamadou Faty, Peter von Rohr, Markus T Friberg, Alexander C Roth, Pedro Gonnet, Gaston Gonnet, and Yves Barral. 2010. “A Role for Codon Order in Translation Dynamics..” *Cell* 141 (2): 355–67. doi:10.1016/j.cell.2010.02.036.

Cao, Song, and Shi-Jie Chen. 2006. “Predicting RNA Pseudoknot Folding Thermodynamics..” *Nucleic Acids Research* 34 (9). Oxford University Press: 2634–52. doi:10.1093/nar/gkl346.

Capon, F, M H Allen, M Ameen, A D Burden, D Tillman, J N Barker, and R C Trembath. 2004. “A Synonymous SNP of the Corneodesmosin Gene Leads to Increased mRNA Stability and Demonstrates Association with Psoriasis Across Diverse Ethnic Groups.” *Hum Mol Genet* 13: 2361–68.

Cataldo, L, M A Mastrangelo, and K C Kleene. 1999. “A Quantitative Sucrose Gradient Analysis of the Translational Activity of 18 mRNA Species in Testes From Adult Mice.” *Mol Hum Reprod* 5: 206–13.

Chamary, J V, and L D Hurst. 2005. “Evidence for Selection on Synonymous Mutations Affecting Stability of mRNA Secondary Structure in Mammals.” *Genome Biology* 6: R75.

Chamary, J V, J L Parmley, and L D Hurst. 2006. “Hearing Silence: Non-Neutral Evolution at Synonymous Sites in Mammals.” *Nat Rev Genet* 7: 98–108.
Chen, Han, Fangqin Lin, Ke Xing, and Xionglei He. 2015. “The Reverse Evolution From Multicellularity to Unicellularity During Carcinogenesis..” *Nature Communications* 6. Nature Publishing Group: 6367. doi:10.1038/ncomms7367.
Chen, Han, Ke Xing, and Xionglei He. 2015. “The dJ/dS Ratio Test Reveals Hundreds of Novel Putative Cancer Drivers..” *Molecular Biology and Evolution* 32 (8): 2181–85. doi:10.1093/molbev/msv083.
Chen, Hua, Nick Patterson, and David Reich. 2010. “Population Differentiation as a Test for Selective Sweeps..” *Genome Research* 20 (3). Cold Spring Harbor Lab: 393–402. doi:10.1101/gr.100545.109.
Chu, Dominique, and Tobias von der Haar. 2012. “The Architecture of Eukaryotic Translation..” *Nucleic Acids Research* 40 (20). Oxford University Press: 10098–106. doi:10.1093/nar/gks825.
Chu, Dominique, David J Barnes, and Tobias von der Haar. 2011. “The Role of tRNA and Ribosome Competition in Coupling the Expression of Different mRNAs in Saccharomyces Cerevisiae..” *Nucleic Acids Research* 39 (15). Oxford University Press: 6705–14. doi:10.1093/nar/gkr300.
Chursov, A, D Frishman, and A Shneider. 2013. “Conservation of mRNA Secondary Structures May Filter Out Mutations in Escherichia Coli Evolution.” *Nucleic Acids Res* 41 (16). England: 7854–60. doi:10.1093/nar/gkt507.
Clote, P, F Ferre, E Kranakis, and D Krizanc. 2005. “Structural RNA Has Lower Folding Energy Than Random RNA of the Same Dinucleotide Frequency.” *RNA (New York, N.Y.)* 11 (5). United States: 578–91. doi:10.1261/rna.7220505.
Cook, L M, B S Grant, I J Saccheri, and J Mallet. 2012. “Selective Bird Predation on the Peppered Moth: the Last Experiment of Michael Majerus..” *Biology Letters* 8 (4). The Royal Society: 609–12. doi:10.1098/rsbl.2011.1136.
Darwin, Charles. 1859. “On the Origin of Species by Means of Natural Selection, or, the Preservation of Favoured Races in the Struggle for Life,” no. Generic. London: J. Murray: 502.
Das, R, and D Baker. 2007. “Automated De Novo Prediction of Native-Like RNA Tertiary Structures.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (37). United States: 14664–69. doi:10.1073/pnas.0703836104.
David, L, W Huber, M Granovskaia, J Toedling, C J Palm, L Bofkin, T Jones, R W Davis, and L M Steinmetz. 2006. “A High-Resolution Map of Transcription in the Yeast Genome.” *Proceedings of the National Academy of Sciences of the United States of America* 103: 5320–25.
de Smit, M H, and J van Duin. 1990. “Secondary Structure of the Ribosome Binding Site Determines Translational Efficiency: a Quantitative Analysis.” *Proceedings of the National Academy of Sciences of the United States of America* 87: 7668–72.
Dela-Moss, L I, W N Moss, and D H Turner. 2014. “Identification of Conserved RNA Secondary Structures at Influenza B and C Splice Sites Reveals Similarities and Differences Between Influenza a, B, and C.” *BMC Research Notes* 7. England: 22–0500–7–22. doi:10.1186/1756-0500-7-22.

Ding, Feng, Shantanu Sharma, Poornima Chalasani, Vadim V Demidov, Natalia E Broude, and Nikolay V Dokholyan. 2008. “Ab Initio RNA Folding by Discrete Molecular Dynamics: From Structure Prediction to Folding Mechanisms..” *RNA (New York, N.Y.)* 14 (6). Cold Spring Harbor Lab: 1164–73. doi:10.1261/rna.894608.

Ding, X, L Jiang, C Ke, Z Yang, C Lei, K Cao, J Xu, et al. 2010. “Amino Acid Sequence Analysis and Identification of Mutations Under Positive Selection in Hemagglutinin of 2009 Influenza a (H1N1) Isolates.” *Virus Genes* 41 (3). United States: 329–40. doi:10.1007/s11262-010-0526-z.

Ding, Y, P Shah, and J B Plotkin. 2012. “Weak 5'-mRNA Secondary Structures in Short Eukaryotic Genes.” *Genome Biology and Evolution* 4 (10). England: 1046–53. doi:10.1093/gbe/evs082.

Ding, Y, W A Lorenz, and J H Chuang. 2012. “CodingMotif: Exact Determination of Overrepresented Nucleotide Motifs in Coding Sequences.” *BMC Bioinformatics* 13. England: 32. doi:10.1186/1471-2105-13-32.

Ding, Yang, Stefan Grünewald, and Peter J Humphries. 2011. “On Agreement Forests.” *Journal of Combinatorial Theory, Series A* 118 (7): 2059–65. doi:10.1016/j.jcta.2011.04.013.

Ding, Yang, William A Lorenz, Ivan Dotu, Evan Senter, and Peter Clote. 2014. “Computing the Probability of RNA Hairpin and Multiloop Formation..” *Journal of Computational Biology : a Journal of Computational Molecular Cell Biology* 21 (3). Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA: 201–18. doi:10.1089/cmb.2013.0148.

