Identifying protein-coding genes and synonymous constraint elements using phylogenetic codon models

by

Michael F. Lin

S.B. (2005), M.Eng. (2006), Massachusetts Institute of Technology

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2012

Signature of Author.....

Department of Electrical Engineering and Computer Science

January 30, 2012

Certified by.....

Manolis Kellis

Associate Professor of Electrical Engineering and Computer Science

Thesis Supervisor

Accepted by.....

Professor Leslie A. Kolodziejski

Chair of the Committee on Graduate Students

Identifying protein-coding genes and synonymous constraint elements using phylogenetic codon models

by

Michael F. Lin

Submitted to the Department of Electrical Engineering and Computer Science
on January 30, 2012 in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy in Computer Science

Abstract

We develop novel methods for comparative genomics analysis of protein-coding genes using phylogenetic codon models, in pursuit of two main lines of biological investigation:

First, we develop PhyloCSF, an algorithm based on empirical phylogenetic codon models to distinguish protein-coding and non-coding regions in multi-species genome alignments. We benchmark PhyloCSF to show that it outperforms other methods, and we apply it to discover novel genes and analyze existing gene annotations in the human, mouse, zebrafish, fruitfly and fungal genomes. We use our predictions to revise the canonical annotations of these genomes in collaboration with GENCODE, FlyBase and other curators. We also reveal a surprisingly widespread mechanism of stop codon readthrough in the fruitfly genome, with additional examples found in mammals. Our work contributes to more-complete gene catalogs and sheds light on fascinating unusual gene structures in the human and other eukaryotic genomes.

Second, we design phylogenetic codon models to detect evolutionary constraint at synonymous sites of mammalian genes. These sites are frequently assumed to evolve neutrally, but increased conservation would suggest they encode additional information overlapping the protein-coding sequence. We produce the first high-resolution catalog of individual human coding regions showing highly conserved synonymous sites across mammals, which we call Synonymous Constraint Elements (SCEs). We locate more than 10,000 SCEs, covering $\sim$2% of synonymous sites, and found within over one-quarter of all human genes. We present evidence that they indeed encode numerous overlapping biological functions, including splicing- and translation-associated regulatory motifs, microRNA target sites, RNA secondary structures, dual-coding genes, and developmental enhancers. We also develop a lineage-specific test which we use to study the evolutionary history of SCEs, and a Bayesian framework that further increases the resolution with which we can identify them. Our methods and datasets can inform future studies on mammalian gene structures, human disease associations, and personal genome interpretation.

Thesis Supervisor: Manolis Kellis
Title: Associate Professor, Department of Electrical Engineering and Computer Science

Contents

Chapter 1

Summary

The sequencing of the human genome, completed ten years ago, was a landmark event in the biomedical sciences – opening whole new avenues to study human biology, revolutionizing our understanding of our own evolution and population history, and creating the potential to drive revolutionary treatments for many diseases (Green et al., 2011; Lander, 2011). Encoded within the human genome are blueprints for the molecular parts that make up our living cells, a regulatory code controlling when and where they are deployed, and the instructions that govern the development of a single fertilized cell into an adult human being. But while this information surely resides within the genome, it is not organized in any neat, immediately decipherable way. Rather, it is the unkempt result of tumultuous trial-and-error over billions of years of evolution. To fulfill the promise of genomics for science and medicine in the 21st century, therefore, it is crucial to construct an accurate and comprehensive catalog of the specific DNA sequences in the genome encoding different biological functions (The ENCODE Project Consortium, 2011).

Computational comparative genomics provides a powerful paradigm for studying genomes at this level. Since natural selection tends to preserve important biological functions over evolutionary time, analyzing patterns of conservation and divergence between the genomes of related species can distinguish functional and non-functional DNA sequences. Furthermore, it can also predict the specific biological functions of conserved sequences, based on *evolutionary signatures* characteristic to different classes of functional elements (Stark et al., 2007). Comparative genomics analysis of species related at appropriate evolutionary distances, such as multiple mammals to study the human genome, can thus contribute to their genome annotations and provide many insights into the function, regulation and evolution of their genomes (Lindblad-Toh et al., 2011).

In this thesis, we will develop and apply novel methods for comparative genomics analysis of protein-coding DNA sequences, in pursuit of two main lines of biological investiga-

tion. First, we will present an algorithm for distinguishing protein-coding and non-coding regions based on cross-species alignments. We show that this algorithm outperforms other methods, and also present numerous applications of the method to discover new genes, revise existing gene annotations, and reveal unusual gene structures in the genomes of several scientifically important species – including human. Second, we will study the unusual phenomenon of evolutionary conservation of synonymous codon sites, due to natural selection on overlapping functions embedded within protein-coding sequences. We develop methods to identify and characterize short coding regions with highly conserved synonymous sites in mammalian genes, which we call *Synonymous Constraint Elements* (SCEs), and use them to produce the first genome-wide, high-resolution annotation of SCEs in the human genome.

These investigations share a common methodological basis in *phylogenetic codon models*, which model the evolution of individual codon sites as a random substitution process unfolding on the phylogenetic tree relating the species under analysis (Anisimova and Kosiol, 2008; Delport et al., 2008). Our research includes formulating appropriate parameterizations of such codon models, estimating their parameters from training data, and applying them to produce well-founded statistical answers to our biological questions. We also utilize machine learning methods such as hidden Markov models and conditional random fields, and a variety of frequentist and Bayesian statistical hypothesis tests.

Overall, our work contributes to a more complete gene catalog in the human genome and those of other species, and a deeper understanding of the expression, regulation and evolution of mammalian gene structures. More generally, it also illustrates the development of novel, rigorous methods for genome-wide analyses, and their application to produce both specific biological insights and datasets that can support many future computational genomics studies.

Acknowledgments

Our work has relied on the advice, support and efforts of many collaborators. Irwin Jungreis, Pouya Kheradpour, Clara Chan, Stefan Washietl, Brian Parker, and Jakob Pedersen are immediate co-authors of recent papers based on the contributions described in this thesis (Lin et al., 2011a,b). We are grateful for helpful discussions and collaborations with Matt Rasmussen, Manuel Garber, Mitch Guttman, Luke Ward, Jason Ernst, Loyal Goff, Andrea Pauli, Madeline Crosby, and Tara Alexander. We have also benefited from the advice of numerous expert scientists and professors, including John Rinn, Kerstin Lindblad-Toh, Bill Gelbart, Sue Celniker, Eric Lander, Katie Pollard, Nick Goldman, Robb

Krumlauf, and Bob Berwick. Major parts of our work have taken place within the context of large community projects such as ENCODE (especially GENCODE), modENCODE, the 12 Drosophila Genomes Consortium, and the 2X mammals project. Lastly, Manolis Kellis has shaped each part of this work in many ways, from initial conception, to the design of specific analyses, the interpretation of results, all the way through final presentation of figures and text. Funding for our work was provided by the National Institutes of Health (U54 HG004555-01) and the National Science Foundation (DBI 0644282).

Chapter 2

Background

2.1 DNA and genomes

As computer scientists, we can understand a *DNA sequence* as a string over an alphabet of four characters, A, G, C, and T, representing four different *nucleotides* that can form a lengthy molecular chain: adenine, guanine, cytosine, and thymine. Within the nucleus of a eukaryotic cell, the DNA sequences are organized into numerous *chromosomes*, lengthy individual DNA molecules wound up compactly with a variety of "packaging" material. The *genome* of an organism is the complete set of DNA sequences that defines its genetic identity – typically, all the chromosomal DNA. Although every organism in a species has its own unique genome, insofar as there is genetic variation within the population, this variation is small enough that it is meaningful to study the genome of a species. In particular, the human genome consists of approximately three billion nucleotides on twenty-four distinct chromosomes – the autosomal chromosomes 1-22 and the sex-determining X and Y chromosomes.

Genomes are a mosaic of many different types of biologically functional elements, only some of which are currently understood. The most well-studied functional elements are *protein-coding genes*, which are instructions for synthesizing *proteins*, the building blocks of all cellular machines. Other genes code not for proteins, but rather molecules made of *RNA*, a chemical cousin of DNA that can perform catalytic and regulatory functions in the cell. Also contained in the genome are cis-*regulatory elements* which, in concert with additional cellular machinery, control when genes are transcribed, the first step in protein synthesis. Some genome sequence is comprised of highly repetitive *heterochromatin* that serves as structural scaffolding for the chromosome (DNA molecule) in the cell. And especially in vertebrate species, genomes contain large amounts of seemingly nonfunctional sequence, left over as historical artifacts of evolution. Some of these include *pseudo-*

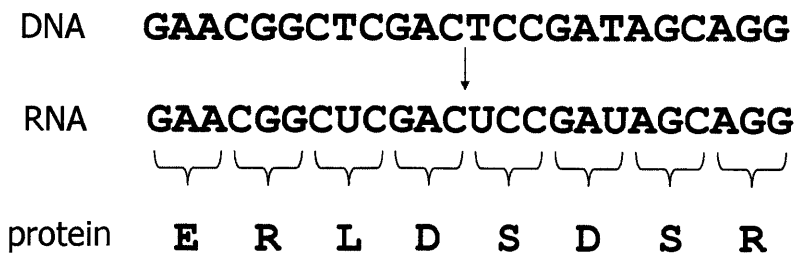


Figure 2-1: Information flow from a protein-coding gene. Protein-coding DNA sequence in the genome is transcribed into RNA, a chemical cousin of DNA. The RNA nucleotides are translated in groups of three (codons) into a sequence of amino acids such as aspartic acid (D) and serine (S), which form the primary structure of a protein.

genes, sequences that were once protein-coding genes but have since ceased to function and degraded, and *transposons*, self-replicating sequences that increase in number over evolutionary time but in most cases serve no useful purpose to the organism.

2.2 Protein-coding genes and the genetic code

In this thesis, we focus on protein-coding genes. A protein-coding gene is a region of the genome that provides other cellular machinery with instructions for synthesizing a specific protein. When protein synthesis begins, the gene, which is an otherwise undistinguished substring in the genome, is copied into RNA, a chemical cousin of DNA, in a process called *transcription*. The transcript is then *translated* into a chain of amino acids, the basic chemical building blocks of proteins. The chain of amino acids, called a *polypeptide*, later folds up in three dimensions to form a functional protein.

With very few exceptions, twenty different amino acids are the building blocks of proteins across all kingdoms of life. The genetic code defines the mapping between sequences in the RNA transcript, and by extension the genome, and amino acids. Every three nucleotides in the transcript specify a single amino acid in the polypeptide. A nucleotide triplet is called a *codon* in this context. For example, the codon TCA corresponds to the amino acid serine. During translation, the cellular machinery reads each codon in the transcript one-by-one and attaches the corresponding amino acid to the nascent polypeptide (Figure 2-1). The genetic code itself, the mapping between codons and amino acids, is nearly universal among known forms of life.

Since there are four DNA nucleotides and three nucleotides in a codon, there are $4^3 = 64$ possible codons, used to specify only twenty different amino acids. The genetic code is *degenerate*: more than one codon may code for a single amino acid. A few amino

acids (such as methionine) have only one corresponding codon, while several (such as serine and alanine) have as many as four *synonymous* codons. There are four special codons that control translation: one *start codon*, which indicates where in the transcript translation should begin, and three *stop codons*, which indicate where translation ends. The start codon, ATG, also codes for methionine when it occurs in the coding region of the transcript. The stop codons, TAA, TAG, and TGA, do not code for any amino acids under normal circumstances, but rather cause translation to terminate.

The transcript of a protein-coding gene typically contains an upstream *untranslated region* (UTR), followed by the actual protein-coding sequence, followed by a downstream UTR. The protein-coding sequence consists of the start codon, followed by a number of codons, followed by one of the three stop codons. The portion of the sequence with no stop codons is referred to as an *open reading frame* (ORF).

2.3 Introns, exons and splicing

In the genomes of simple organisms such as bacteria, the RNA transcript of a gene directly corresponds to the sequence in the genome. However, in higher forms of life, the relationship of the final transcript sequence to the genomic sequence is more complicated. In these genomes, including all the species studied in this thesis, the protein-coding sequence may not occur as a contiguous open reading frame in the genome, but rather may be partitioned into several pieces separated by non-coding sequence. The coding portions of these genes are called *exons*, and the intervening non-coding portions are called introns. When the transcript is copied from the genome, it subsequently undergoes *splicing*, where the introns are cut out from the transcript to form a contiguous open reading frame. Specifically, processing machinery in the cell recognizes *splice sites* in the transcript, which are short sequences (8-20 nucleotides in length) that flank the introns, and recruit enzymes that cut the introns out of the transcript. There are two distinct types of splice sites: *acceptor sites*, which occur at the beginning of an exon, and *donor sites*, which occur at the end of an exon. The transcript after splicing is called the *messenger RNA* (mRNA), while before splicing it is called the *pre-mRNA* (Figure 2-2).

There are a several biological reasons for splicing. Most notably, it allows for *alternative splicing*: the selective inclusion or exclusion of individual exons during transcript processing. Alternative splicing allows one gene to encode several different proteins, and is thought to be a major progenitor of structural and functional protein diversity from a comparatively limited number of protein-coding genes.

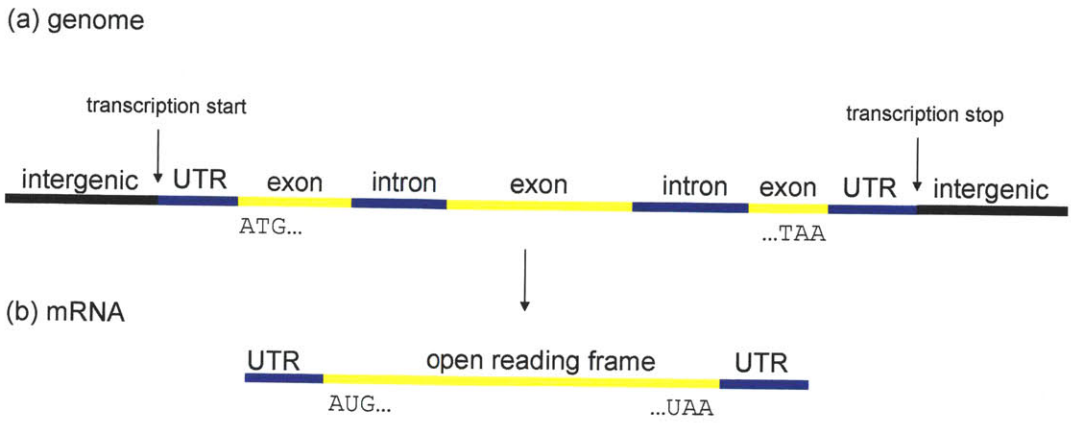


Figure 2-2: Structure of a eukaryotic protein-coding gene. (a) Transcription from the genomic DNA into an RNA transcript starts and ends at certain points in the genome. Introns in the RNA transcript are spliced out to form (b) the final mRNA. The open reading frame (ORF) between the ATG start codon and the stop codon (in this case TAA) codes for the protein sequence as shown in Figure 2-1. The ORF is flanked by untranslated regions (UTR).

2.4 Gene annotations and transcriptome sequencing

Extensive research and manpower has gone into producing *gene annotations* for the human genome, marking up the locations of protein-coding genes and their constituent exons on the chromosomal sequences. The best annotation databases combine several types of evidence to locate and define gene structures, including known protein sequences, experimental data from mRNA transcripts, computational data including comparative genomics analyses, and sometimes extensive manual curation. High-quality annotations for the ~22,000 protein-coding genes in the human genome are available from a variety of sources (Pruitt et al., 2002; Harrow et al., 2006; Hsu et al., 2006; Pruitt et al., 2009; Flicek et al., 2010) and are generally accurate and comprehensive enough for routine downstream use in genome-wide analyses – although further refinement continues. High-quality gene annotations are also available for many important model organisms in the biological sciences, such as yeast and the fruitfly.

One recent major advance in gene annotation was the advent of high-throughput transcriptome sequencing (*RNA-seq*), an experimental protocol in which the sequences of virtually all mRNAs expressed in a sample of cells can be determined quickly and inexpensively (under two weeks and several thousand dollars), albeit with somewhat noisy and error-prone results (Ozsolak and Milos, 2011). RNA-seq has provided evidence for many novel genes in the human genome (Guttman et al., 2010; The ENCODE Project Consortium, 2011; Cabili et al., 2011). Relatively few of these seem to be protein-coding,

however, reflecting the extensive earlier efforts to annotate protein-coding genes. RNA-seq is also providing detailed structures for many novel alternative splicing isoforms of known genes, both coding and non-coding.

2.5 Comparative genomics

The genomes of modern species – many of which have now been sequenced – are a product of their evolution over millions of years. Thus, comparative analysis of the genomes of related species can yield insights into their evolution, and conversely, our knowledge of evolutionary theory can guide the comparative analysis of genomes. As we will explore in this thesis, *comparative genomics* provides a powerful way to locate and study functional elements, such as protein-coding genes, in the genomes analyzed. Underlying this is the observation that functional elements evolve subject to natural selection, whereas nonfunctional sequences undergo essentially random mutation. Comparative analysis of the genomes of related species, which evolve largely independently since their divergence, can reveal selection on functional elements. Most apparently, functional elements generally show higher sequence identity among related genomes than nonfunctional sequence, since any individual mutation in a functional element is more likely to reduce the fitness of the organism than a mutation in a nonfunctional element. More specifically than indicating that a sequence *is* functional, however, comparative analysis can reveal distinctive evolutionary signatures that are clues to *how* the sequence is functioning. This thesis will investigate such signatures within protein-coding genes. However, before it is possible to compare genomes at such a fine granularity, several major challenges must be addressed.

Sequence alignment. A classic problem in computational biology is sequence alignment. Given two or more related biological sequences (for example, DNA sequences of related genes in two different species), a sequence alignment algorithm computes the optimal pattern of insertions, deletions, and nucleotide substitutions in each sequence in order to match up the parts of the sequences that are similar to one another (Figure 2-3). Sequence alignment among biological sequences that are known to be related is a reasonably well-understood algorithmic problem, and there are a variety of tools available for this purpose.

Genome alignment. Sequence alignment is not by itself sufficient for comparative genomics. To understand why, consider that while the human genome has 24 distinct chromosomes, mice have 21, and dogs have 39. While the DNA sequence for each of these chromosomes is known, no simple mapping exists between the chromosomes of the different species. If a certain chromosome in human contains a certain set of genes, those

```
human TCATTTCACTCATAGGTTTAT---ATTTCTCAGAGTTCTTTGAGCTAAA
  dog TCTTTTCAC----AGGATTATCCAAATTGC-AAAGTTCATTGAGCAGAG
mouse TTCTTTCAC----GGTTTATTAGGATTCCCAAAAGTGCTTTAAACAAAA
      *  ******    *** ****    *** *  * *** * ** * *  *
```

Figure 2-3: Example of a sequence alignment. A sequence alignment algorithm computes the optimal placement of gaps (indicated by dashes) in order to line up several biological sequences, highlighting their similarities and differences. The stars indicate perfect matches down the corresponding column of the alignment. Gaps are also referred to as "indels" referring to the inference that, during evolution, some sequence was inserted or deleted at that point in the sequence.

genes might be dispersed across several different chromosomes in other species, and vice versa. Moreover, individual genes, and even whole chromosomes, can be duplicated or lost during evolution. Thus, in addition to a tool for nucleotide-level sequence alignment, it is necessary to have a methodology for *genome alignment*, determining at a large scale which parts of related genomes correspond to each other, in order to comparatively analyze those genomes. The effectiveness of any fine-grained comparative genomic analysis is strongly dependent on the completeness and quality of the genome alignments for the species under analysis. Genome alignment, however, remains a research area with many unresolved problems, especially on large genomes such as those of the mammals. Existing tools are widely used in comparative genomics, but are known to have certain limitations (Blanchette, 2007).

Choosing informant species. Finally, it is crucial to carefully choose the species to compare, so that their genomes are far enough diverged that the conservation of functional elements, such as genes, is measurable against the background of random mutation, but not so far that the desired elements are not well-conserved across those species. For example, the genome of the chimpanzee is so similar to that of a human that a comparative analysis of the two is virtually uninformative for finding conserved genes. On the other hand, human and yeast are so distant that probably only a modest subset of their genes are even shared.

In the case of the human genome, it is thought that eutherian mammals outside of primates, such as dog and mouse, are appropriate candidates for comparative genomic analysis (Lindblad-Toh et al., 2011). The vast majority of human genes should be conserved in these species, while the nonfunctional sequence has diverged substantially more. The common ancestor of these species is thought to have lived about 125 million years ago.

2.6 Phylogenetic codon models

Equipped with genome alignments, gene annotations and other relevant data, we now require a quantitative methodological framework in which to carry out comparative genomics analysis of protein-coding genes. In this thesis we use *phylogenetic codon models*, which are statistical models of the evolution of a single codon within a protein-coding gene, and specifically of the process of *codon substitution* during evolution. The key types of codon substitutions are illustrated in Figure 2-4.

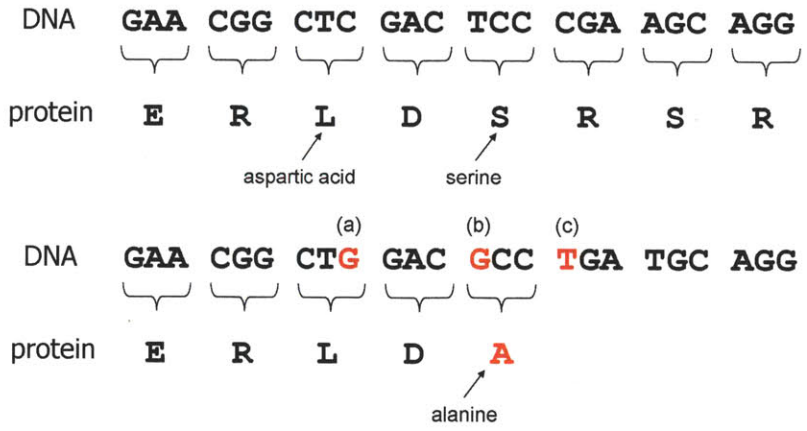


Figure 2-4: Possible effects of codon substitutions in protein-coding sequences. (a) synonymous substitutions lead to no change in the protein sequence because the new codon corresponds to the same amino acid as the old codon. (b) missense substitutions cause a different amino acid to be incorporated at the corresponding position. Some amino acid substitutions are more favorable than others due to biochemical similarities and differences between the different amino acids. (c) nonsense substitutions change a sense codon into a stop codon, truncating the protein sequence.

Intuitively, these different types of substitutions are likely to have differing fitness effects on the resulting protein and the organism as a whole. A nonsense substitution, for example, truncates the protein and probably renders it useless if it occurs early in the coding sequence. Some missense substitutions could also severely disrupt the structure or function of the protein product, while others – for example, a substitution to a chemically similar amino acid – might have only a modest effect. Synonymous codon substitutions are often assumed to be evolutionarily neutral, as they are likely to have negligible fitness effects unless the protein-coding nucleotide sequence encodes additional information – an unusual phenomenon which we will investigate in this thesis.

Since different codon substitutions have different fitness effects, evolution will preserve them at different rates over time. Phylogenetic codon models provide a rigorous mathe-

matical framework to capture these rates and estimate them from data, while accounting for the evolutionary relationships among the species being studied and the inherent uncertainty about their extinct ancestors. They have been reviewed recently in Anisimova and Kosiol (2008) and Delport et al. (2008), and we present them briefly here.

2.6.1 Generative process

The unit of observation for a phylogenetic codon model consists of an individual codon (three nucleotides) within a gene of the *reference* species (the main species of interest), and its alignment with corresponding codons in the informant genomes, together called a *codon alignment site*. These codons might be conserved, or the site may exhibit one or more codon substitutions between the reference and informants, or among the informants. A phylogenetic codon model posits that the aligned codons of an individual site are the result of a stochastic process of codon substitution, beginning from a common ancestral codon which then evolves on the *phylogenetic tree* relating the species under analysis. Starting from a single common ancestor, the phylogenetic tree specifies a sequence of binary *speciations* or branchings resulting in the several modern species, where evolution is assumed to occur independently within the two lineages resulting from a speciation. Each *leaf* of this tree thus represents an extant (modern) species, an internal *node* represents an extinct ancestral species, and the *root* represents the extinct common ancestor of all the species under analysis. Figure 2-5 shows a simple phylogenetic tree of three species, and why it is important to use a detailed model of the phylogeny, rather than treating the codons in the alignment as independent observations.

The substitution process can be further characterized by specifying the probability of any codon substitution $i \to j$ along each branch of the tree, with no substitution $(i \to i)$ among the possibilities. These probabilities can be tabulated in *substitution matrices* $\mathbf{P}^{(b)}$, where $\mathbf{P}^{(b)}_{ij}$ is $\Pr(j|i)$, the probability that the child at the end of branch b has codon j, conditioned on its parent having i. Branches corresponding to close phylogenetic relationships tend to have substitution matrices with high probabilities of no substitution, along the diagonal, while more distant relationships lead to relatively larger entries off the diagonal. The relative magnitude of an off-diagonal entry reflects the evolutionary favorability of the corresponding codon substitution.

Finally, there is also a prior probability distribution over which codon is present in root node. (The model may include all 64 possible codons, but often excludes the three stop codons since they do not appear within open reading frames.)

We can now specify a stochastic process to generate an alignment site, which is essentially a Bayesian network for the phylogeny. First, choose a root codon x according to

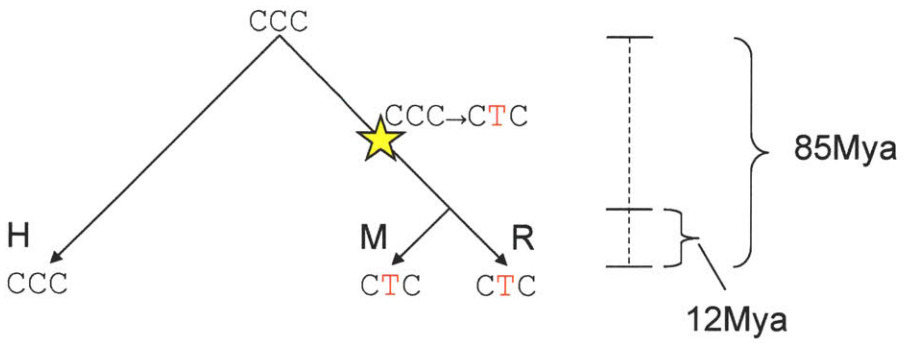


Figure 2-5: Several observed codon substitutions may be due to a single evolutionary event. The diagram shows the phylogenetic tree of human (H), mouse (M), and rat (R). In this scenario, an ancestral species had the codon CCC. A single substitution to CTC occurs on the rodent lineage before mouse and rat diverge. Using the human sequence as the reference, we observe codon substitutions in both mouse and rat informants, but these actually arise from a single evolutionary event. Without modeling the phylogeny, therefore, we would tend to "double count" evidence in the alignment. (Given only the observed codons, it cannot be ruled out that two separate substitution events occurred in both mouse and rat since they diverged, but this is much less likely than the depicted single event.)

the prior distribution. Then, generate the codon y in one of the root's immediate children, by sampling from the conditional distribution represented in row x of the substitution matrix for that branch. Repeat this for the root's other child, using the substitution matrix for the respective branch, independently of the first branch (conditional on x). Then apply this same procedure recursively to the root's children to generate the grandchildren, and so on to generate codons for each node of the tree. Finally, read the alignment site from the leaves of the tree.

2.6.2 Likelihood calculation

Given an assignment of codons to each node in the tree (including the root, internal nodes, and leaves), we can easily compute the joint probability of this configuration arising from the above process, by multiplying the prior probability of the root codon with the probability of the event on each branch implied by the configuration. However, this is not directly useful because our genome alignments only give us the assignments of the leaves, as the internal nodes represent extinct species.

Instead, we can determine the marginal probability of an assignment of codons to the leaves only, by summing the joint probability over all possible assignments of the internal

nodes. In a model for 61 codons and n extant species, there are 61^{n-1} such assignments to sum over. Fortunately, the belief propagation algorithm provides an efficient dynamic programming solution to perform this marginalization. In the field of phylogeny, this is usually known as *Felsenstein's algorithm* (Felsenstein, 2004). This marginalization approach also provides a natural way to deal with missing data – for example, when a gene cannot be aligned to certain informant genomes, or when codons have been deleted in informant genomes – by effectively eliminating the corresponding variables from the calculation.

In a sequence alignment of several codons (e.g. an exon or a complete ORF), each codon site is considered independent and identically distributed, i.e. their probabilities can be computed individually as above and multiplied together.

2.6.3 Parameterizations

As described thus far, a model for 61 codons and n extant species (leaves) requires $61 \cdot 60 \cdot (2n-2) + 60$ free parameters – all entries of the substitution matrices on all branches, and the prior distribution over root codons. To analyze dozens of species, it is impractical to specify this many parameters *a priori*, and also infeasible to reliably estimate all of them, even given genome-wide training data. It is therefore necessary to constrain the parameterization, by stating each parameter algebraically in terms of a smaller number of parameters. A standard approach with phylogenetic codon models is to constrain the substitution matrices $\mathbf{P}^{(b)}$ to arise from a *continuous-time Markov process* (CTMP) with 61 states representing the codons, and an instantaneous *rate matrix* common to all branches. Specifically, the substitution matrix for each branch ($\mathbf{P}^{(b)}$) is computed from the CTMP rate matrix ($\mathbf{Q}$) and a scalar *branch length* parameter (t_b) according to a matrix exponential formula: $\mathbf{P}^{(b)} = \exp(\mathbf{Q}t_b)$. The off-diagonal entries of $\mathbf{Q}$, q_{ij}, are nonnegative free parameters, and the diagonal entries are $q_{ii} = -\sum_{k \neq i} q_{ik}$. An arbitrary degree of freedom between the absolute branch lengths and the absolute rate matrix entries is conventionally eliminated by scaling the rate matrix so that $\frac{1}{61}\sum_i q_{ii} = 1$. Loosely speaking, the branch lengths reflect elapsed times between speciations in the phylogenetic tree, while $\mathbf{Q}$, which is shared throughout the tree, specifies the rate of each possible codon substitution per unit branch length. This model has $61 \cdot 60 + (2n-2) - 1 + 60 - 1$ free parameters ($\mathbf{Q}$, the branch lengths, and the prior distribution, minus the arbitrary degree of freedom).

The rate matrix is usually further constrained to represent a *reversible* CTMP. In this case, $\mathbf{Q}$ can be decomposed so that its off-diagonal entries are given by $q_{ij} = \pi_j s_{ij}$, where π_j are frequencies of the codons representing the stationary distribution of the

reversible process, and s_{ij} are symmetric "exchangeabilities" of the codons ($s_{ij} = s_{ji}$). The stationary distribution can also be used as the prior distribution over ancestral codons, resulting in $\frac{61 \cdot 60}{2} + (2n - 2) - 1$ parameters.

The stationary distribution and exchangeabilities are often parameterized even further, and many such parameterizations have been explored – the simplest in common use with only $2n + 9$ parameters (the MG94–F3×4 model). Such models can be estimated from the alignment of a single gene, and their individual parameters often have an insightful biological interpretation in terms of the fundamental evolutionary forces of mutation and natural selection. One of the most important is ω, also known as d_N/d_S, which is a scale factor on non-synonymous rates relative to synonymous rates. A reduced rate of non-synonymous substitutions ($\omega < 1$) suggests *purifying selection* on the encoded amino acid; an accelerated rate ($\omega > 1$) suggests *diversifying selection*; and a similar rate ($\omega \approx 1$) suggests neutral evolution (assuming synonymous substitutions are neutral).

2.6.4 Parameter estimation

However the codon model is parameterized, it is usually desirable to estimate the parameters from data – often an individual ORF, and sometimes a sample of sites from all ORFs genome-wide. Maximum likelihood and Bayesian parameter estimation strategies are both widely used for codon models. If the number of parameters and the dataset size are small (e.g. a dozen parameters for a single ORF), maximum likelihood estimates can be computed by cyclic coordinate ascent. Larger parameterizations and datasets may demand gradient ascent or expectation-maximization approaches, sometimes with fast approximations for certain steps in these procedures. The dependence of the likelihood on the parameters through the matrix exponential necessitates the use of spectral decompositions of the rate and substitution matrices in these methods (Arvestad and Bruno, 1997; Holmes and Rubin, 2002; Siepel and Haussler, 2004; Hobolth and Jensen, 2005; Kloosterman et al., 2006; Kiryu, 2011). Bayesian parameter estimation may require numerical integration or Markov chain Monte Carlo approaches. This thesis will utilize several of these parameter estimation strategies.