Dirks, Robert M, and Niles A Pierce. 2004. “An Algorithm for Computing Nucleic Acid Base-Pairing Probabilities Including Pseudoknots..” *Journal of Computational Chemistry* 25 (10). Wiley Subscription Services, Inc., A Wiley Company: 1295–1304. doi:10.1002/jcc.20057.

Drummond, D Allan, and Claus O Wilke. 2009. “The Evolutionary Consequences of Erroneous Protein Synthesis..” *Nature Reviews.Genetics* 10 (10). Nature Publishing Group: 715–24. doi:10.1038/nrg2662.

Duan, J, M S Wainwright, J M Comeron, N Saitou, A R Sanders, J Gelernter, and P V Gejman. 2003. “Synonymous Mutations in the Human Dopamine Receptor D2 (DRD2) Affect mRNA Stability and Synthesis of the Receptor.” *Hum Mol Genet* 12: 205–16.

Duret, L. 2002. “Evolution of Synonymous Codon Usage in Metazoans.” *Curr Opin Genet Dev* 12: 640–49.

Duret, L, and D Mouchiroud. 1999. “Expression Pattern and, Surprisingly, Gene Length Shape Codon Usage in Caenorhabditis, Drosophila, and Arabidopsis.” *Proceedings of the National Academy of Sciences of the United States of America* 96 (8). UNITED STATES: 4482–87.

Edgar, R C. 2004. “MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput.” *Nucleic Acids Res* 32 (5). England: 1792–97. doi:10.1093/nar/gkh340.

Egea, Raquel, Sònia Casillas, and Antonio Barbadilla. 2008. “Standard and Generalized McDonald-Kreitman Test: a Website to Detect Selection by Comparing Different

Classes of DNA Sites..” *Nucleic Acids Research* 36 (Web Server issue): W157–62. doi:10.1093/nar/gkn337.
Eisenberg, E, and E Y Levanon. 2003. “Human Housekeeping Genes Are Compact.” *Trends in Genetics : TIG* 19 (7). England: 362–65. doi:10.1016/S0168-9525(03)00140-9.
Enard, David, Philipp W Messer, and Dmitri A Petrov. 2014. “Genome-Wide Signals of Positive Selection in Human Evolution..” *Genome Research* 24 (6). Cold Spring Harbor Lab: 885–95. doi:10.1101/gr.164822.113.
Ewens, W J. 1972. “The Sampling Theory of Selectively Neutral Alleles..” *Theoretical Population Biology* 3 (1): 87–112.
Excoffier, L, T Hofer, and M Foll. 2009. “Detecting Loci Under Selection in a Hierarchically Structured Population..” *Heredity* 103 (4). Nature Publishing Group: 285–98. doi:10.1038/hdy.2009.74.
Eyre-Walker, A. 1996. “Synonymous Codon Bias Is Related to Gene Length in Escherichia Coli: Selection for Translational Accuracy?.” *Molecular Biology and Evolution* 13 (6). UNITED STATES: 864–72.
Eyre-Walker, A, and M Bulmer. 1993. “Reduced Synonymous Substitution Rate at the Start of Enterobacterial Genes.” *Nucleic Acids Res* 21: 4599–4603.
Fabian, Daniel K, Martin Kapun, Viola Nolte, Robert Kofler, Paul S Schmidt, Christian Schlötterer, and Thomas Flatt. 2012. “Genome-Wide Patterns of Latitudinal Differentiation Among Populations of Drosophila Melanogaster From North America..” *Molecular Ecology* 21 (19): 4748–69. doi:10.1111/j.1365-294X.2012.05731.x.
Fairbrother, W G, R F Yeh, P A Sharp, and C B Burge. 2002. “Predictive Identification of Exonic Splicing Enhancers in Human Genes.” *Science (New York, N.Y.)* 297 (5583). United States: 1007–13. doi:10.1126/science.1073774.
Fariello, María Inés, Simon Boitard, Hugo Naya, Magali Sancristobal, and Bertrand Servin. 2013. “Detecting Signatures of Selection Through Haplotype Differentiation Among Hierarchically Structured Populations..” *Genetics* 193 (3). Genetics Society of America: 929–41. doi:10.1534/genetics.112.147231.
Fay, Justin C. 2011. “Weighing the Evidence for Adaptation at the Molecular Level..” *Trends in Genetics : TIG* 27 (9): 343–49. doi:10.1016/j.tig.2011.06.003.
Foley, Shawn W, Lee E Vandivier, Pavel P Kuksa, and Brian D Gregory. 2015. “Transcriptome-Wide Measurement of Plant RNA Secondary Structure..” *Current Opinion in Plant Biology* 27 (October): 36–43. doi:10.1016/j.pbi.2015.05.021.
Fontana, W, D A Konings, P F Stadler, and P Schuster. 1993. “Statistics of RNA Secondary Structures.” *Biopolymers* 33 (9). UNITED STATES: 1389–1404. doi:10.1002/bip.360330909.
Frellsen, Jes, Ida Moltke, Martin Thiim, Kanti V Mardia, Jesper Ferkinghoff-Borg, and Thomas Hamelryck. 2009. “A Probabilistic Model of RNA Conformational Space..” Edited by Paul Gardner. *PLoS Computational Biology* 5 (6). Public Library of Science: e1000406. doi:10.1371/journal.pcbi.1000406.

Fu, Y X. 1997. “Statistical Tests of Neutrality of Mutations Against Population Growth, Hitchhiking and Background Selection..” *Genetics* 147 (2). Genetics Society of America: 915–25.

Fu, Y X, and W H Li. 1993. “Statistical Tests of Neutrality of Mutations..” *Genetics* 133 (3). Genetics Society of America: 693–709.

Fumagalli, Matteo, Manuela Sironi, Uberto Pozzoli, Anna Ferrer-Admetlla, Anna Ferrer-Admettla, Linda Pattini, and Rasmus Nielsen. 2011. “Signatures of Environmental Genetic Adaptation Pinpoint Pathogens as the Main Selective Pressure Through Human Evolution..” Edited by Joshua M Akey. *PLoS Genetics* 7 (11). Public Library of Science: e1002355. doi:10.1371/journal.pgen.1002355.

Garud, Nandita R, Philipp W Messer, Erkan O Buzbas, and Dmitri A Petrov. 2015. “Recent Selective Sweeps in North American Drosophila Melanogaster Show Signatures of Soft Sweeps..” Edited by Gregory P Copenhaver. *PLoS Genetics* 11 (2). Public Library of Science: e1005004. doi:10.1371/journal.pgen.1005004.

Gerashchenko, Maxim V, Alexei V Lobanov, and Vadim N Gladyshev. 2012. “Genome-Wide Ribosome Profiling Reveals Complex Translational Regulation in Response to Oxidative Stress..” *Proceedings of the National Academy of Sciences of the United States of America* 109 (43): 17394–99. doi:10.1073/pnas.1120799109.

Gingold, Hila, and Yitzhak Pilpel. 2011. “Determinants of Translation Efficiency and Accuracy..” *Molecular Systems Biology* 7 (1). EMBO Press: 481–81. doi:10.1038/msb.2011.14.

Goldman, N, and Z Yang. 1994. “A Codon-Based Model of Nucleotide Substitution for Protein-Coding DNA Sequences.” *Molecular Biology and Evolution* 11: 725–36.

Grossman, S R, I Shlyakhter, E K Karlsson, E H Byrne, S Morales, G Frieden, E Hostetter, et al. 2010. “A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection.” *Science (New York, N.Y.)* 327 (5967). United States: 883–86. doi:10.1126/science.1183863.

Grossman, Sharon R, Kristian G Andersen, Ilya Shlyakhter, Shervin Tabrizi, Sarah Winnicki, Angela Yen, Daniel J Park, et al. 2013. “Identifying Recent Adaptations in Large-Scale Genomic Data..” *Cell* 152 (4): 703–13. doi:10.1016/j.cell.2013.01.035.

Gu, W, M Li, Y Xu, T Wang, J H Ko, and T Zhou. 2014. “The Impact of RNA Structure on Coding Sequence Evolution in Both Bacteria and Eukaryotes.” *BMC Evolutionary Biology* 14. England: 87–2148–14–87. doi:10.1186/1471-2148-14-87.

Gu, W, T Zhou, and C O Wilke. 2010. “A Universal Trend of Reduced mRNA Stability Near the Translation-Initiation Site in Prokaryotes and Eukaryotes.” *PLoS Computational Biology* 6: e1000664.

Gultyaev, Alexander P, Anton Tsyganov-Bodounov, Monique IJ Spronken, Sander van der Kooij, Ron AM Fouchier, and René CL Olsthoorn. 2014. “RNA Structural Constraints in the Evolution of the Influenza a Virus Genome NP Segment.” *RNA Biology* 11 (7): 942–52. doi:10.4161/rna.29730.

Guo, H, N T Ingolia, J S Weissman, and D P Bartel. 2010. “Mammalian microRNAs Predominantly Act to Decrease Target mRNA Levels.” *Nature* 466: 835–40.

Guydosh, Nicholas R, and Rachel Green. 2014. “Dom34 Rescues Ribosomes in 3' Untranslated Regions..” *Cell* 156 (5): 950–62. doi:10.1016/j.cell.2014.02.006.

Haasl, Ryan J, and Bret A Payseur. 2015. “Fifteen Years of Genomewide Scans for Selection: Trends, Lessons and Unaddressed Genetic Sources of Complication..” *Molecular Ecology*, July, n/a–n/a. doi:10.1111/mec.13339.

Han, Lide, and Mark Abney. 2013. “Using Identity by Descent Estimation with Dense Genotype Data to Detect Positive Selection..” *European Journal of Human Genetics : EJHG* 21 (2). Nature Publishing Group: 205–11. doi:10.1038/ejhg.2012.148.

Hanchard, Neil A, Kirk A Rockett, Chris Spencer, Graham Coop, Margaret Pinder, Muminatou Jallow, Martin Kimber, Gil McVean, Richard Mott, and Dominic P Kwiatkowski. 2006. “Screening for Recently Selected Alleles by Analysis of Human Haplotype Similarity..” *American Journal of Human Genetics* 78 (1): 153–59. doi:10.1086/499252.

Hancock, Angela M, Benjamin Brachi, Nathalie Faure, Matthew W Horton, Lucien B Jarymowycz, F Gianluca Sperone, Chris Toomajian, Fabrice Roux, and Joy Bergelson. 2011. “Adaptation to Climate Across the Arabidopsis Thaliana Genome..” *Science* 334 (6052). American Association for the Advancement of Science: 83–86. doi:10.1126/science.1209244.

Harris, H. 1966. “Enzyme Polymorphisms in Man..” *Proceedings of the Royal Society of London. Series B, Biological Sciences* 164 (995): 298–310.

Havgaard, Jakob H, Elfar Torarinsson, and Jan Gorodkin. 2007. “Fast Pairwise Structural RNA Alignments by Pruning of the Dynamical Programming Matrix..” *PLoS Computational Biology* 3 (10). Public Library of Science: 1896–1908. doi:10.1371/journal.pcbi.0030193.

Hendrickson, D G, D J Hogan, H L McCullough, J W Myers, D Herschlag, J E Ferrell, and P O Brown. 2009. “Concordant Regulation of Translation and mRNA Abundance for Hundreds of Targets of a Human microRNA.” *PLoS Biol* 7: e1000238.

Hendry, A P. 2013. “Key Questions in the Genetics and Genomics of Eco-Evolutionary Dynamics..” *Heredity* 111 (6). Nature Publishing Group: 456–66. doi:10.1038/hdy.2013.75.

Hershberg, R, and D A Petrov. 2008. “Selection on Codon Bias.” *Annu Rev Genet* 42: 287–99.

Hirsh, A E, H B Fraser, and D P Wall. 2005. “Adjusting for Selection on Synonymous Sites in Estimates of Evolutionary Distance.” *Molecular Biology and Evolution* 22: 174–77.

Hoekstra, H E, R J Hirschmann, R A Bundey, P A Insel, and J P Crossland. 2006. “A Single Amino Acid Mutation Contributes to Adaptive Beach Mouse Color Pattern.” *Science (New York, N.Y.)* 313 (5783). United States: 101–4. doi:10.1126/science.1126121.

Hofacker, I L, M Fekete, and P F Stadler. 2002. “Secondary Structure Prediction for Aligned RNA Sequences.” *Journal of Molecular Biology* 319 (5). England: Elsevier Science Ltd: 1059–66. doi:10.1016/S0022-2836(02)00308-X.

Hofacker, I L, W Fontana, P F Stadler, L S Bonhoeffer, M Tacker, and P Schuster. 1994. “Fast Folding and Comparison of RNA Secondary Structures.” *Monatshefte Fur Chemie* 125: 167–88.