2.6.5 Limitations

The phylogenetic codon models developed in this thesis rest upon a widely-used and historically-successful set of modeling assumptions that balance the goals of biological plausibility and accuracy with the constraints of computational tractability and practicality of parameter estimation. Nonetheless, we will briefly mention some of the inherent

limitations of these assumptions, many of which are the subject of ongoing research (Anisimova and Kosiol, 2008; Delport et al., 2008).

- The assumption that individual codon sites are statistically independent is mathematically convenient, but cannot possibly capture molecular interactions between amino acids in the protein, which is typically folded into an intricate three-dimensional structure.

- We model codon evolution as a stationary, reversible continuous-time Markov process that is homogeneous on the phylogenetic tree; that is, that the instantaneous rates of random codon substitution remained fixed during the evolution of the extant species and their ancestors. This cannot reflect the varying environments and selective pressures these lineages undoubtedly experienced over many millions of years of evolution. (Towards the end of this thesis, we do consider a specific relaxation of the homogeneity assumption, to discover lineage-specific changes in selective pressures on synonymous sites.)

- We do not explicitly model insertions and deletions within the nucleotide sequence during evolution, which arise frequently over the evolutionary distances we consider. Our calculations marginalize over such events, effectively ignoring them, but in fact they can yield powerful evolutionary signatures that are largely orthogonal to codon substitutions.

- We assume a fixed phylogenetic tree topology in our analyses of individual genes, corresponding to the species tree topology. In fact however, the most appropriate tree topology can vary from gene to gene, due to effects such as gene duplication, incomplete lineage sorting and horizontal gene transfer. At the same time, correct determination of the species tree topology can be complicated by long branch attraction or homoplasy (convergent evolution).

Chapter 3

PhyloCSF: A comparative genomics method to distinguish protein-coding and non-coding regions

3.1 Introduction

High-throughput transcriptome sequencing (RNA-seq) is yielding precise structures for novel transcripts in many species, including mammals (Guttman et al., 2010). Accurate computational methods are needed to classify these transcripts and the corresponding genomic exons as protein-coding or non-coding, even if the transcript models are incomplete or if they only reveal novel exons of already-known genes. In addition to classifying novel transcript models, such methods also have applications in evaluating and revising existing gene annotations (Kellis et al., 2003; Lin et al., 2007; Clamp et al., 2007; Butler et al., 2009; Pruitt et al., 2009), and as input features for *de novo* gene structure predictors (Brent, 2008; Alioto and Guig, 2009). We have previously (Lin et al., 2008a) compared numerous methods for determining whether an exon-length nucleotide sequence is likely to be protein-coding or non-coding, including single-sequence metrics that analyze the genome of interest only, and comparative genomics metrics that use alignments of orthologous regions in the genomes of related species.

Among the comparative methods benchmarked in our previous work, one of our original contributions was the Codon Substitution Frequencies (CSF) metric, which assigns a score to each codon substitution observed in the input alignment based on the relative frequency of that substitution in known coding and non-coding regions. We showed that CSF is highly effective, performing competitively with a phylogenetic modeling approach with much less computational expense, and indeed we have applied it successfully in flies (Stark

et al., 2007; Lin et al., 2007), fungi (Butler et al., 2009), and mammals (Clamp et al., 2007; Guttman et al., 2009, 2010). However, as discussed in our previous study, CSF has certain drawbacks arising from its *ad hoc* scheme for combining evidence from multiple species. For example, it makes only partial use of the evidence available in a multi-species alignment, and it produces a score lacking a precise theoretical interpretation, meaningful only relative to its empirical distributions in known coding and non-coding regions.

Here we introduce a rigorous reformulation of CSF, which frames the evaluation of a given alignment as a statistical model comparison problem, choosing between phylogenetic models estimated from known coding and non-coding regions as the best explanation for the alignment. This new "PhyloCSF" method fully leverages multiple alignments in a phylogenetic framework, produces meaningful likelihood ratios as its output, and rests upon a sound theoretical foundation for statistical model comparison. Benchmarking on the classification datasets from our original study, we show that PhyloCSF outperforms all of the other methods we had previously considered.

PhyloCSF is applicable for assessing the coding potential of transcript models or individual exons in an assembled genome that can be aligned to one or more informant genomes at appropriate phylogenetic distances. To estimate parameters in the underlying statistical models, the approach also requires that the genome of interest, or one of the informant genomes, have existing coding gene annotations of reasonably good quality. We describe several initial applications of the method in such settings, which illustrate how it can contribute to new genome annotation strategies based on RNA-seq.

3.2 Approach

PhyloCSF is based on the well-established theoretical framework for statistical phylogenetic model comparison. In this context, phylogenetic models are generative probabilistic models that produce alignments of molecular sequences, based on a prior distribution over a common ancestral sequence, the topology and branch lengths of a phylogenetic tree relating the descendants, and a substitution process along each branch giving the rates (per unit branch length) at which each character changes to any other. In phylogenetic model comparison, we wish to choose between two competing models as the better explanation for a given alignment. A standard approach is to decide based on the *likelihood ratio* between the two models, which quantifies how much more probable the alignment is under one model than under the other. This general approach has been used to explore many different aspects of the evolution of protein-coding genes, as recently reviewed in Anisimova and Kosiol (2008) and Delport et al. (2008).

To distinguish coding and non-coding regions, we design one phylogenetic model to represent the evolution of codons in protein-coding genes, and another to represent the evolution of nucleotide triplet sites in non-coding regions. These models may have one or more parameters θ that adjust them to the genomic region of interest, e.g. the neutral substitution rate or G+C content. To analyze a given alignment A of extant sequences, we first determine the probability of the alignment under the maximum likelihood estimate (MLE) of the parameters for the coding model, $p_C = \max_{\theta_C} \Pr(A|\,\text{Coding}, \theta_C)$. We similarly estimate the alignment's probability under the non-coding model, $p_N = \max_{\theta_N} \Pr(A|\,\text{Non-coding}, \theta_N)$. Finally, we decide if the alignment is more likely to represent a protein-coding region or a non-coding region based on the log-likelihood ratio $\Lambda = \log \frac{p_C}{p_N}$. The cutoff can be chosen to achieve a certain level of statistical significance, based on known asymptotic convergence properties of the log-likelihood ratio statistic (Whelan and Goldman, 1999; Ota et al., 2000; Vuong, 1989), or it can be chosen empirically based on classification performance in a test set; we use the latter strategy in this work.

The d_N/d_S test

A standard method for detecting purifying selection on protein-coding sequences is to test for evidence that non-synonymous substitutions occur at a slower rate than synonymous substitutions. In the widely used PAML implementation of this test (Yang and Nielsen, 1998; Yang, 2007), the vector of codon frequencies π and the ratio of transition to transversion rates κ are the only parameters used to determine all triplet substitution rates in the background/non-coding model, while the coding model additionally supposes that non-synonymous codon substitution rates are reduced relative to synonymous rates by a scale factor ω (also called d_N/d_S). PAML takes the phylogenetic tree topology as input, and estimates the branch lengths, π, κ, and ω for each alignment. The log-likelihood ratio between the coding and non-coding models can then be obtained from PAML's output. (For detecting purifying selection, the log-likelihood ratio is set to zero if the estimated $\omega \geq 1$.)

Our previous work (Lin et al., 2008a) showed this to be one of the best comparative methods for distinguishing coding and non-coding regions, outperforming our CSF metric according to standard classification error measures. Notably however, the d_N/d_S test performed worse than CSF for short regions (≤ 180nt). This is not surprising since PAML was designed for evolutionary analysis of complete open reading frames, not short exon-length regions, which probably provide too little information to reliably estimate both the branch lengths and codon frequencies in addition to the two rate parameters.

PhyloCSF

PhyloCSF differs from the standard d_N/d_S test in two main ways. First, it takes advantage of recent advances in phylogenetic codon models that enable much more detailed representations of coding and non-coding sequence evolution. Specifically, while the d_N/d_S test uses only a few parameters to model the rates of all possible codon substitutions (Yang and Nielsen, 1998), PhyloCSF uses *empirical codon models* (ECMs) based on several thousand parameters modeling these rates (Kosiol et al., 2007) – one ECM estimated from alignments of many known coding regions, and another ECM from non-coding regions. By comparing these two rich evolutionary models, PhyloCSF can observe many additional informative features of a given alignment compared to the d_N/d_S test. For example, the coding ECM captures not only the decreased overall rate of non-synonymous substitutions, but also the different rates of specific non-synonymous substitutions reflecting the chemical properties of the amino acids. (Earlier codon modeling approaches also incorporate amino acid distances, e.g. Goldman and Yang (1994), but to our knowledge, these are not widely used for distinguishing coding and non-coding regions.) Also, our ECMs explicitly model the extreme difference in the rates of nonsense substitutions (giving rise to stop codons) in coding and non-coding regions.

Second, PhyloCSF also takes advantage of genome-wide training data to provide prior information about the branch lengths in the phylogenetic tree and the codon frequencies, rather than attempting to re-estimate these *a priori* in each individual alignment. PhyloCSF assumes a fixed tree "shape" based on the genome-wide MLEs of the branch lengths, and estimates only two scale factors ρ_C and ρ_N, applied uniformly to all of the branch lengths in the coding and non-coding models respectively, for each individual region analyzed. This allows PhyloCSF to accommodate some region-specific rate variation and reduces its sensitivity to the absolute degree of conservation – without needing to estimate many parameters for each alignment, which may be difficult for short regions.

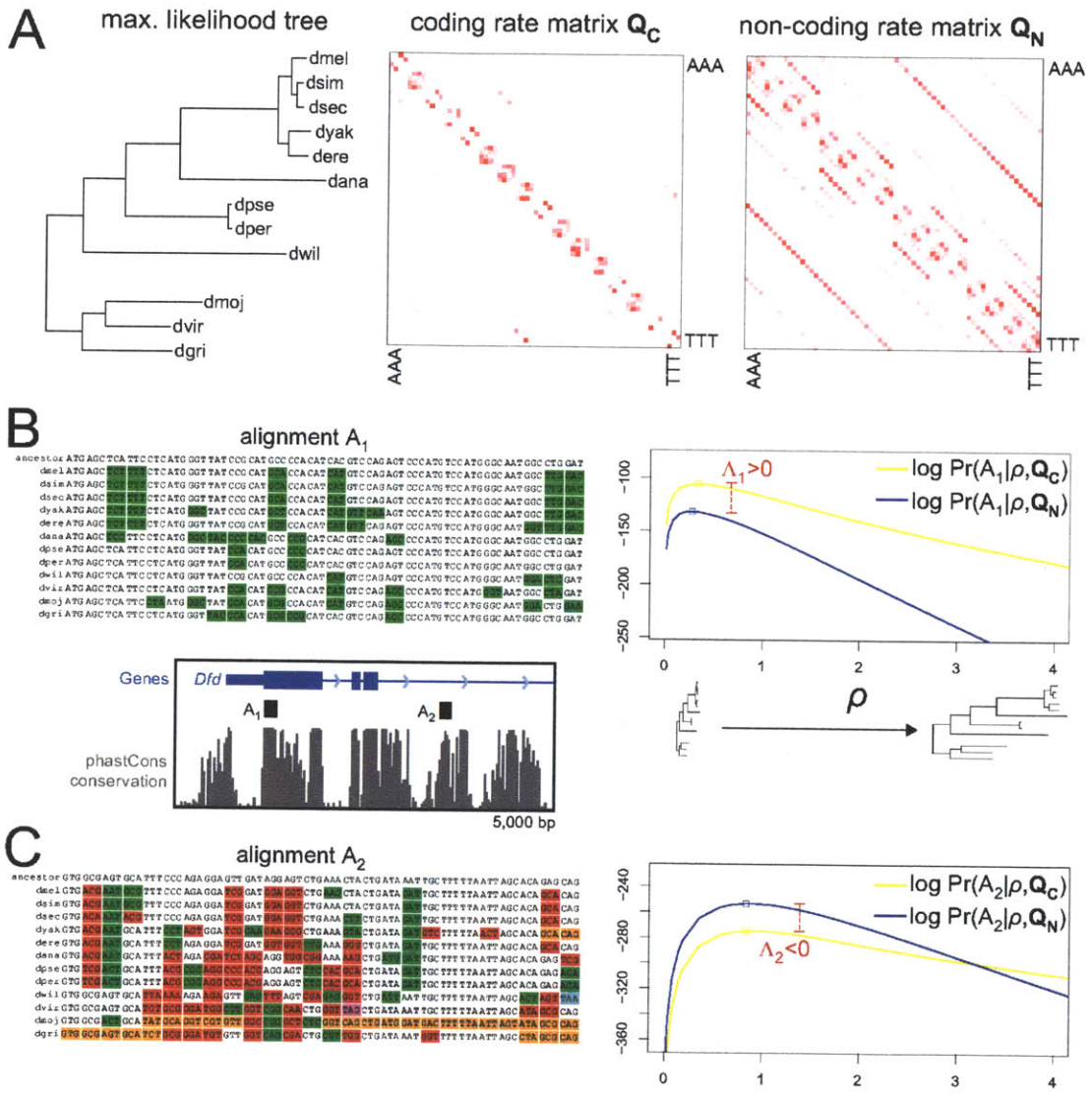
To summarize (Figure 3-1), PhyloCSF relies on two ECMs fit to genome-wide training data, which include estimates for the branch lengths, codon frequencies, and codon substitution rates for alignments of known coding and non-coding regions. To evaluate a given nucleotide sequence alignment, PhyloCSF (1) determines the MLE of the scale factor ρ on the branch lengths for each of these models, (2) computes the likelihood of each model (the probability of the alignment under the model) using the MLE of the scale factor, and (3) reports the log-likelihood ratio between the coding and non-coding models.

In detail, PhyloCSF's trained parameters include a phylogenetic tree (with branch lengths) and two 64-by-64 codon rate matrices $\mathbf{Q_C}$ and $\mathbf{Q_N}$ representing coding and

non-coding sequence evolution, respectively, as reversible, homogeneous, continuous-time Markov processes. To evaluate a given alignment, we first evaluate the likelihood of the coding model as follows. We define an alignment-specific parameter ρ_C that operates as a scale factor applied to all of the branch lengths in the predefined tree. Given a setting of ρ_C, the substitution probability matrix along any branch with length t is given by $\mathbf{P} = \exp(t\rho_C\mathbf{Q_C})$. We can then compute the probability of the full alignment using Felsenstein (2004)'s algorithm – assuming independence of the codon sites, using the equilibrium frequencies implicit in $\mathbf{Q_C}$ as the prior distribution over the common ancestral sequence, and marginalizing out any gapped or ambiguous codons. We numerically maximize this probability over ρ_C to obtain the likelihood of the coding model p_C. We then evaluate the likelihood of the non-coding model p_N in the same way, using $\mathbf{Q_N}$ and an independent scale factor ρ_N, and report the log-likelihood ratio $\Lambda = \log \frac{p_C}{p_N}$ as the result.

To estimate the phylogenetic tree and the empirical rate matrices $\mathbf{Q_C}$ and $\mathbf{Q_N}$ for the species of interest, we rely on sequence alignments of many known coding and random non-coding regions. Given this genome-wide training data, we estimate the parameters for the coding and non-coding models by maximum likelihood, using an expectation-maximization approach. The E-step is carried out as previously described (Holmes and Rubin, 2002; Siepel and Haussler, 2004). In each M-step, we update the ECM exchangeability parameters using a spectral approximation method (Arvestad and Bruno, 1997) and the branch lengths by numerical optimization of the expected log-likelihood function (Siepel and Haussler, 2004). Meanwhile, the codon/triplet frequencies are fixed to their empirical averages in the training examples, and we assume a fixed species tree topology.

Figure 3-1 *(facing page)*: PhyloCSF method overview. **(A)** PhyloCSF uses phylogenetic codon models estimated from genome-wide training data based on known coding and non-coding regions. These models include a phylogenetic tree and codon substitution rate matrices $\mathbf{Q_C}$ and $\mathbf{Q_N}$ for coding and non-coding regions, respectively, shown here for 12 *Drosophila* species. $\mathbf{Q_C}$ captures the characteristic evolutionary signatures of codon substitutions in conserved coding regions, while $\mathbf{Q_N}$ captures the typical evolutionary rates of triplet sites in non-coding regions. **(B)** PhyloCSF applied to a short region from the first exon of the *D. melanogaster* homeobox gene *Dfd*. The alignment of this region shows only synonymous substitutions compared to the inferred ancestral sequence (green). Using the maximum likelihood estimate of a scale factor ρ applied to the assumed branch lengths, the alignment has higher probability under the coding model than the non-coding model, resulting in a positive log-likelihood ratio Λ. **(C)** PhyloCSF applied to a conserved region within a *Dfd* intron. In contrast to the exonic alignment, this region shows many non-synonymous substitutions (red), nonsense substitutions (blue, purple), and frameshifts (orange). The alignment has lower probability under the coding model, resulting in a negative score.