Hoffman, Michael M, and Ewan Birney. 2007. “Estimating the Neutral Rate of Nucleotide Substitution Using Introns..” *Molecular Biology and Evolution* 24 (2): 522–31. doi:10.1093/molbev/msl179.
Hoffman, Michael M, and Ewan Birney. 2010. “An Effective Model for Natural Selection in Promoters..” *Genome Research* 20 (5). Cold Spring Harbor Lab: 685–92. doi:10.1101/gr.096719.109.
Hubby, J L, and R C Lewontin. 1966. “A Molecular Approach to the Study of Genic Heterozygosity in Natural Populations. I. the Number of Alleles at Different Loci in Drosophila Pseudoobscura.” *Genetics* 54 (2). UNITED STATES: 577–94.
Huber, Christian D, Magnus Nordborg, Joachim Hermisson, and Ines Hellmann. 2014. “Keeping It Local: Evidence for Positive Selection in Swedish Arabidopsis Thaliana..” *Molecular Biology and Evolution* 31 (11). Oxford University Press: 3026–39. doi:10.1093/molbev/msu247.
Hudson, R R, M Kreitman, and M Aguade. 1987. “A Test of Neutral Molecular Evolution Based on Nucleotide Data..” *Genetics* 116 (1). Genetics Society of America: 153–59.
Hurowitz, E H, and P O Brown. 2003. “Genome-Wide Analysis of mRNA Lengths in Saccharomyces Cerevisiae.” *Genome Biology* 5: R2.
Hurst, Laurence D. 2002. “The Ka/Ks Ratio: Diagnosing the Form of Sequence Evolution..” *Trends in Genetics : TIG* 18 (9): 486.
Hutchinson, E C, J C von Kirchbach, J R Gog, and P Digard. 2010. “Genome Packaging in Influenza a Virus.” *The Journal of General Virology* 91 (Pt 2). England: 313–28. doi:10.1099/vir.0.017608-0.
Ingolia, N T, L F Lareau, and J S Weissman. 2011. “Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes.” *Cell* 147: 789–802.
Ingolia, N T, S Ghaemmaghami, J R Newman, and J S Weissman. 2009. “Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling.” *Science* 324: 218–23.
Ingolia, Nicholas T. 2014. “Ribosome Profiling: New Views of Translation, From Single Codons to Genome Scale..” *Nature Reviews.Genetics* 15 (3). Nature Publishing Group: 205–13. doi:10.1038/nrg3645.
Jackson, Richard J, Christopher U T Hellen, and Tatyana V Pestova. 2010. “The Mechanism of Eukaryotic Translation Initiation and Principles of Its Regulation..” *Nature Reviews.Molecular Cell Biology* 11 (2). Nature Publishing Group: 113–27. doi:10.1038/nrm2838.
Jambhekar, A, and J L Derisi. 2007. “Cis-Acting Determinants of Asymmetric, Cytoplasmic RNA Transport.” *RNA (New York, N.Y.)* 13 (5). United States: 625–42. doi:10.1261/rna.262607.
Jiang, T, S D Kennedy, W N Moss, E Kierzek, and D H Turner. 2014. “Secondary Structure of a Conserved Domain in an Intron of Influenza a M1 mRNA.” *Biochemistry* 53 (32). United States: 5236–48. doi:10.1021/bi500611j.
Jones, Eric, Travis Oliphant, and Peterson Pearu. 2001. “SciPy: Open Source Scientific Tools for Python.”

Jonikas, Magdalena A, Randall J Radmer, Alain Laederach, Rhiju Das, Samuel Pearlman, Daniel Herschlag, and Russ B Altman. 2009. “Coarse-Grained Modeling of Large RNA Molecules with Knowledge-Based Potentials and Structural Filters..” *RNA (New York, N.Y.)* 15 (2). Cold Spring Harbor Lab: 189–99. doi:10.1261/rna.1270809.

Jukes Th, Cantor C R. 1969. “Evolution of Protein Molecules.” In *Mammalian Protein Metabolism: Volume III*, edited by H N Munro, 1:21–132. New York: Academic Press.

Karlsson, Elinor K, Dominic P Kwiatkowski, and Pardis C Sabeti. 2014. “Natural Selection and Infectious Disease in Human Populations..” *Nature Reviews.Genetics* 15 (6). Nature Publishing Group: 379–93. doi:10.1038/nrg3734.

Keller, T E, S D Mis, K E Jia, and C O Wilke. 2012. “Reduced mRNA Secondary-Structure Stability Near the Start Codon Indicates Functional Genes in Prokaryotes.” *Genome Biology and Evolution* 4: 80–88.

Kerpedjiev, Peter, Christian Höner Zu Siederdissen, and Ivo L Hofacker. 2015. “Predicting RNA 3D Structure Using a Coarse-Grain Helix-Centered Model..” *RNA (New York, N.Y.)* 21 (6). Cold Spring Harbor Lab: 1110–21. doi:10.1261/rna.047522.114.

Kim, Yuseob, and Rasmus Nielsen. 2004. “Linkage Disequilibrium as a Signature of Selective Sweeps..” *Genetics* 167 (3). Genetics Society of America: 1513–24. doi:10.1534/genetics.103.025387.

Kim, Yuseob, and Wolfgang Stephan. 2002. “Detecting a Local Signature of Genetic Hitchhiking Along a Recombining Chromosome..” *Genetics* 160 (2). Genetics Society of America: 765–77.

Kimchi-Sarfaty, C, J M Oh, I W Kim, Z E Sauna, A M Calcagno, S V Ambudkar, and M M Gottesman. 2007. “A ‘Silent’ Polymorphism in the MDR1 Gene Changes Substrate Specificity.” *Science* 315 (5811). United States: 525–28. doi:10.1126/science.1135308.

Kimura, M. 1977. “Preponderance of Synonymous Changes as Evidence for the Neutral Theory of Molecular Evolution.” *Nature* 267: 275–76.

Kimura, Motoo. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge Cambridgeshire ; New York: Cambridge University Press.

Knudsen, B, and J Hein. 2003. “Pfold: RNA Secondary Structure Prediction Using Stochastic Context-Free Grammars.” *Nucleic Acids Research* 31 (13). England: 3423–28.

Kosakovsky Pond, S L, A F Poon, A J Leigh Brown, and S D Frost. 2008. “A Maximum Likelihood Method for Detecting Directional Evolution in Protein Sequences and Its Application to Influenza a Virus.” *Molecular Biology and Evolution* 25 (9). United States: 1809–24. doi:10.1093/molbev/msn123.

Kryazhimskiy, Sergey, Jonathan Dushoff, Georgii A Bazykin, and Joshua B Plotkin. 2011. “Prevalence of Epistasis in the Evolution of Influenza a Surface Proteins..” Edited by Harmit S Malik. *PLoS Genetics* 7 (2). Public Library of Science: e1001301. doi:10.1371/journal.pgen.1001301.

Kudla, G, A W Murray, D Tollervey, and J B Plotkin. 2009. “Coding-Sequence Determinants of Gene Expression in Escherichia Coli.” *Science (New York, N.Y.)* 324 (5924). United States: 255–58. doi:10.1126/science.1170160.

Kudla, G, L Lipinski, F Caffin, A Helwak, and M Zylicz. 2006. “High Guanine and Cytosine Content Increases mRNA Levels in Mammalian Cells.” *PLoS Biol* 4 (6). United States: e180. doi:10.1371/journal.pbio.0040180.

Kural, D, Y Ding, J Wu, A M Korpi, and J H Chuang. 2009. “COMIT: Identification of Noncoding Motifs Under Selection in Coding Sequences.” *Genome Biology* 10 (11). England: R133. doi:10.1186/gb-2009-10-11-r133.

Kwiatkowski, Dominic P. 2005. “How Malaria Has Affected the Human Genome and What Human Genetics Can Teach Us About Malaria..” *American Journal of Human Genetics* 77 (2): 171–92. doi:10.1086/432519.

Kwok, Chun Kit, Yin Tang, Sarah M Assmann, and Philip C Bevilacqua. 2015. “The RNA Structurome: Transcriptome-Wide Structure Probing with Next-Generation Sequencing..” *Trends in Biochemical Sciences* 40 (4): 221–32. doi:10.1016/j.tibs.2015.02.005.

Lachance, Joseph, and Sarah A Tishkoff. 2013. “Population Genomics of Human Adaptation..” *Annual Review of Ecology Evolution and Systematics* 44 (1). Annual Reviews: 123–43. doi:10.1146/annurev-ecolsys-110512-135833.

Lackner, Daniel H, Traude H Beilharz, Samuel Marguerat, Juan Mata, Stephen Watt, Falk Schubert, Thomas Preiss, and Jürg Bähler. 2007. “A Network of Multiple Regulatory Layers Shapes Gene Expression in Fission Yeast..” *Molecular Cell* 26 (1): 145–55. doi:10.1016/j.molcel.2007.03.002.

Lacsina, J R, G LaMonte, C V Nicchitta, and J T Chi. 2011. “Polysome Profiling of the Malaria Parasite Plasmodium Falciparum.” *Mol Biochem Parasitol* 179: 42–46.

Langley, Charles H, Kristian Stevens, Charis Cardeno, Yuh Chwen G Lee, Daniel R Schrider, John E Pool, Sasha A Langley, et al. 2012. “Genomic Variation in Natural Populations of Drosophila Melanogaster..” *Genetics* 192 (2). Genetics Society of America: 533–98. doi:10.1534/genetics.112.142018.

Lewontin, R C, and J Krakauer. 1973. “Distribution of Gene Frequency as a Test of the Theory of the Selective Neutrality of Polymorphisms..” *Genetics* 74 (1). Genetics Society of America: 175–95.

Li, G W, E Oh, and J S Weissman. 2012. “The Anti-Shine-Dalgarno Sequence Drives Translational Pausing and Codon Choice in Bacteria.” *Nature* 484: 538–41. doi:10.1038/nature10965.

Lin, Z, and W H Li. 2012. “Evolution of 5' Untranslated Region Length and Gene Expression Reprogramming in Yeasts.” *Molecular Biology and Evolution* 29: 81–89.

Luksza, Marta, and Michael Lässig. 2014. “A Predictive Fitness Model for Influenza..” *Nature* 507 (7490). Nature Publishing Group: 57–61. doi:10.1038/nature13087.

Lyngsø, R B, and C N Pedersen. 2000. “RNA Pseudoknot Prediction in Energy-Based Models..” *Journal of Computational Biology : a Journal of Computational Molecular Cell Biology* 7 (3-4). Mary Ann Liebert, Inc.: 409–27. doi:10.1089/106652700750050862.

Majerus, M E N. 1998. *Melanism*. Oxford University Press, USA.

Mathews, D H, and D H Turner. 2002. “Dynalign: an Algorithm for Finding the Secondary Structure Common to Two RNA Sequences.” *Journal of Molecular Biology* 317 (2). England: Elsevier Science Ltd: 191–203. doi:10.1006/jmbi.2001.5351.

McCandlish, David M, Etienne Rajon, Premal Shah, Yang Ding, and Joshua B Plotkin. 2013. “The Role of Epistasis in Protein Evolution..” *Nature* 497 (7451). Nature Publishing Group: E1–2–discussionE2–3. doi:10.1038/nature12219.

McDonald, J H, and M Kreitman. 1991. “Adaptive Protein Evolution at the Adh Locus in Drosophila.” *Nature* 351: 652–54.

McManus, C Joel, Gemma E May, Pieter Spealman, and Alan Shteyman. 2014. “Ribosome Profiling Reveals Post-Transcriptional Buffering of Divergent Gene Expression in Yeast..” *Genome Research* 24 (3). Cold Spring Harbor Lab: 422–30. doi:10.1101/gr.164996.113.

Messer, Philipp W, and Dmitri A Petrov. 2013. “Population Genomics of Rapid Adaptation by Soft Selective Sweeps..” *Trends in Ecology & Evolution* 28 (11): 659–69. doi:10.1016/j.tree.2013.08.003.

Messer, Philipp W, and Richard A Neher. 2012. “Estimating the Strength of Selective Sweeps From Deep Population Diversity Data..” *Genetics* 191 (2). Genetics Society of America: 593–605. doi:10.1534/genetics.112.138461.

Milot, E, F M Mayer, D H Nussey, M Boisvert, F Pelletier, and D Reale. 2011. “Evidence for Evolution in Response to Natural Selection in a Contemporary Human Population.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (41). United States: 17040–45. doi:10.1073/pnas.1104210108.

Moss, W N, L I Dela-Moss, S F Priore, and D H Turner. 2012. “The Influenza a Segment 7 mRNA 3' Splice Site Pseudoknot/Hairpin Family.” *RNA Biology* 9 (11). United States: 1305–10. doi:10.4161/rna.22343.

Moss, W N, S F Priore, and D H Turner. 2011. “Identification of Potential Conserved RNA Secondary Structure Throughout Influenza a Coding Regions.” *RNA (New York, N.Y.)* 17 (6). United States: 991–1011. doi:10.1261/rna.2619511.

Muse, S V, and B S Gaut. 1994. “A Likelihood Approach for Comparing Synonymous and Nonsynonymous Nucleotide Substitution Rates, with Application to the Chloroplast Genome.” *Molecular Biology and Evolution* 11: 715–24.

Nackley, A G, S A Shabalina, I E Tchivileva, K Satterfield, O Korchynskyi, S S Makarov, W Maixner, and L Diatchenko. 2006. “Human Catechol-O-Methyltransferase Haplotypes Modulate Protein Expression by Altering mRNA Secondary Structure.” *Science (New York, N.Y.)* 314 (5807). United States: 1930–33. doi:10.1126/science.1131262.

Nawrocki, E P, D L Kolbe, and S R Eddy. 2009. “Infernal 1.0: Inference of RNA Alignments.” *Bioinformatics (Oxford, England)* 25 (10). England: 1335–37. doi:10.1093/bioinformatics/btp157.

Neher, Richard A, Colin A Russell, and Boris I Shraiman. 2014. “Predicting Evolution From the Shape of Genealogical Trees..” Edited by Gil McVean. *eLife* 3. eLife Sciences Publications Limited: e01914. doi:10.7554/eLife.03568.

Nei, M, and T Gojobori. 1986. “Simple Methods for Estimating the Numbers of Synonymous and Nonsynonymous Nucleotide Substitutions.” *Molecular Biology and Evolution* 3 (5). UNITED STATES: 418–26.