A
max. likelihood tree
dmel
dsim
dsec
dyak
dere
dana
dpse
dper
dwil
dmoj
dvir
dgri
coding rate matrix $\mathbf{Q_C}$
non-coding rate matrix $\mathbf{Q_N}$
AAA
TTT
B
alignment A_1
Genes
Dfd
A_1
A_2
phastCons conservation
5,000 bp
Λ_1>0
log Pr($A_1|\rho,\mathbf{Q_C}$)
log Pr($A_1|\rho,\mathbf{Q_N}$)
ρ
C
alignment A_2
Λ_2<0
log Pr($A_2|\rho,\mathbf{Q_C}$)
log Pr($A_2|\rho,\mathbf{Q_N}$)

3.3 PhyloCSF outperforms other methods

We used the datasets and benchmarks from our previous study (Lin et al., 2008a) to evaluate our new method. Briefly, the datasets consist of known protein-coding regions and randomly selected non-coding regions (about 50,000 total regions) in the genome of the fruitfly *Drosophila melanogaster*, aligned with eleven other *Drosophila* species using MULTIZ (Drosophila 12 Genomes Consortium, 2007; Stark et al., 2007; Blanchette et al., 2004). The lengths of the regions in both the coding and non-coding sets match the length distribution of fly coding exons. Consistent with our previous work, we trained and applied PhyloCSF on this dataset using four-fold cross-validation, to ensure that any observed performance differences are not due to overfitting. We assessed the results by examining ROC curves and computing the minimum average error (MAE), the average false positive and false negative rates at the cutoff that minimizes this average. To compare the power of the methods for short exons specifically, we additionally computed these benchmarks only for the 37% of examples from 30 to 180 nucleotides in length (a range including three-quarters of mammalian coding exons).

These benchmarks show that PhyloCSF outperforms the other comparative methods we previously benchmarked (Figure 3-2, left column), effectively dominating them all at good sensitivity/specificity tradeoffs. PhyloCSF's overall MAE is 19% lower than that of the Reading Frame Conservation metric, 15% lower than that of our older CSF method, and 8% lower than the d_N/d_S test's. PhyloCSF also clearly outperforms the other methods for short exons (Figure 3-2, bottom row), with an MAE 11% lower than the d_N/d_S test's. Our previous study (Lin et al., 2008a) showed that these comparative methods in turn outperform single-species metrics based on sequence composition (e.g. interpolated Markov models (Delcher et al., 1999) and the Z curve discriminant (Gao and Zhang, 2004)).

We also compared PhyloCSF to the other methods using only pairwise alignments between *D. melanogaster* and *D. ananassae*, which we previously showed to be the best single informant for this purpose. PhyloCSF also dominates other methods in these pairwise alignments (Figure 3-2, right column), with MAE 24% lower than the next-best method's for all aligned regions (d_N/d_S test) and 21% lower for the short regions (CSF). Some of PhyloCSF's greater relative advantage in the pairwise case arises from its ability to produce an informative score even for regions lacking alignment with *D. ananassae*, based on the composition of the *D. melanogaster* region and the codon frequencies included in PhyloCSF's generative ECMs.

Overall, these benchmarks show that PhyloCSF provides superior power to distinguish fly coding and non-coding regions based on either multi-species or pairwise genome

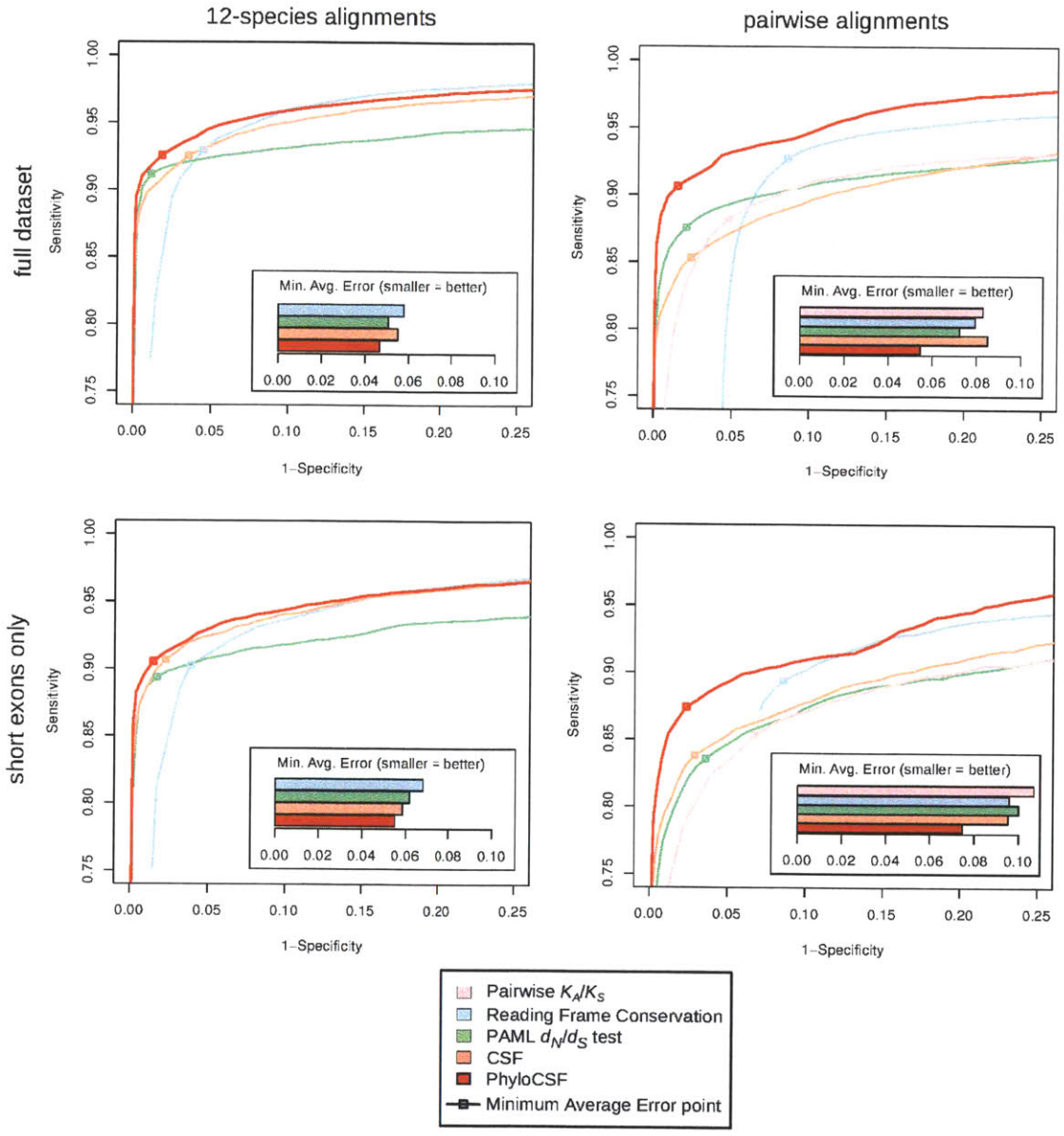


Figure 3-2: PhyloCSF performance benchmarks. ROC plots and error measures for several methods to distinguish known protein-coding and randomly selected non-coding regions in *D. melanogaster*. The top row of plots shows the results for our full dataset of approximately 50,000 regions matching the fly exon length distribution, while the bottom row of plots is based on the 37% of these regions between 30 and 180 nucleotides in length. The left-hand plots show the performance of the methods applied to multiple alignments of 12 fly genomes, while the right-hand plots use pairwise alignments between *D. melanogaster* and *D. ananassae*. PhyloCSF effectively dominates the other methods.

alignments.

3.4 Implementation

To facilitate the use of PhyloCSF by the community, we provide an implementation that evaluates input sequence alignments in Multi-FASTA format and reports the resulting log-likelihood ratios in units of decibans. We also provide the ECMs and other parameter settings for several phylogenies.

The software provides two additional noteworthy features. First, it can evaluate all open reading frames (ORFs) above a given length, in three or six reading frames, within each alignment. Thus, it can delimit likely protein-coding ORFs within transcript models that include untranslated regions. Second, the software also provides a simplified method, similar to the aforementioned d_N/d_S test, that does not require extensive training data from known coding and non-coding regions. This method is less accurate than the full ECM comparison presented here, as it reduces all substitutions to either synonymous or non-synonymous, but it can be used in settings where genome alignments are available but high-quality existing gene annotations are lacking, as may increasingly be the case outside of well-studied phylogenies such as mammals and flies.

The Objective Caml source code and executables for GNU/Linux and Mac OS X are available at:
`http://compbio.mit.edu/PhyloCSF`

3.5 Discussion

We have introduced PhyloCSF, a comparative genomics method for distinguishing protein-coding and non-coding regions, and shown that it outperforms previous methods. In addition to its superior discriminatory power, PhyloCSF is far more theoretically attractive than our older CSF and other *ad hoc* metrics, relying on a formal statistical comparison of phylogenetic codon models. However, we note that PhyloCSF and CSF produce highly correlated scores (Pearson coefficient 0.95 in our dataset), and the new method is much slower (though both have linear algorithmic complexity in the number of species).

PhyloCSF can provide an important building block in future computational strategies for genome annotation based on RNA-seq. Other pieces of the puzzle include methods to reconstruct transcript models based on short reads, complementary metrics of coding potential based on primary sequence composition or indel patterns, database search tools to identify similarity to known proteins and non-coding RNAs, and *de novo* gene structure

predictors that may be able to identify lowly- or rarely-expressed genes. Major challenges remain in integrating these methods into coherent pipelines, harmonizing the results with existing genome annotation databases, and coping with the uneven coverage and relatively high error rate of current high-throughput sequencing technologies (Ozsolak and Milos, 2011).

Chapter 4

Applications of PhyloCSF and related methods in fungal, fly and vertebrate genomes

In this chapter, we summarize how we have used PhyloCSF and related methods to discover novel genes, revise existing gene annotations, and reveal unusual gene structures in several different eukaryotic genomes. This work has arisen from extensive collaborations with other scientists studying these species, providing access to RNA-seq and other experimental data, and often the curators who maintain the canonical gene annotations for their genomes. For historical reasons, some of this work was carried out using CSF, the previously mentioned method that preceded PhyloCSF. As discussed in the preceding chapter, PhyloCSF is more mathematically rigorous and more accurate than CSF, but the two methods serve the same essential purpose and can be applied in similar settings. Our previous work included benchmarking CSF and a variety of other metrics of coding potential (Lin et al., 2008a).

4.1 Revisiting the protein-coding gene catalog of Drosophila melanogaster

The fruitfly *Drosophila melanogaster* has played a central role in science's understanding of genetics and the development of the bilateral animal body plan. Following the sequencing and alignment of twelve *Drosophila* genomes (Drosophila 12 Genomes Consortium, 2007; Stark et al., 2007), we collaborated with FlyBase and the Berkeley Drosophila Genome Project to use CSF and other comparative metrics in an extensive revision of its gene

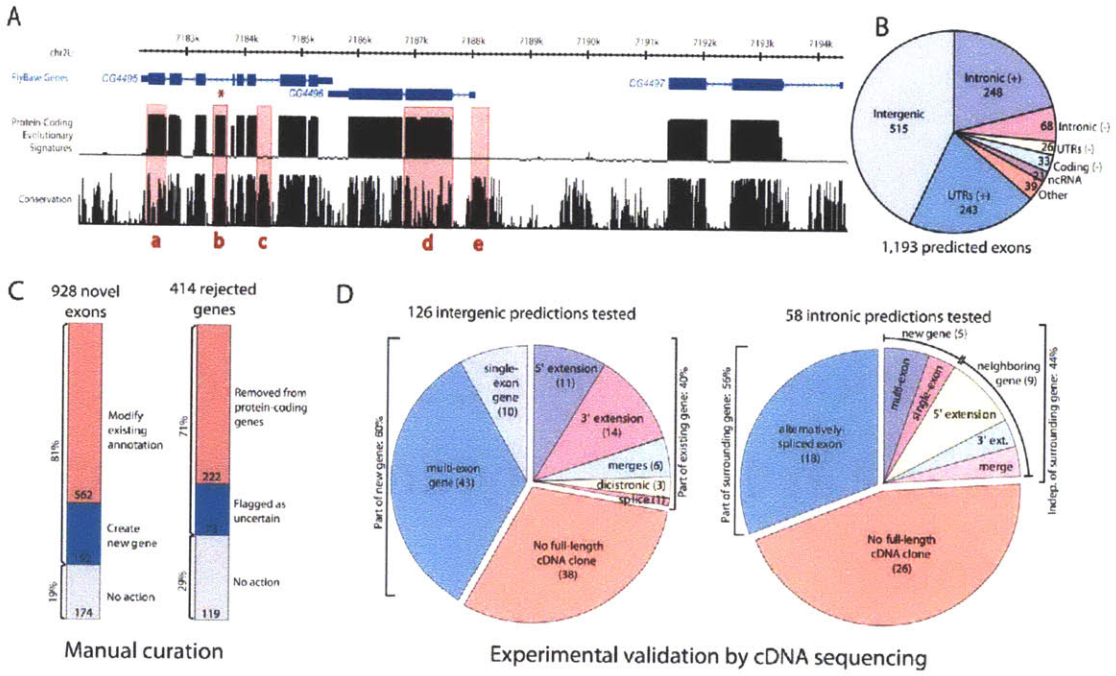


Figure 4-1: New protein-coding exons predicted by evolutionary signatures, examined by manual curation, and validated by cDNA sequencing. **(A)** The "Evolutionary Signatures" track shows the posterior probability of a protein-coding state in an exon finder incorporating CSF. The "Conservation" track shows the analogous quantity from a model measuring nucleotide conservation only (Siepel, 2005). Note the high protein-coding scores of known exons despite lower nucleotide conservation (a,d), the low protein-coding scores of conserved non-coding regions (c,e), and the prediction of a novel exon within an intron of *CG4495* (b), subsequently validated (see Figure 4-2). Rendered by the UCSC Genome Browser (Kent et al., 2002). **(B)** Distribution of 1193 new exon predictions throughout the genome. **(C)** Newly predicted exons were examined by manual curation, 81% leading to new and modified FlyBase gene annotations. Additionally, curation of genes rejected by evolutionary signatures led to the recognition of hundreds of spurious annotations. **(D)** A sample of predicted new exons was tested by cDNA sequencing with inverse PCR. Surprisingly, 44% of the validated predictions in "intronic" regions revealed a transcript independent of the surrounding gene, and 40% of the validated predictions in "intergenic" regions were part of existing genes. See Figure 4-2 for examples.

annotations (Lin et al., 2007) – ultimately affecting about 10% of all genes in this species.

First, we implemented a simple protein-coding exon finding algorithm for the fly genome by incorporating CSF into a conditional random field (CRF), a type of probabilistic graphical model. A CRF is similar to a hidden Markov model (HMM) in that it observes features of the input sequence in order to generate an optimal segmentation of it – in this case, segmenting the genome into protein-coding exons and non-coding regions – but, unlike at HMM, it can directly incorporate non-probabilistic discriminative features like CSF (Gross et al., 2007; DeCaprio et al., 2007). We trained this algorithm

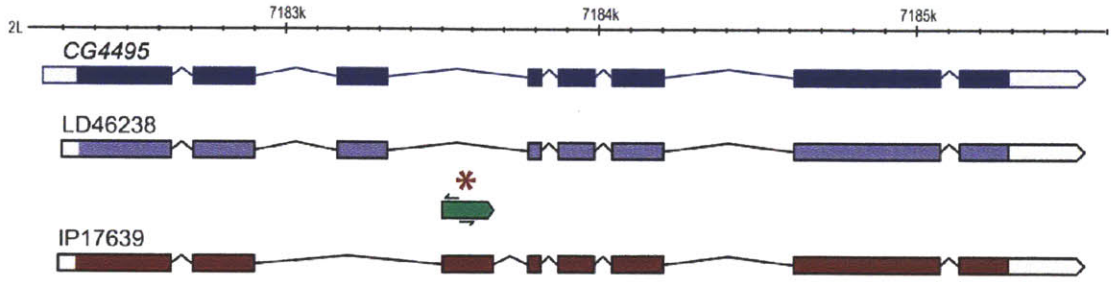
using the existing FlyBase gene annotations as training data, and then used it to predict 1193 novel protein-coding exons. Of these, 983 were subjected to manual curation and experimental validation, resulting in the incorporating of 83% into updated FlyBase annotations (Figure 4-1). The inverse PCR experimental validation protocol allowed the recovery of many novel complete gene structures based on the individual exon predictions (Figure 4-2). With manual curation, this led to a total of 150 new gene models and 84 changes to existing annotations.

We also identified 414 existing gene annotations lacking any evidence of protein-coding sequence evolution in the aligned species, suggesting either that they are very novel or rapidly evolving genes, or else simply spurious annotations. Upon manual examination, 222 were deleted from the annotations or recategorized as non-coding genes, and many of the remainder were marked as being of uncertain quality. Lastly, we used the high resolution provided by CSF to propose detailed refinements to existing gene models, including moving 413 translation start sites, adjusting over 900 exon boundaries, changing the reading frame of translation in 5 genes, and identifying several recent nonsense and frameshift mutations (Figure 4-3).

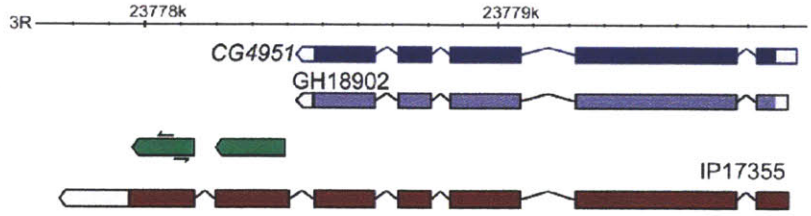
Later, the modENCODE project produced evidence for 1938 novel genes using both inverse PCR and RNA-seq in several cell types (The modENCODE Consortium et al., 2010). We applied PhyloCSF to these transcript models to identify conserved protein-coding genes among them, using the mode that evaluates all ORFs in the transcript and reports the best-scoring. As expected, relatively few of the novel loci appeared to represent protein-coding genes: 57 contained complete, nicely conserved ORFs, and an additional 81 appeared to represent incomplete coding gene models, as they appear to contain conserved coding sequences but lack clear translation start and stop sites. The remaining 1800 novel genes are likely to represent largely non-coding genes.

Figure 4-2 *(facing page)*: Experimental validation of predicted novel exons in *D. melanogaster*. **(A)** Alternatively spliced transcripts – Exon Shuffling. An mRNA transcript sequence (red) recovered by targeting the exon prediction (green) using inverse PCR provides evidence for an alternative transcript of the annotated gene *CG4495* (blue). **(B)** Coding sequence extension of *CG4951*. **(C)** New spliced, interleaved gene. Validation of predicted exons reveals a novel gene encoded antisense to introns of known genes. **(D)** Novel overlapping genes. An inverse PCR clone provides support for a novel gene antisense to *Rad9*, with a region of exonic overlap such that 45 amino acids in each putative protein are encoded on opposite strands of the same DNA.

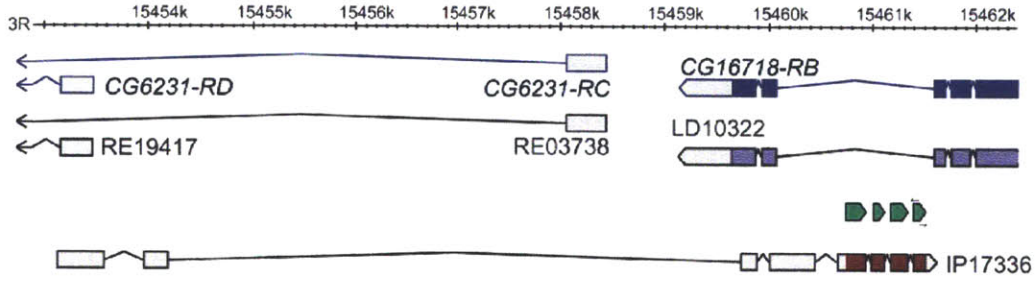
A. New alternatively spliced isoform



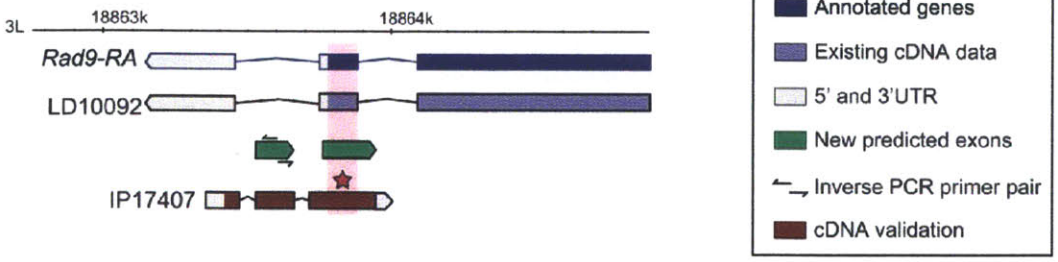
B. New 3' exons extend existing gene model



C. cDNA validation reveals interleaved genes



D. cDNA validation reveals overlapping coding regions



4.2 Revising gene annotations in additional species

4.2.1 Human and other vertebrate genomes

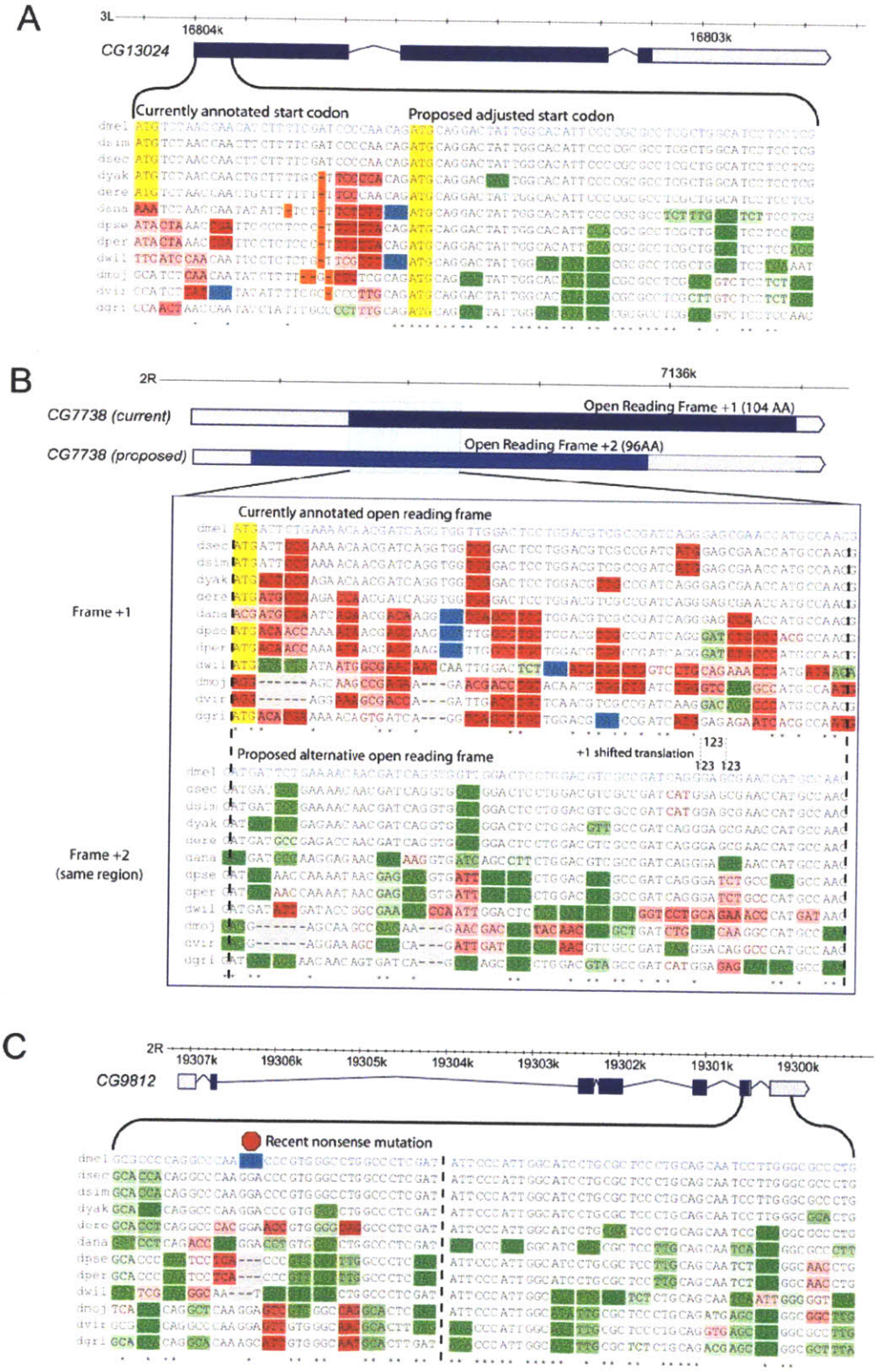
We have used CSF and PhyloCSF to predict novel genes in the human and mouse genomes, using alignments of dozens of mammals. We used our CRF exon finder to predict 3788 candidate novel, conserved exons in the human genome, most of which have additional support in the form of RNA-seq evidence or similarity to known protein sequences (Lindblad-Toh et al., 2011). These predictions are being used along with other computational and experimental sources in the GENCODE project to create a manually-curated reference human genome annotation (Harrow et al., 2006). Also within the context of GENCODE, we applied PhyloCSF to transcript models reconstructed from RNA-seq on 16 human tissues to identify several novel coding genes (Figure 4-4), despite the extensive decade-long efforts to annotate the human genome.

CSF and PhyloCSF have also played an important role in recognizing thousands of long *non*-coding RNAs with diverse biological functions in both the human and mouse genomes, by distinguishing coding and non-coding genes in RNA-seq transcript datasets (Guttman et al., 2009, 2010; Hung et al., 2011; Cabili et al., 2011). Similarly, we used PhyloCSF in a pipeline to identify long non-coding RNAs expressed during embryonic development in the zebrafish *Danio rerio*, another important model organism for studying vertebrate development (Pauli et al., 2011).

4.2.2 The pathogenic fungus Candida albicans

Candida albicans is a common human pathogen, causing millions of fungal infections each year. A complete and accurate annotation of its genome can support research on novel treatments for such infections as well as studies of its basic biology. The recent sequencing of eight *Candida* genomes provided an opportunity to revisit its gene annotations, similar to our previous work in flies.

Figure 4-3 *(facing page)*: Examples of adjustments to existing annotations based on evolutionary signatures. **(A)** Translation start adjustment. The annotated coding sequence begins at the indicated ATG, but the informant species show frameshifts, nonsense mutations, and nonconservative substitutions in the immediately downstream region. Strikingly, however, coding signatures begin at a slightly downstream ATG. **(B)** Incorrect reading frame annotated. The transcript model contains two overlapping reading frames, the slightly longer of which is annotated as the coding sequence; but the evolutionary signatures clearly show that the other is the frame under selection. **(C)** Nonsense mutation in (the sequenced strain of) *D. melanogaster*.


A
3L
16804k
16803k
CG13024
Currently annotated start codon
Proposed adjusted start codon
B
2R
7136k
CG7738 (current)
Open Reading Frame +1 (104 AA)
CG7738 (proposed)
Open Reading Frame +2 (96AA)
Currently annotated open reading frame
Frame +1
Proposed alternative open reading frame
+1 shifted translation
123
123 123
Frame +2
(same region)
C
2R
19307k
19306k
19305k
19304k
19303k
19302k
19301k
19300k
CG9812
Recent nonsense mutation

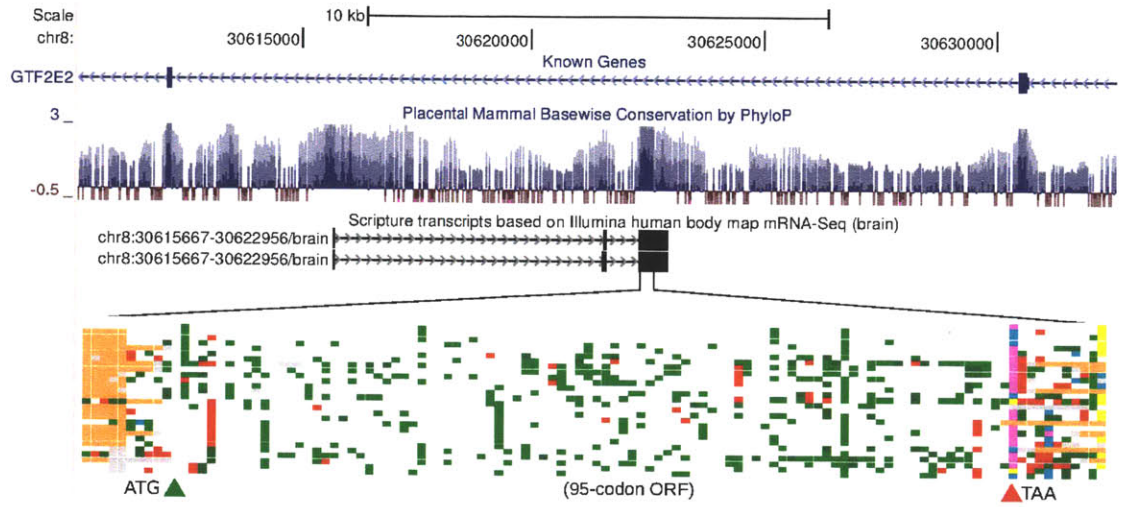


Figure 4-4: A novel human coding gene found using RNA-seq and PhyloCSF. Transcriptome reconstruction by Scripture (Guttman et al., 2010) based on brain RNA-seq data provided by Illumina, Inc. produced two alternative transcript models lying antisense to an intron of *GTF2E2*, a known protein-coding gene. PhyloCSF identified a 95-codon ORF in the third exon of this transcript, highly conserved across placental mammals. The color schematic illustrates the genome alignment of 29 placental mammals for this ORF, indicating conservation (white), synonymous and conservative codon substitutions (green), other non-synonymous codon substitutions (red), stop codons (blue/magenta/yellow), and frame-shifted regions (orange). Despite its unmistakable protein-coding evolutionary signatures, the ORF's translation shows no sequence similarity to known proteins.

We generated whole-genome alignments of the eight genomes using MULTIZ (Blanchette et al., 2004), and then trained our CRF exon finder using existing gene annotations from the Candida Genome Database (CGD) as training data (Inglis et al., 2011). The CRF identified 91 candidate open reading frames exhibiting high CSF scores and other characteristics of protein-coding genes. Upon manual examination by CGD curators, 84 of these were ultimately incorporated into the *C. albicans* genome annotation. We also identified 222 annotated genes lacking any evidence of protein-coding sequence evolution in the aligned species; on manual examination, 181 of these were found to have no experimental support and were flagged as "dubious" in CGD. CSF also proved useful for identifying likely sequencing errors in the *C. albicans* genome assembly, leading to 173 corrections (Butler et al., 2009).

4.2.3 The fission yeast Schizosaccharomyces pombe

The fission yeast *Schizosaccharomyces pombe* is a model organism in eukaryotic cell biology. Recent RNA-seq experiments provided evidence for ~800 novel transcript models in its genome. We generated whole-genome alignments of *S. pombe* with *S. octosporus*, *S. japonicus* and *S. cryophilus* using MULTIZ, trained PhyloCSF using existing gene annotations, and then applied it to the novel RNA-seq transcript models using the mode that evaluates all ORFs in the transcript and reports the best-scoring. This identified 89 novel protein-coding genes that are highly conserved among these species, even though they do not show primary sequence similarity to known proteins. The RNA-seq reconstruction also suggested 400 revisions to the exon-intron structures of existing gene annotations. PhyloCSF showed that the revised transcript models strongly tend to have higher scores than the originals, confirming the quality of the proposed revisions (Rhind et al., 2011).

4.3 Discovery of unusual gene structures

In addition to improving annotations of canonical protein-coding genes, the discriminatory power and high resolution provided by CSF and PhyloCSF allowed us to recognize a variety of unusual phenomena that have not been amenable to systematic genome-wide discovery, including stop codon readthrough, polycistronic transcripts, and translational frameshifts (Lin et al., 2007).