Nei, Masatoshi, Yoshiyuki Suzuki, and Masafumi Nozawa. 2010. “The Neutral Theory of Molecular Evolution in the Genomic Era..” *Annual Review of Genomics and Human Genetics* 11 (1). Annual Reviews: 265–89. doi:10.1146/annurev-genom-082908-150129.

Nelson, Martha I, and Edward C Holmes. 2007. “The Evolution of Epidemic Influenza..” *Nat Rev Genet* 8 (3). Nature Publishing Group: 196–205. doi:10.1038/nrg2053.

Nemeroff, M E, U Utans, A Krämer, and R M Krug. 1992. “Identification of Cis-Acting Intron and Exon Regions in Influenza Virus NS1 mRNA That Inhibit Splicing and Cause the Formation of Aberrantly Sedimenting Presplicing Complexes..” *Molecular and Cellular Biology* 12 (3). American Society for Microbiology (ASM): 962–70.

Neverov, Alexey D, Sergey Kryazhimskiy, Joshua B Plotkin, and Georgii A Bazykin. 2015. “Coordinated Evolution of Influenza a Surface Proteins..” Edited by Harmit S Malik. *PLoS Genetics* 11 (8). Public Library of Science: e1005404. doi:10.1371/journal.pgen.1005404.

Nielsen, Rasmus, Melissa J Hubisz, Ines Hellmann, Dara Torgerson, Aida M Andrés, Anders Albrechtsen, Ryan Gutenkunst, et al. 2009. “Darwinian and Demographic Forces Affecting Human Protein Coding Genes..” *Genome Research* 19 (5). Cold Spring Harbor Lab: 838–49. doi:10.1101/gr.088336.108.

Nielsen, Rasmus, Scott Williamson, Yuseob Kim, Melissa J Hubisz, Andrew G Clark, and Carlos Bustamante. 2005. “Genomic Scans for Selective Sweeps Using SNP Data..” *Genome Research* 15 (11). Cold Spring Harbor Lab: 1566–75. doi:10.1101/gr.4252305.

Oh, E, A H Becker, A Sandikci, D Huber, R Chaba, F Gloge, R J Nichols, et al. 2011. “Selective Ribosome Profiling Reveals the Cotranslational Chaperone Action of Trigger Factor in Vivo.” *Cell* 147: 1295–1308.

Parisien, M, and F Major. 2008. “The MC-Fold and MC-Sym Pipeline Infers RNA Structure From Sequence Data.” *Nature* 452 (7183). England: 51–55. doi:10.1038/nature06684.

Pedersen, J S, G Bejerano, A Siepel, K Rosenbloom, K Lindblad-Toh, E S Lander, J Kent, W Miller, and D Haussler. 2006. “Identification and Classification of Conserved RNA Secondary Structures in the Human Genome.” *PLoS Computational Biology* 2 (4). United States: e33. doi:10.1371/journal.pcbi.0020033.

Pelz, Hans-Joachim, Simone Rost, Mirja Hünerberg, Andreas Fregin, Ann-Charlotte Heiberg, Kristof Baert, Alan D MacNicoll, et al. 2005. “The Genetic Basis of Resistance to Anticoagulants in Rodents..” *Genetics* 170 (4). Genetics Society of America: 1839–47. doi:10.1534/genetics.104.040360.

Penel, S, A M Arigon, J F Dufayard, A S Sertier, V Daubin, L Duret, M Gouy, and G Perriere. 2009. “Databases of Homologous Gene Families for Comparative Genomics.” *BMC Bioinformatics* 10 Suppl 6: S3.

Plotch, S J, and R M Krug. 1986. “In Vitro Splicing of Influenza Viral NS1 mRNA and NS1-Beta-Globin Chimeras: Possible Mechanisms for the Control of Viral mRNA

Splicing..” *Proceedings of the National Academy of Sciences of the United States of America* 83 (15). National Academy of Sciences: 5444–48.
Plotkin, J B, and G Kudla. 2011. “Synonymous but Not the Same: the Causes and Consequences of Codon Bias.” *Nature Reviews.Genetics* 12 (1). England: 32–42. doi:10.1038/nrg2899.
Plotkin, Joshua B. 2010. “Transcriptional Regulation Is Only Half the Story..” *Molecular Systems Biology* 6 (1). EMBO Press: 406. doi:10.1038/msb.2010.63.
Pool, John E, Russell B Corbett-Detig, Ryuichi P Sugino, Kristian A Stevens, Charis M Cardeno, Marc W Crepeau, Pablo Duchen, et al. 2012. “Population Genomics of Sub-Saharan Drosophila Melanogaster: African Diversity and Non-African Admixture..” Edited by Harmit S Malik. *PLoS Genetics* 8 (12). Public Library of Science: e1003080. doi:10.1371/journal.pgen.1003080.
Popenda, Mariusz, Marta Szachniuk, Maciej Antczak, Katarzyna J Purzycka, Piotr Lukasiak, Natalia Bartol, Jacek Blazewicz, and Ryszard W Adamiak. 2012. “Automated 3D Structure Composition for Large RNAs..” *Nucleic Acids Research* 40 (14). Oxford University Press: e112–12. doi:10.1093/nar/gks339.
Prendergast, J G, and C A Semple. 2011. “Widespread Signatures of Recent Selection Linked to Nucleosome Positioning in the Human Lineage.” *Genome Research* 21 (11). United States: 1777–87. doi:10.1101/gr.122275.111.
Priore, S F, E Kierzek, R Kierzek, J R Baman, W N Moss, L I Dela-Moss, and D H Turner. 2013. “Secondary Structure of a Conserved Domain in the Intron of Influenza a NS1 mRNA.” *PloS One* 8 (9). United States: e70615. doi:10.1371/journal.pone.0070615.
Priore, S F, W N Moss, and D H Turner. 2012. “Influenza a Virus Coding Regions Exhibit Host-Specific Global Ordered RNA Structure.” *PloS One* 7 (4). United States: e35989. doi:10.1371/journal.pone.0035989.
Priore, S F, W N Moss, and D H Turner. 2013. “Influenza B Virus Has Global Ordered RNA Structure in (+) and (-) Strands but Relatively Less Stable Predicted RNA Folding Free Energy Than Allowed by the Encoded Protein Sequence.” *BMC Research Notes* 6. England: 330–0500–6–330. doi:10.1186/1756-0500-6-330.
Qin, X, S Ahn, T P Speed, and G M Rubin. 2007. “Global Analyses of mRNA Translational Control During Early Drosophila Embryogenesis.” *Genome Biology* 8: R63.
Rao, Y S, Z F Wang, X W Chai, G Z Wu, M Zhou, Q H Nie, and X Q Zhang. 2010. “Selection for the Compactness of Highly Expressed Genes in Gallus Gallus.” *Biology Direct* 5. England: 35. doi:10.1186/1745-6150-5-35.
Reid, D W, and C V Nicchitta. 2012. “Primary Role for Endoplasmic Reticulum-Bound Ribosomes in Cellular Translation Identified by Ribosome Profiling.” *J Biol Chem* 287: 5518–27.
Reinhardt, Josie A, Bryan Kolaczkowski, Corbin D Jones, David J Begun, and Andrew D Kern. 2014. “Parallel Geographic Variation in Drosophila Melanogaster..” *Genetics* 197 (1). Genetics Society of America: 361–73. doi:10.1534/genetics.114.161463.
Ren, Jihong, Baharak Rastegari, Anne Condon, and Holger H Hoos. 2005. “HotKnots: Heuristic Prediction of RNA Secondary Structures Including Pseudoknots..” *RNA*