4.3.1 Stop codon readthrough

The protein-coding evolutionary signatures identified by CSF and PhyloCSF usually terminate abruptly at the stop codon ending the ORF, or shortly upstream – as expected, since the downstream region usually does not encode a protein sequence. When we used CSF to analyze these regions in *D. melanogaster*, however, we discovered 149 genes in which the evolutionary signatures strongly suggest that translation continues well past a deeply-conserved, in-frame stop codon (Figure 4-5A). This suggests that many fly proteins have biologically functional isoforms that are expressed by stop codon readthrough, an unusual process in which the ribosome ignores a stop codon and continues translation of the downstream region. In many cases, the evolutionary signatures continue exactly until the next downstream in-frame stop codon, although there were a few cases in which *two* stop codons appeared to be bypassed.

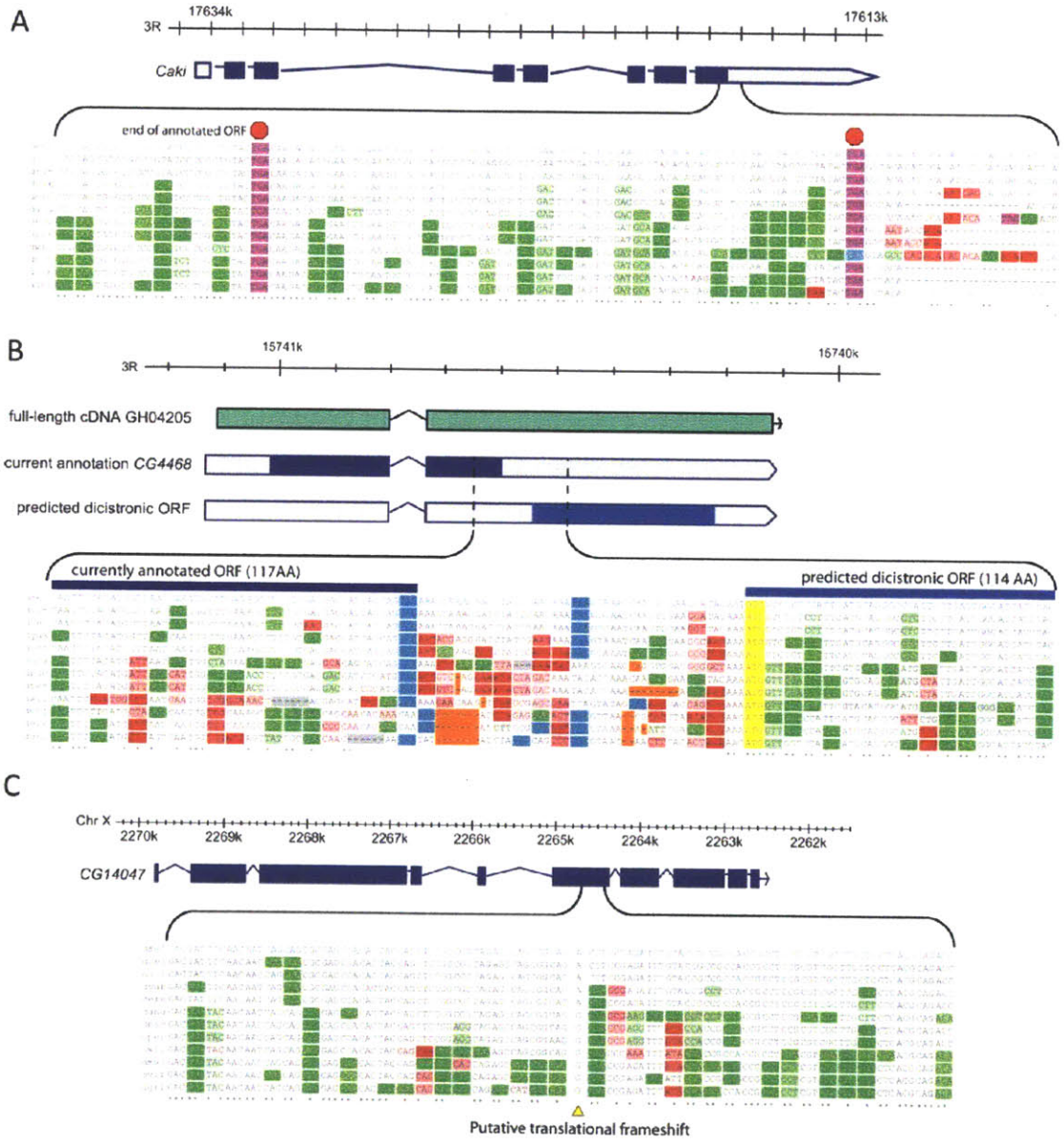
Our evidence for a mechanism of stop codon readthrough active in hundreds of fly genes was extremely surprising, as only a handful of cases were previously known. Follow-up work using PhyloCSF expanded our original list of 149 fly readthrough genes to 283 candidates, investigated various mechanistic explanations, studied their evolutionary history, experimentally validated several candidates in fly embryos, and found additional examples in other species, including human (Jungreis et al., 2011).

4.3.2 Polycistronic mRNAs

Polycistronic mRNAs are single processed transcripts containing several nonoverlapping ORFs, each of which is individually translated. We used CSF to search for complete (start-to-stop) fly ORFs that show clear signs of protein-coding selection and are fully contained within the untranslated region of an existing transcript model (Figure 4-5B). This strategy rediscovered 85 of 115 euchromatic dicistronic transcripts previously annotated in FlyBase (73%) and predicted an additional 135 putative ORFs in 123 genes. Thus, our results potentially double the number of known dicistronic transcripts in the *D. melanogaster* genome.

4.3.3 Programmed ribosomal frameshifting

Programmed ribosomal frameshifts occur when the ribosome is induced to skip one or two nucleotides, changing the reading frame of translation for the downstream region. This potentially allows expression of alternative protein sequences in a regulated fashion, depending on whether the frameshift actually occurs during translation. We found four locations in fly transcripts where protein-coding selection abruptly shifts from one reading


A
17634k
3R
17613k
Caki
end of annotated ORF
B
15741k
3R
15740k
full-length cDNA GH04205
current annotation CG4468
predicted dicistronic ORF
currently annotated ORF (117AA)
predicted dicistronic ORF (114 AA)
C
Chr X
2270k
2269k
2268k
2267k
2266k
2265k
2264k
2263k
2262k
CG14047
Putative translational frameshift

frame to another, that are not readily explained as an incorrect transcript model or a species- or lineage-specific mutation. In these cases, the comparative evidence appears to indicate that a conserved translational frameshift occurs (Figure 4-5C). One case in *Apc* is mediated by a highly conserved RNA structure, as subsequently verified experimentally based on our prediction (Baranov et al., 2011).

Figure 4-5 *(facing page)*: Unusual fly gene structures identified using CSF. **(A)** A well-conserved 30aa ORF immediately following the stop codon of the gene *Caki* suggests translational readthrough. Note the perfect conservation of the putative readthrough stop codon, the "wobble" of the downstream stop codon, and the abrupt loss of conservation following the downstream stop codon, typical of a true translation stop. **(B)** A well-conserved ORF within the annotated 3' UTR of *CG4468* suggests a dicistronic transcript structure. Note the region of poor conservation extending precisely from the upstream stop codon to the downstream start codon, suggesting separate translation of the two ORFs. **(C)** An abrupt change in the reading frame upon which selection appears to act within an exon of *CG14047* is suggestive of a "programmed" translational frameshift.

Chapter 5

Locating synonymous constraint elements in mammalian genomes

5.1 Introduction

It is often assumed that synonymous sites within protein-coding open reading frames (ORFs) evolve neutrally, since mutations in them do not change the amino acid translation. But, in fact, ORFs in many species simultaneously encode additional functional sequence elements within the codon sequence, often with strong evolutionary constraint on the synonymous sites (Chamary et al., 2006; Itzkovitz and Alon, 2007). For example, mammalian ORFs are known to encode exonic splicing enhancers and silencers (Chen and Manley, 2009), microRNA target sites (Lewis et al., 2005; Hurst, 2006), A-to-I recoding sites (Rueter et al., 1999; Bass, 2002), and transcriptional enhancers (Lang et al., 2005; Nguyen et al., 2007; Lampe et al., 2008; Tumpel et al., 2008; Dong et al., 2009). Several previous studies have observed strong genome-wide trends toward increased evolutionary constraint on such overlapping functional elements, by averaging across many loci pooled together (Baek, 2005; Xing and Lee, 2006; Chen, 2005; Down et al., 2006; Goren et al., 2006; Parmley, 2005; Robins et al., 2008; Kural et al., 2009). However, these averaging approaches generally did not have the power to locate individual overlapping functional elements within specific genes.

In this chapter, we use the discovery power provided by alignments of 29 mammalian genomes to provide a systematic annotation of individual functional elements embedded within protein-coding regions throughout the human genome. Since the average codon site in these multiple sequence alignments shows about four synonymous substitutions, we predict that overlapping functional elements will individually stand out as short, localized regions with exceptionally few synonymous substitutions – in much the same way that

widely used methods such as GERP, phastCons, phyloP, and SiPhy locate conserved functional elements within a background of neutral nucleotide-level sequence evolution (Cooper, 2005; Siepel, 2005; Margulies et al., 2007; Garber et al., 2009; Pollard et al., 2009).

However, detecting overlapping evolutionary constraints within protein-coding ORFs is more difficult than detecting general nucleotide-level constraints, for two main reasons. First, since the majority of nucleotide sites in a typical human ORF are already highly conserved among mammals due to their protein-coding function, we must expect less statistical power to detect increased conservation for overlapping sequence elements of a given length. Second, it is important to account precisely for the protein-coding constraints on each nucleotide site by modeling the evolutionary process at the codon level, rather than analyzing the conservation of individual nucleotide sites independently of those surrounding them.

To address these challenges, we present a novel adaptation of statistical phylogenetic codon models widely used in evolutionary analysis of protein-coding genes (for recent reviews, see Anisimova and Kosiol (2008) and Delport et al. (2008)), which locates short windows within alignments of known human ORFs showing significantly reduced rates of synonymous substitution. Applying this new method to the 29-species alignments, we confidently locate more than 10,000 such regions, typically with 70%–90% reduced synonymous rates, down to a resolution of just nine codons. These putative "synonymous constraint elements" contain only $\sim$2% of all synonymous sites, but are found within more than a quarter of all human protein-coding genes.

A few previous studies also sought to locate individual mammalian SCEs on a genome-wide basis (Bejerano et al., 2004; Parmley and Hurst, 2007b; Schattner, 2006; Suzuki and Saitou, 2011), though generally with less sensitivity, resolution, and methodological rigor. Other related studies, including some multispecies approaches, have analyzed only a small fraction of mammalian genes (Hurst and Pal, 2001; Chen and Blanchette, 2007; Parmley and Hurst, 2007b; Lin et al., 2008b), while we undertake a comprehensive genome-wide analysis. A few methods have been developed to identify examples of certain known classes of overlapping functional elements with predictable evolutionary signatures, including dual-coding ORFs (Chung et al., 2007; Ribrioux et al., 2008) and RNA secondary structures (Pedersen, 2004a,b). These are complementary to the rate-based approach we take here. Lastly, our codon modeling methods are related to previous efforts to incorporate synonymous rate variation in such models (Pond, 2005; Mayrose et al., 2007; Singh et al., 2007; Yang and Nielsen, 2008; Zhou et al., 2010; Rubinstein et al., 2011). These studies have tended to focus on codon usage bias and how rate variation affects detection

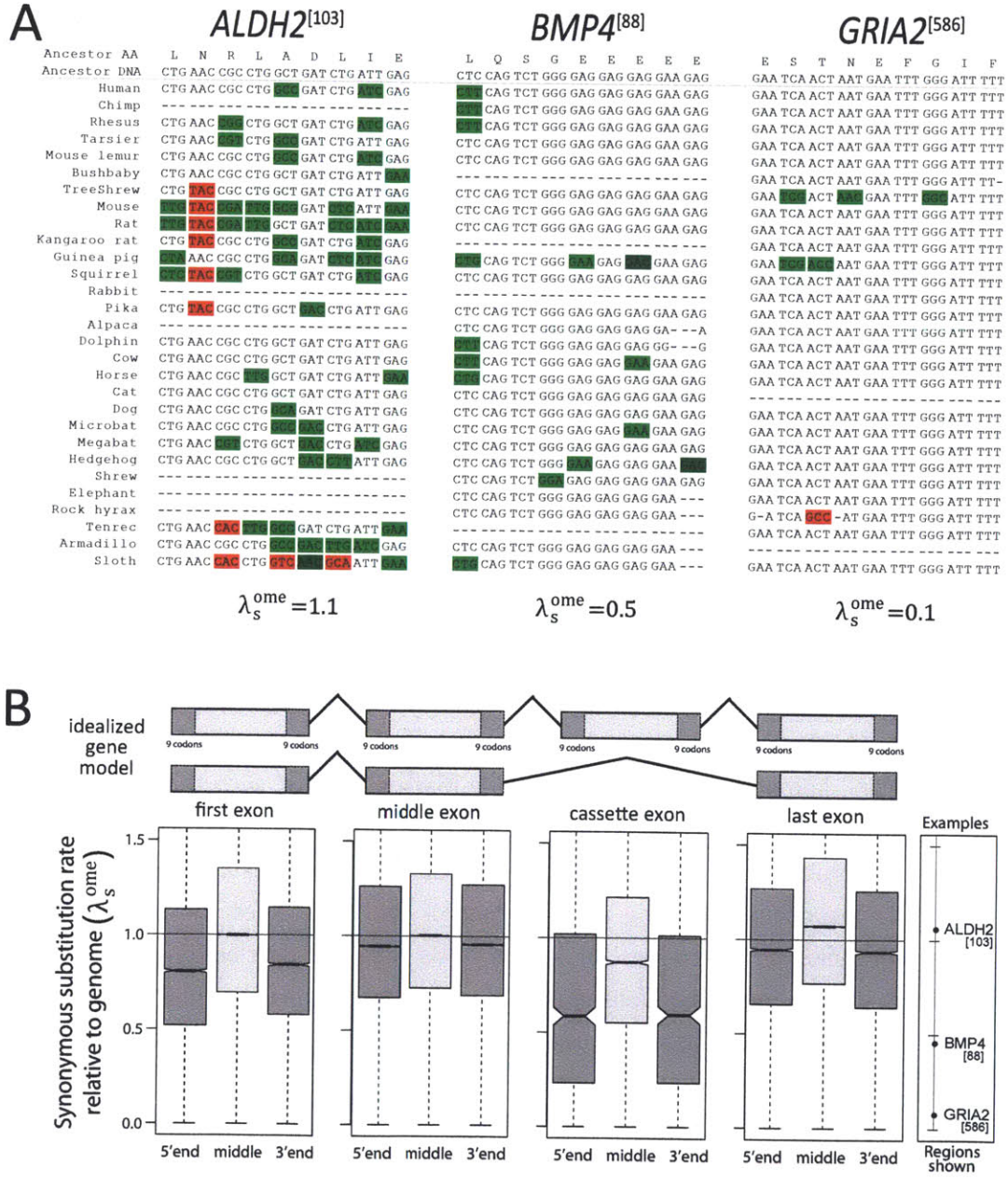
of positive selection at the peptide level, rather than our goal of reliable genome-wide detection of mammalian SCEs.

5.2 Estimating synonymous substitution rates in short windows within ORFs

Our method uses phylogenetic codon models to find short windows within multi-species alignments of known ORFs that exhibit unusually low rates of synonymous substitution as measured by d_S, a composite rate commonly used to summarize the relevant parameters of such models (Yang and Bielawski, 2000). Specifically, our method analyzes any window of adjacent codon sites within an alignment to compute the maximum likelihood estimate of a parameter λ_s, which is a scale factor on d_S, indicating how much slower or faster synonymous substitutions have occurred in that window relative to a null model representing typical protein-coding sequence evolution. For example, a particular window with $\lambda_s = 0.5$ is estimated to have evolved with a synonymous substitution rate only one-half that of the null model average (window $d_S = \lambda_s \times$ average d_S) (Figure 5-1A). Our method is designed to estimate λ_s for short windows of adjacent sites – in this study, nine to 30 codons. (The method also estimates λ_n, the analogous parameter for non-synonymous substitutions, which we use to exclude potentially misaligned regions.)