(New York, N.Y.) 11 (10). Cold Spring Harbor Lab: 1494–1504. doi:10.1261/rna.7284905.
Reuter, Jessica S, and David H Mathews. 2010. "RNAstructure: Software for RNA Secondary Structure Prediction and Analysis.." *BMC Bioinformatics* 11 (1). BioMed Central Ltd: 129. doi:10.1186/1471-2105-11-129.
Reuveni, S, I Meilijson, M Kupiec, E Ruppin, and T Tuller. 2011. "Genome-Scale Analysis of Translation Elongation with a Ribosome Flow Model." *PLoS Computational Biology* 7 (9). United States: e1002127. doi:10.1371/journal.pcbi.1002127.
Rivas, E, and S R Eddy. 1999. "A Dynamic Programming Algorithm for RNA Structure Prediction Including Pseudoknots.." *Journal of Molecular Biology* 285 (5): 2053–68. doi:10.1006/jmbi.1998.2436.
Sabeti, Pardis C, David E Reich, John M Higgins, Haninah Z P Levine, Daniel J Richter, Stephen F Schaffner, Stacey B Gabriel, et al. 2002. "Detecting Recent Positive Selection in the Human Genome From Haplotype Structure.." *Nature* 419 (6909). Nature Publishing Group: 832–37. doi:10.1038/nature01140.
Sabeti, Pardis C, Patrick Varilly, Ben Fry, Jason Lohmueller, Elizabeth Hostetter, Chris Cotsapas, Xiaohui Xie, et al. 2007. "Genome-Wide Detection and Characterization of Positive Selection in Human Populations.." *Nature* 449 (7164). Nature Publishing Group: 913–18. doi:10.1038/nature06250.
Sawyer, S A, and D L Hartl. 1992. "Population Genetics of Polymorphism and Divergence." *Genetics* 132: 1161–76.
Sella, G, D A Petrov, M Przeworski, and P Andolfatto. 2009. "Pervasive Natural Selection in the Drosophila Genome?." *PLoS Genetics* 5 (6). United States: e1000495. doi:10.1371/journal.pgen.1000495.
Shah, P, and M A Gilchrist. 2010. "Effect of Correlated tRNA Abundances on Translation Errors and Evolution of Codon Usage Bias." *PLoS Genetics* 6.
Shah, P, and M A Gilchrist. 2011. "Explaining Complex Codon Usage Patterns with Selection for Translational Efficiency, Mutation Bias, and Genetic Drift." *Proceedings of the National Academy of Sciences of the United States of America* 108: 10231–36.
Shah, Premal, Yang Ding, Malwina Niemczyk, Grzegorz Kudla, and Joshua B Plotkin. 2013. "Rate-Limiting Steps in Yeast Protein Translation.." *Cell* 153 (7): 1589–1601. doi:10.1016/j.cell.2013.05.049.
Shapiro, B A. 1988. "An Algorithm for Comparing Multiple RNA Secondary Structures." *Computer Applications in the Biosciences : CABIOS* 4 (3). ENGLAND: 387–93.
Shapiro, B A, and K Z Zhang. 1990. "Comparing Multiple RNA Secondary Structures Using Tree Comparisons." *Computer Applications in the Biosciences : CABIOS* 6 (4). ENGLAND: 309–18.
Sharp, P M, and W H Li. 1987. "The Codon Adaptation Index--a Measure of Directional Synonymous Codon Usage Bias, and Its Potential Applications." *Nucleic Acids Res* 15: 1281–95.
Sharp, P M, L R Emery, and K Zeng. 2010. "Forces That Influence the Evolution of Codon Bias." *Philos Trans R Soc Lond B Biol Sci* 365: 1203–12.

Sharp, P M, M Averof, A T Lloyd, G Matassi, and J F Peden. 1995. “DNA Sequence Evolution: the Sounds of Silence.” *Philos Trans R Soc Lond B Biol Sci* 349: 241–47.
Shen, L X, J P Basilion, and V P Jr Stanton. 1999. “Single-Nucleotide Polymorphisms Can Cause Different Structural Folds of mRNA.” *Proceedings of the National Academy of Sciences of the United States of America* 96: 7871–76.
Sherman, D J, T Martin, M Nikolski, C Cayla, J L Souciet, and P Durrens. 2009. “Genolevures: Protein Families and Synteny Among Complete Hemiascomycetous Yeast Proteomes and Genomes.” *Nucleic Acids Res* 37: D550–54.
Shriver, Mark D, Giulia C Kennedy, Esteban J Parra, Heather A Lawson, Vibhor Sonpar, Jing Huang, Joshua M Akey, and Keith W Jones. 2004. “The Genomic Distribution of Population Substructure in Four Populations Using 8,525 Autosomal SNPs..” *Human Genomics* 1 (4). BioMed Central: 274–86. doi:10.1186/1479-7364-1-4-274.
Smith, Gavin J D, Dhanasekaran Vijaykrishna, Justin Bahl, Samantha J Lycett, Michael Worobey, Oliver G Pybus, Siu Kit Ma, et al. 2009. “Origins and Evolutionary Genomics of the 2009 Swine-Origin H1N1 Influenza a Epidemic..” *Nature* 459 (7250). Nature Publishing Group: 1122–25. doi:10.1038/nature08182.
Stamatakis, A. 2006. “RAxML-VI-HPC: Maximum Likelihood-Based Phylogenetic Analyses with Thousands of Taxa and Mixed Models.” *Bioinformatics (Oxford, England)* 22 (21). England: 2688–90.
Stearns, S C, S G Byars, D R Govindaraju, and D Ewbank. 2010. “Measuring Selection in Contemporary Human Populations.” *Nature Reviews.Genetics* 11 (9). England: 611–22. doi:10.1038/nrg2831.
Stoletzki, Nina, and Adam Eyre-Walker. 2007. “Synonymous Codon Usage in Escherichia Coli: Selection for Translational Accuracy..” *Molecular Biology and Evolution* 24 (2). Oxford University Press: 374–81. doi:10.1093/molbev/msl166.
Suzuki, Y. 2006. “Natural Selection on the Influenza Virus Genome.” *Molecular Biology and Evolution* 23 (10). United States: 1902–11.
Swofford, David L. 2003. “{PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods). Version 4.}.” Sinauer associates.
Tajima, F. 1989. “Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism..” *Genetics* 123 (3). Genetics Society of America: 585–95.
Tajima, F. 1993. “Simple Methods for Testing the Molecular Evolutionary Clock Hypothesis..” *Genetics* 135 (2). Genetics Society of America: 599–607.
Tishkoff, S A, F A Reed, A Ranciaro, B F Voight, C C Babbitt, J S Silverman, K Powell, et al. 2007. “Convergent Adaptation of Human Lactase Persistence in Africa and Europe.” *Nature Genetics* 39 (1). United States: 31–40. doi:10.1038/ng1946.
Tuller, T, A Carmi, K Vestsigian, S Navon, Y Dorfan, J Zaborske, T Pan, O Dahan, I Furman, and Y Pilpel. 2010. “An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation.” *Cell* 141: 344–54.
Tuplin, A, D J Evans, and P Simmonds. 2004. “Detailed Mapping of RNA Secondary Structures in Core and NS5B-Encoding Region Sequences of Hepatitis C Virus by RNase Cleavage and Novel Bioinformatic Prediction Methods.” *The Journal of General Virology* 85 (Pt 10). England: 3037–47. doi:10.1099/vir.0.80141-0.

Tuplin, A, J Wood, D J Evans, A H Patel, and P Simmonds. 2002. “Thermodynamic and Phylogenetic Prediction of RNA Secondary Structures in the Coding Region of Hepatitis C Virus.” *RNA (New York, N.Y.)* 8 (6). United States: 824–41.

Tusche, C, L Steinbruck, and A C McHardy. 2012. “Detecting Patches of Protein Sites of Influenza a Viruses Under Positive Selection.” *Molecular Biology and Evolution* 29 (8). United States: 2063–71. doi:10.1093/molbev/mss095.

Van Rossum, G, and F Drake. 2001. “Python Reference Manual.”

Vandivier, Lee E, Rafael Campos, Pavel P Kuksa, Ian M Silverman, Li-San Wang, and Brian D Gregory. 2015. “Chemical Modifications Mark Alternatively Spliced and Uncapped Messenger RNAs in Arabidopsis..” *The Plant Cell* 27 (11). American Society of Plant Biologists: 3024–37. doi:10.1105/tpc.15.00591.

Vitalis, R, K Dawson, and P Boursot. 2001. “Interpretation of Variation Across Marker Loci as Evidence of Selection..” *Genetics* 158 (4). Genetics Society of America: 1811–23.

Vitti, Joseph J, Sharon R Grossman, and Pardis C Sabeti. 2013. “Detecting Natural Selection in Genomic Data..” *Annual Review of Genetics* 47 (1). Annual Reviews: 97–120. doi:10.1146/annurev-genet-111212-133526.

Voight, Benjamin F, Sridhar Kudaravalli, Xiaoquan Wen, and Jonathan K Pritchard. 2006. “A Map of Recent Positive Selection in the Human Genome..” Edited by Laurence Hurst. *PLoS Biol* 4 (3). Public Library of Science: e72. doi:10.1371/journal.pbio.0040072.

Wang, Eric T, Greg Kodama, Pierre Baldi, and Robert K Moyzis. 2006. “Global Landscape of Recent Inferred Darwinian Selection for Homo Sapiens..” *Proceedings of the National Academy of Sciences of the United States of America* 103 (1). National Acad Sciences: 135–40. doi:10.1073/pnas.0509691102.

Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. “RNA-Seq: a Revolutionary Tool for Transcriptomics..” *Nature Reviews.Genetics* 10 (1). Nature Publishing Group: 57–63. doi:10.1038/nrg2484.

Watterson, G A. 1978. “The Homozygosity Test of Neutrality..” *Genetics* 88 (2). Genetics Society of America: 405–17.

Weinberg, David E, Premal Shah, Stephen W Eichhorn, Jeffrey A Hussmann, Joshua B Plotkin, and David P Bartel. 2015. “Improved Ribosome-Footprint and mRNA Measurements Provide Insights Into Dynamics and Regulation of Yeast Translation.” doi:10.1101/021501.

Yang, Z. 2000. “Maximum Likelihood Estimation on Large Phylogenies and Analysis of Adaptive Evolution in Human Influenza Virus A.” *Journal of Molecular Evolution* 51 (5). UNITED STATES: 423–32. doi:10.1007/s002390010105.

Zanini, F, and R A Neher. 2013. “Quantifying Selection Against Synonymous Mutations in HIV-1 Env Evolution.” *Journal of Virology* 87 (21). United States: 11843–50. doi:10.1128/JVI.01529-13.

Zeng, Kai, Suhua Shi, and Chung-I Wu. 2007. “Compound Tests for the Detection of Hitchhiking Under Positive Selection..” *Molecular Biology and Evolution* 24 (8): 1898–1908. doi:10.1093/molbev/msm119.

Zeng, Kai, Yun-Xin Fu, Suhua Shi, and Chung-I Wu. 2006. “Statistical Tests for Detecting Positive Selection by Utilizing High-Frequency Variants..” *Genetics* 174 (3). Genetics Society of America: 1431–39. doi:10.1534/genetics.106.061432.

Zhang, Chun, Dione K Bailey, Tarif Awad, Guoying Liu, Guoliang Xing, Manqiu Cao, Venu Valmeekam, et al. 2006. “A Whole Genome Long-Range Haplotype (WGLRH) Test for Detecting Imprints of Positive Selection in Human Populations..” *Bioinformatics (Oxford, England)* 22 (17): 2122–28. doi:10.1093/bioinformatics/btl365.

Zhao, Xiaqing, Alan O Bergland, Emily L Behrman, Brian D Gregory, Dmitri A Petrov, and Paul S Schmidt. 2015. “Global Transcriptional Profiling of Diapause and Climatic Adaptation in Drosophila Melanogaster..” *Molecular Biology and Evolution*, November. Oxford University Press, msv263. doi:10.1093/molbev/msv263.

Zhao, Yunjie, Yangyu Huang, Zhou Gong, Yanjie Wang, Jianfen Man, and Yi Xiao. 2012. “Automated and Fast Building of Three-Dimensional RNA Structures..” *Scientific Reports* 2. Nature Publishing Group: 734. doi:10.1038/srep00734.

Zhou, T, and C O Wilke. 2011. “Reduced Stability of mRNA Secondary Structure Near the Translation-Initiation Site in dsDNA Viruses.” *BMC Evolutionary Biology* 11: 59.

Zinshteyn, Boris, and Wendy V Gilbert. 2013. “Loss of a Conserved tRNA Anticodon Modification Perturbs Cellular Signaling..” Edited by Gregory P Copenhaver. *PLoS Genetics* 9 (8). Public Library of Science: e1003675. doi:10.1371/journal.pgen.1003675.

Zuker, M. 1989. “On Finding All Suboptimal Foldings of an RNA Molecule.” *Science (New York, N.Y.)* 244 (4900). UNITED STATES: 48–52.