In detail, our relative rate estimation procedure uses, as a parameter, any standard phylogenetic codon model $M = \langle T, Q \rangle$ where T specifies the topology and branch lengths of a phylogenetic tree and Q is a reversible 61×61 rate matrix describing codon evolution as a stationary, homogeneous, continuous-time Markov process, such that the transition probability matrix for any branch with length t is given by $P = \exp(Qt)$. Given such a model, the probability of any alignment of extant sequences can be computed using Felsenstein's algorithm, assuming independence of the codon sites and using the equilibrium frequencies of the codons implicit in Q as the prior distribution over the root. Nonaligned, gapped, or stop codons in any informant species are marginalized out so that they are irrelevant to the probability, using standard techniques for statistical phylogenetic models (Felsenstein, 2004).

To analyze a given alignment window of several codons, we wish to obtain maximum likelihood estimates of the synonymous (and non-synonymous) rates relative to the null model M. Our approach is to hold T fixed and estimate a new window-specific rate matrix Q^{wnd} by numerically maximizing the probability of the given alignment window jointly over two nonnegative parameters λ_s and λ_n, where the entries of Q^{wnd} relative to the entries of Q are given by


A
ALDH2[103]
BMP4[88]
GRIA2[586]
Ancestor AA
Ancestor DNA
Human
Chimp
Rhesus
Tarsier
Mouse lemur
Bushbaby
TreeShrew
Mouse
Rat
Kangaroo rat
Guinea pig
Squirrel
Rabbit
Pika
Alpaca
Dolphin
Cow
Horse
Cat
Dog
Microbat
Megabat
Hedgehog
Shrew
Elephant
Rock hyrax
Tenrec
Armadillo
Sloth
λ_s^{ome}=1.1
λ_s^{ome}=0.5
λ_s^{ome}=0.1
B
idealized gene model
9 codons
first exon
middle exon
cassette exon
last exon
Examples
Synonymous substitution rate relative to genome (λ_s^{ome})
1.5
1.0
0.5
0.0
5'end
middle
3'end
ALDH2 [103]
BMP4 [88]
GRIA2 [586]
Regions shown

$$q_{ij}^{\text{wnd}} = \begin{cases} \lambda_s \times q_{ij} & \text{if } i \neq j \text{ and } \text{aa}_i = \text{aa}_j \\ \lambda_n \times q_{ij} & \text{if } i \neq j \text{ and } \text{aa}_i \neq \text{aa}_j \\ -\sum_{k \neq i} q_{ik}^{\text{wnd}} & \text{otherwise } (i = j) \end{cases}$$

where aa_i denotes the amino acid translation of codon i. Thus, λ_s represents a scale factor on the synonymous rates specified by Q, and similarly for λ_n on the non-synonymous rates. Importantly, while Q is typically normalized to unity mean rate of replacement at equilibrium, we do *not* renormalize Q^{wnd}. Since T is held fixed, this allows λ_s and λ_n to control the absolute synonymous and non-synonymous rates, respectively. Assuming Q is reversible, it is easy to verify that Q^{wnd} is reversible with the same equilibrium frequencies π_j as Q, by considering the decomposition $q_{ij} = \pi_j \times s_{ij}$ for symmetric "exchangeabilities" s_{ij} and noting that λ_s and λ_n scale the entries symmetrically.

Using this approach, we reduced the parameter estimation problem in each window to a mere two-dimensional optimization by reusing ORF- or ORFeome-wide estimates of many other phylogenetic model parameters, for which we probably could not obtain reliable joint estimates based only on a few codon sites (Anisimova et al., 2001; Suzuki and Nei, 2002; Schmid and Yang, 2008; Nozawa et al., 2009). This is similar to the approach used by the Sitewise Likelihood Ratio method of Massingham (2004), which reuses ORF-wide parameter estimates while estimating $\omega = d_N/d_S$ at individual sites. We reuse common parameter estimates while effectively estimating both d_N and d_S in windows of several sites.

Figure 5-1 *(facing page)*: **(A)** Examples of local synonymous rate variation in alignments of 29 placental mammals for short nine-codon windows within the open reading frames (ORFs) of three known human protein-coding genes (ALDH2, BMP4, and GRIA2) with brackets denoting starting codon position within each ORF of shown alignment. (Bright green) Synonymous substitutions with respect to the inferred ancestral sequence; (dark green) conservative amino acid substitutions; (red) other non-synonymous substitutions. The estimated parameter some denotes the rate of synonymous substitution within these selected windows relative to genome-wide averages. For example, the nine-codon window starting at codon 88 of the BMP4 ORF shows $\lambda_s^{\text{ome}} = 0.5$, corresponding to an estimated synonymous substitution rate 50% below the genome average. **(B)** Variation in the estimated synonymous rate at different positions with respect to exon boundaries and translation start and stop, across all CCDS ORFs. For each class of regions, box-and-whisker plots show the observed distribution of λ_s^{ome}, including the median (middle horizontal bars), middle 50% range (boxes), extreme values (whiskers), and whether medians differ with high statistical confidence (nonoverlapping notches between two boxes). Estimated synonymous rates tend to be significantly reduced at the 5' and 3' ends of exons, and dramatically reduced in alternatively spliced exons, likely reflecting widespread splicing regulatory elements embedded within protein-coding regions.

5.3 Statistical significance of synonymous rate reduction

To locate synonymous constraint, we wish not only to find a small estimate of λ_s in a given window, but also to show that the reduction is not explained by random effects. We can associate a statistical significance with our estimate of λ_s for any window, using standard techniques for testing the goodness-of-fit of statistical phylogenetic models. Specifically, we can perform a likelihood ratio test (LRT) to assess precisely whether a small estimate of λ_s explains an observed alignment window better than the average rates assumed by the null model. Then, if a window has a significantly reduced s according to this test, we infer that its synonymous sites have probably been constrained by natural selection acting on an overlapping functional element. This is very similar to likelihood methods for nucleotide-level constraint detection (Garber et al., 2009; Pollard et al., 2009), extended to codon models so that we can disentangle the different evolutionary pressures on synonymous and non-synonymous sites.

The LRT provides an elegant way to avoid certain potential pitfalls in detecting individual regions with reduced synonymous rates. For example, it accounts for the uncertainty in rate estimates based on the exact set of informant species aligned for each window. To illustrate, consider two windows, one with estimated $\lambda_s = 0.5$ with all 29 species aligned, and the other with $\lambda_s = 0.1$ but only a human/chimpanzee alignment available. Even though the estimated λ_s is lower in the second window, it is almost surely less significant by the LRT, because it is based on far less informative data. The LRT also accounts for the expected constraint at each individual site based on the amino acid it encodes. For example, consider a hypothetical window coding exclusively for conserved methionine and tryptophan residues, which are encoded by non-degenerate codons (ATG and TGG, respectively). By definition, this window does not exhibit any synonymous substitutions and might therefore appear to have a very low synonymous substitution rate. But a reduced estimate of λ_s in this window would not be considered significant, because it does not provide a better explanation for the conservation of the non-degenerate sites.

We also designed our method to control for background variation in sequence composition and evolutionary rates across the genome (Lercher et al., 2001; Williams and Hurst, 2002; Fox et al., 2008), as well as the possibility of selection on diffuse effects that can constrain all or most of an ORF, such as transcript structural stability or codon bias for translation efficiency (Chamary et al., 2006). To account for these, we evaluated each window against *two* null models, one representing the average sequence composition and evolutionary rates of the entire "ORFeome" (to obtain λ_s^{ome}), and the second

estimated specifically from the individual ORF containing each window (to obtain λ_s^{ORF}). By identifying statistically significant rate reductions with respect to both null models, we required windows of interest to be exceptional with respect to genome-wide averages, on one hand, and also not explained by local biases in composition, rates, and codon usage, on the other hand.

In detail, to test the significance of the adjusted rate estimates in any alignment window, we evaluated the likelihoods of several models:

1. The null model $M(\lambda_s = 1; \lambda_n = 1)$
2. $\lambda_s = 1; \lambda_n$ estimated by maximum likelihood
3. λ_s and λ_n jointly estimated by maximum likelihood $(0 \leq \lambda_s \leq 1)$
4. $\lambda_s = 1; \lambda_n$ estimated by maximum likelihood $(\lambda_n > 1)$

The likelihood ratios of nested models can then be used to perform the different significance tests we described. For example, the primary test for $\lambda_s < 1$ compares model 3 to model 2, and the test for $\lambda_n > 1$ compares model 4 to model 1. Specifically, let $\Pr(A|M, \lambda_s = a, \lambda_n = b)$ denote the probability of the alignment window A, given the null model M and with Q^{wnd} configured with $\lambda_s = a, \lambda_n = b$ for specified A and B, computed using Felsenstein's algorithm. The likelihood ratio Λ between models 3 and 2 is:

$$\Lambda = \frac{\max_{a,b \geq 0} \Pr(A|M, \lambda_s = a, \lambda_n = b)}{\max_{b \geq 0} \Pr(A|M, \lambda_s = 1, \lambda_n = b)}$$

To formally compare two models, we follow the standard frequentist approach for phylogenetic model comparison by assuming that, when the null model holds, the statistic $-2 \log \Lambda$ converges in distribution to the χ^2 distribution with one degree of freedom. We then report a P-value for each window by halving the χ^2 distribution tail probability corresponding to lods (Ota et al., 2000).

5.4 Application to human genes

To annotate likely overlapping functional elements in human genes, we applied our local synonymous rate estimation procedure to sliding windows in all open reading frames in the human Consensus Coding Sequence (CCDS) catalog, a conservative set containing the ~85% of human gene annotations that are unanimously agreed on by the major gene catalogs (Pruitt et al., 2009). The vast majority of CCDS ORFs are aligned across at least

15 of the 29 placental mammals used in this study, with an average of four synonymous substitutions per codon site.

Specifically, we used ORF annotations from the 2009-03-27 build of CCDS (Pruitt et al., 2009) for NCBI version 36 of the reference human genome assembly. When CCDS annotates multiple isoforms of a single locus (as defined by multiple CCDS IDs in overlapping chromosomal regions sharing the same HGNC gene symbol), only the isoform with the longest coding sequence was analyzed. We extracted the alignments for these ORFs from the MULTIZ whole-genome alignments of 44 vertebrate species, generated by UCSC Genome Bioinformatics and used throughout the initial analysis of the 2× mammals data set. Only the "rows" of the MULTIZ alignments corresponding to the 29 available placental mammals were used, and the whole-genome alignments were "spliced" as necessary to produce an alignment of the complete human ORF.

Next, we defined the null models for use in the analysis. The relative rate estimation procedure described above does not make any assumptions about how the null model was originally estimated (except for reversibility of Q). In fact, we explored two different ways to estimate Q before providing it to this procedure. The first uses a parameterization equivalent to the M0 model of PAML, based on estimates of ω and κ, the transition/transversion rate ratio (Goldman and Yang, 1994; Yang et al., 2000). The second is an empirical codon model (ECM) that essentially amounts to an independent estimate for every entry in the 61×61 rate matrix (under the reversibility constraint), not restricted to single-nucleotide instantaneous substitutions (Kosiol et al., 2007). Comparing the two approaches, we found the ECM parameterization to be clearly superior to M0 for our purposes: it achieved better fit to the training data based on BIC and AIC scores, led to tighter distributions of λ_s and λ_n, and resulted in a more conservative overall test for synonymous constraint (slightly fewer rejections of the null hypothesis at any significance level). The ECM approach also has the advantage that it accounts for the CpG hypermutability effect and any other sequence-specific rate biases, to the extent possible under the assumption of independence between codon sites. The results described below are all based on the ECM parameterization.

With both parameterizations, the ORFeome-wide null model was fit to a random sample of 5% of the codon sites in autosomal CCDS ORFs. A separate chromosome-specific null model was used for genes on chromosome X, based on all codon sites on that chromosome. The ORF-specific null models were estimated from the complete alignment of each ORF. The topology of the mammalian species tree proposed in the 2× mammals analysis (Lindblad-Toh et al., 2011) was used in all models, while the branch lengths were re-estimated for each ORF in the ORF-specific models.

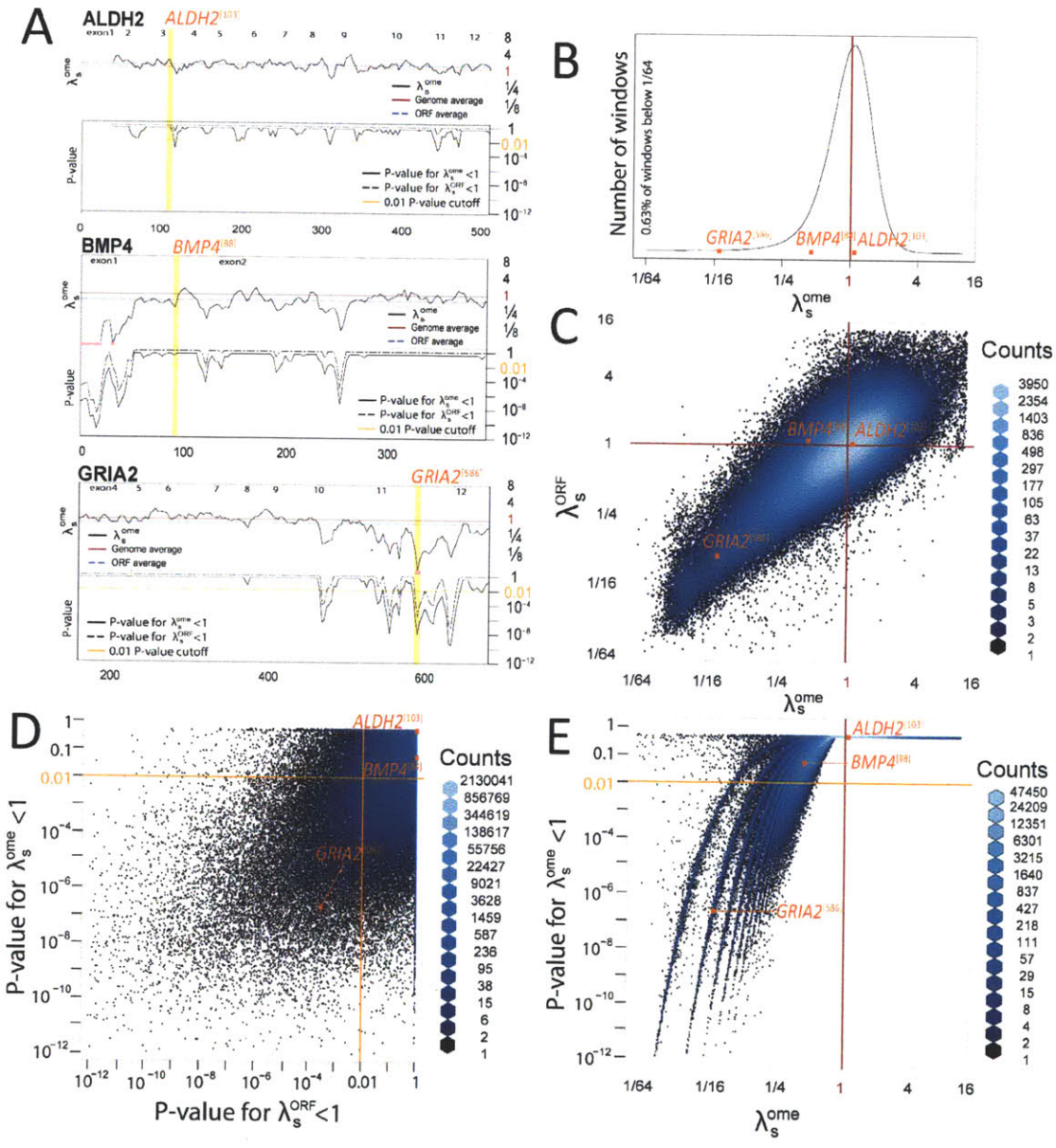
5.5 Genome-wide trends in synonymous rate variation

Before attempting to locate individual regions of synonymous constraint, we were immediately able to confirm a few notable genome-wide trends in synonymous rate variation that have been observed previously using different methods. Specifically, we observed marked reductions in average synonymous rates at the boundaries of coding exons, suggesting widespread evolutionary constraint on overlapping translation and splicing regulatory elements (Smith and Hurst, 1999; Baek, 2005; Xing and Lee, 2006; Chen, 2005; Parmley, 2005; Parmley and Hurst, 2007a). For example, the median λ_s^{ome} estimate for the first nine codons following the start codon in each ORF is only 0.81, indicating that the first several codons in a typical mammalian gene appear to have tolerated synonymous substitutions at a rate 19% below average ($P < 10^{-15}$, Mann-Whitney U test) (Figure 5-1B). The median estimated synonymous rate is also reduced by 13% in nine-codon windows spanning exon-exon junctions ($P < 10^{-15}$). Strikingly, the first nine to 12 codons of alternatively spliced cassette exons annotated in CCDS typically have an estimated synonymous rate 41% below average ($P < 10^{-15}$). We will revisit the possible translation and splicing regulatory roles suggested by these overall trends after establishing the statistical significance of the synonymous constraint in each individual window.

5.6 Synonymous constraint elements in CCDS ORFs

We next applied LRTs to identify individual windows that show statistically significant evidence of reduced synonymous substitution rates (Figure 5-2). Using three window sizes of nine, 15, and 30 codons, and sliding across each ORF by one-third of the window length, we selected windows passing three likelihood ratio tests, for the hypotheses that λ_s^{ome} is significantly below one, that λ_s^{ORF} is also significantly below one, and that λ_n^{ome}, the relative rate of non-synonymous substitutions, is not significantly above one. To account for the multiplicity of windows tested, we corrected the primary tests on λ_s^{ome} using the Benjamini and Hochberg false discovery rate (FDR)-controlling method (Benjamini and Hochberg, 1995), which tolerates local positive correlations (Benjamini and Yekutieli, 2001; Storey, 2003) , requiring estimated FDR < 0.01. The secondary tests on λ_s^{ORF} were Bonferroni-corrected for the number of windows tested in each ORF, requiring corrected $P < 0.01$. Benchmarks with simulated and permuted data confirmed the robustness of this strategy.

At the intermediate window length of 15 codons, 1.7% of the windows in CCDS ORFs meet these criteria. Overlapping significant windows collapse into 10,757 separate regions throughout human ORFs, covering 2.8% of all approximately 28 million CCDS


A
ALDH2
ALDH2[103]
exon1
Genome average
ORF average
P-value
P-value for λs ome <1
P-value for λs ORF <1
0.01 P-value cutoff
BMP4
BMP4[88]
exon2
GRIA2
GRIA2[586]
B
Number of windows
0.63% of windows below 1/64
GRIA2[586]
BMP4[88]
ALDH2[103]
λs ome
C
Counts
3950
2354
1403
836
498
297
177
105
63
37
22
13
8
5
3
2
1
λs ORF
D
P-value for λs ome <1
P-value for λs ORF<1
2130041
856769
344619
138617
55756
22427
9021
3628
1459
587
236
95
38
15
6
2
1
E
47450
24209
12351
6301
3215
1640
837
427
218
111
57
29
15
8
4
2
1

coding nucleotide positions. More than one-third of CCDS genes (6033/16,939) contain at least one such region. Notably, although the test threshold is $\lambda_s^{\text{ome}} < 1$, the median λ_s^{ome} among the windows passing the test is only 0.23, corresponding to a 77% reduced rate of synonymous substitution compared to the genome-wide average. Furthermore, the

---

Figure 5-2 *(facing page)*:
Identifying individual windows with statistically significant synonymous constraint. **(A)** Estimated synonymous rate relative to genome average (λ_s^{ome}) and corresponding P-value for the hypothesis $\lambda_s^{\text{ome}} < 1$ evaluated in nine-codon windows along the entire protein-coding regions of ALDH2, BMP4, and GRIA2, highlighting the windows corresponding to the three examples in Figure 1. For each plot, the top portion shows the λ_s^{ome} estimate for each window (black curve), the genome average (red line at $\lambda_s^{\text{ome}} = 1$), and the ORF average (blue dashed line). The bottom portion shows the statistical significance of the reduction in the synonymous rate estimate in each window, accounting for evidence in the cross-species alignments, using a likelihood ratio test for the hypothesis $\lambda_s^{\text{ome}} < 1$ (continuous black curve, using the genome average as the null model), and for the hypothesis $\lambda_s^{\text{ORF}} < 1$ (dashed black curve, using the ORF average as the null model). (Vertical gray lines) Exon boundaries; (orange) regions where λ_s^{ome} drops below 1/16th toward the 5' end of BMP4 and the 3' end of GRIA2. **(B)** Overall distribution of λ_s^{ome} estimates for all nine-codon windows across all CCDS genes. Heavy left tail indicates an excess of windows with very low estimated synonymous rates, shifting the mean ($\lambda_s^{\text{ome}} = 1$) to the left of the distribution mode, which likely represents neutral rates. **(C)** Comparison of synonymous rates estimated relative to genome-wide (λ_s^{ome}) and ORF-specific (λ_s^{ORF}) null models, each point denoting one nine-codon window, and density of overlapping points denoted by color. Joint distribution shows that low λ_s^{ome} estimates also usually correspond to low λ_s^{ORF} estimates, and therefore that the heavy tail observed in B does not reflect regional or ORF-wide deceleration, but instead localized constraints in small windows within each ORF, also visible in the three examples of A. **(D)** Comparison of P-values for synonymous rate reduction with respect to genome-wide (y-axis) and ORF-specific (x-axis) null models. Candidate synonymous constraint windows are selected when synonymous rate reductions are significant at $P < 0.01$ with respect to both null models (orange lines). Note that many windows are significant with respect to one null model but not the other. **(E)** Correspondence between λ_s^{ome} and the associated significance estimate for the each nine-codon window. The visible stripes in this plot arise from windows that are perfectly conserved except for one, two, three, or more synonymous substitutions observed in the extant species, while the position along each stripe reflects variation in the λ_s^{ome} estimate and its significance, determined by the species coverage, codon composition, and observed codon substitutions in each window. **(B–E)** The three example regions highlighted in A are shown in each distribution and density plot, with horizontal and vertical axes aligned. The orange line in plots A, D, and E denotes the statistical significance cutoff of $P < 0.01$, and the red line in plots A, B, C, and E denotes the genome-wide average $\lambda_s^{\text{ome}} = 1$ and $\lambda_s^{\text{ORF}} = 1$ for B. The *ALDH2*[103] synonymous rate is not significantly reduced either relative to the genome or to the *ALDH2* ORF; *BMP4*[88] is reduced relative to the genome but not relative to its ORF, which shows an overall reduced rate; *GRIA2*[586] is $> 80\%$ reduced relative to both the genome and its ORF, resulting in significant P-values for both.

estimates of λ_s^{ORF} in significant windows are also very low (best-fit line $\lambda_s^{\text{ORF}} = 0.78\lambda_s^{\text{ome}} + 0.06$ with $R^2 = 0.81$), confirming that these regions are generally not explained by ORF- or region-specific variation in sequence composition or evolutionary rates. Finally, the locally estimated synonymous and non-synonymous rates are not strongly correlated (Pearson coefficient between λ_s^{ORF} and λ_n^{ORF} of 0.04 in all windows, and 0.06 in significant windows), suggesting that our method largely succeeds in disentangling evolutionary pressures on synonymous and non-synonymous sites, when controlling for regional biases in composition and rates. (Due to such biases, the estimates of λ_s^{ome} and λ_n^{ome} do correlate somewhat, with Pearson coefficient 0.24 in all windows and 0.22 in significant windows.)

Windows with significant synonymous constraints are not unusually enriched on any of the individual human autosomes, although they are about twofold depleted on chromosome 19, which frequently stands out in genome-wide analyses owing to several unusual properties. In particular, the apparent depletion of synonymous constraints on this chromosome may be due to a calibration bias in our method arising from the chromosome's above-average G+C content and mutation rates (Lercher et al., 2001; Castresana, 2002), or it could reflect a genuine biological tendency related to the chromosome's unusually large complement of genes from a few tandem families (Grimwood et al., 2004). We analyzed ORFs on the X chromosome using a null model estimated from coding sites on that chromosome only and found that its resulting proportion of significant windows is lower than most autosomes, but still much greater than chromosome 19. We did not analyze the fewer than 100 protein-coding genes on the Y chromosome due to their eccentric, fast-evolving properties (Hughes et al., 2005, 2010; Kuroki et al., 2006).

In the longer windows of 30 codons, our method has increased statistical power to detect synonymous constraints since it combines evidence from more sites, and a larger proportion of windows reach significance with somewhat higher typical synonymous rate estimates, although they collapse into fewer separate regions. Conversely, a smaller proportion of the shorter nine-codon windows reach significance, with even lower estimated synonymous rates (Table 5.1). We also attempted our analysis with even smaller windows of six and three codons, but vanishingly few reached significance. Evidently, in these very short windows, even perfect conservation across the available species is usually not adequate to infer synonymous constraints using our current methodology and alignments.

Since the three window sizes lead to different trade-offs between resolution and discovery power, it is reasonable to expect them to identify somewhat different sets of regions as significant. In fact, of the regions obtained by collapsing overlapping significant windows at the 15-codon resolution, 24% are not detected at either the longer or shorter resolution. Similarly, 33% of the 30-codon regions and 28% of the nine-codon regions

Resolution (window size)	Short (9 codons)	Intermediate (15 codons)	Long (30 codons)
No. of windows tested	2,915,773	1,727,202	842,475
Proportion of windows significant	1.04%	1.72%	2.75%
Median significant λ_s^{ome}	0.1361	0.2299	0.3584
Median significant λ_s^{ORF}	0.1443	0.2572	0.4120
Maximum significant λ_s^{ome}	0.5497	0.6323	0.7134
Maximum significant λ_s^{ORF}	0.4682	0.5805	0.6520
No. of nonoverlapping synonymous constraint elements	11,882	10,757	8933
Proportion of 27,812,282 CCDS nucleotide positions within a synonymous constraint element	1.79%	2.82%	4.48%
Proportion of 16,939 CCDS ORFs containing a synonymous constraint element	35.8%	35.6%	33.3%

Table 5.1: Sliding windows in CCDS ORFs were tested for significantly reduced synonymous substitution rate estimates at different resolutions.

are detected only at those resolutions. As expected, the intermediate 15-codon resolution has the most overlap with the others, including 67% of the 30-codon and 72% of the nine-codon regions.

Hereafter, we refer to the collapsed significant regions as "synonymous constraint elements" (SCEs) and undertake numerous downstream analyses to show that they correspond to overlapping functional elements with diverse biological functions. We will perform most of these analyses based on SCEs identified at the 15-codon resolution, since they include most of the other sets, but we will also use the nine-codon and 30-codon resolutions based on the expected length of different types of overlapping functional elements. Similarly, we expect that each can be useful in different contexts for future follow-up studies.

5.7 Sequence composition and codon usage in SCEs

Since a major design goal for this study was to control for the specific codon sequence in each window, we carefully examined the sequence composition of the regions reported as significant. They do show certain subtle biases in sequence composition compared to other CCDS protein-coding sequences. For example, they have 3.4% lower G+C content at first and third codon positions (but 1.6% lower at second). CpG dinucleotides span 3.2% of second and third codon positions, compared to 2.4% in other coding regions, but they are slightly depleted spanning codon positions (1,2) and (3,1). Encoded serine residues are more frequent (9.2% of sites, compared to 8.2%) and tryptophan is less frequent (1.0% vs 1.2%). These and other nucleotide, dinucleotide and amino acid biases are statistically significant, but none represent more than a 1.4-fold enrichment or depletion, and most far

less. To put this in perspective, human ORFs show comparable or greater compositional variation from chromosome to chromosome: for example, ORFs on chromosome 3 are 1.4-fold enriched for ApA dinucleotides across codon positions (3,1), and chromosome 19 ORFs are twofold depleted for TpA's spanning codon positions (2,3).

Other compositional properties of the SCEs allowed us to rule out certain possible artifactual explanations for their low divergence across mammals, at least as predominant effects. First, short tandem and microsatellite repeats within coding regions, which can be maintained through evolution by non-selective processes (Richard et al., 2008), might resemble codons with conserved synonymous sites, but SCEs are depleted for such repeats relative to CCDS coding regions (0.87% vs. 2.03% of positions identified by TRF or the appropriate subset of RepeatMasker annotations). Second, biases in DNA mismatch repair can lead to a reduced apparent neutral substitution rate within some genomic isochores (Chamary et al., 2006), but the lower G+C content at third codon positions in SCEs is contrary to the trend in regions that have experienced such biased conversions (Galtier and Duret, 2007). Third, codon usage biases related to translational efficiency or background compositional effects can reduce synonymous divergence without additional overlapping function, and while the relative usage frequencies of synonymous codons do tend to differ between SCEs and other coding regions, the overall codon usage in SCEs is actually slightly less biased than average: the Effective Number of Codons (ENC) (Wright, 1990; Fuglsang, 2005) in all SCEs taken together is 56.8, higher than in non-significant regions (54.4). Similarly, the ENC in a typical ORF containing an SCE (median 51.2) is slightly higher than in other ORFs (median 50.1). In contrast, the ENC is markedly reduced in ORFs with strong codon bias owing to isochore-specific G+C composition (Chamary et al., 2006).

In summary, the SCEs show significant but not extreme differences in sequence composition and codon usage compared to other coding regions, which probably reflect the sequence-dependent biological nature of the overlapping functional elements they encode. They may also partly reflect biases in our phylogenetic models arising from contextual effects they do not accurately captured, but the observed compositional biases do not suggest debilitating shortcomings of our overall approach. In particular, the lack of increased codon usage bias in SCEs indicates that, as intended, our method excluded regions explained by this effect.

5.8 Discussion

In this chapter, we showed that thousands of human protein-coding genes contain short regions with conspicuously low estimated synonymous rates across placental mammals. We located between 9000 and 12,000 SCEs, depending on the resolution chosen, within one-third of all CCDS ORFs (Table 5.1), or well over one-quarter of all human protein-coding genes. These SCEs are likely to encode a variety of overlapping functional elements, as we investigate in the next chapter.

One major challenge in the design of our analysis, shared with nucleotide-level constraint detection methods, lies in the estimation of null models. Nucleotide-level methods typically calibrate their null models on presumptively neutral regions such as ancestral repeats or fourfold degenerate sites, but these were obviously not suitable for our purposes because our null models must capture the typical evolutionary rates of both synonymous and non-synonymous sites in coding regions. Therefore, we simply calibrated our null models to the ORFeome-wide or ORF-level background, averaging in any unusually evolving sites. If we assume that purifying selection is much more common than positive selection in synonymous sites (Resch et al., 2007), then this approach leads to a somewhat conservative test for synonymous constraint, providing one of our countermeasures against the possibility of background rate variation leading to spurious inferences of selection. More accurate null model calibration is an important direction for future investigation, perhaps by explicitly modeling statistical distributions of synonymous rates (Pond, 2005; Rodrigue et al., 2008).

Still, despite our prudently conservative null model calibration, we were able to achieve far greater discovery power than previous efforts to locate regions of synonymous constraint in mammalian genes (Schattner, 2006; Parmley and Hurst, 2007b), which identified at most $\sim$2% of our SCEs, and at much lower resolution. This is attributable both to the many informant species now available and to the rigorous phylogenetic methodology we devised to take advantage of them, based on maximum likelihood estimates of the synonymous substitution rates in short windows and formal statistical tests for their reduction. Naturally, this methodology can accommodate additional sequenced genomes and improved assemblies and alignments as they become available.

Chapter 6

Initial survey of biological functions encoded by synonymous constraint elements

6.1 Introduction

In this chapter, we attempt to attribute specific biological functions to many of the ~10,000 SCEs identified in the previous chapter, including splicing and translational regulation, dual-coding regions, RNA secondary structures, and developmental enhancers. While numerous individual known examples and previously-observed statistical trends in synonymous rates provide a great deal of existing evidence that SCEs play roles in these processes (Chamary et al., 2006), our high-resolution catalog of the exact locations of SCEs throughout the genome provides a wealth of opportunities to gain new biological insights in both large categories and individual examples. This initial survey is, however, only a first step to revealing the biological roles of the individual SCEs.

6.2 Characteristics of genes containing SCEs

We first studied overall properties of the 6033 genes containing SCEs (at 15-codon resolution). Compared to the remaining CCDS genes, the typical gene containing an SCE has a much longer ORF (median 558 vs. 356 codons). This is not actually longer than expected based on drawing genes randomly weighted by their ORF length, but the genes containing SCEs also have more introns (median nine vs. five), lengthier individual introns (1727 nt vs. 1261 nt), and they span much larger genomic regions (36,000 nt vs. 11,000 nt), suggesting that the overall length distribution is entangled with the well-established cor-

relations between gene length and other relevant characteristics including conservation, functional categories, and expression levels (Castillo-Davis et al., 2002; Urrutia, 2003; Stanley et al., 2006; Pozzoli et al., 2007). The genes containing SCEs also appear to be under stronger purifying selection on their amino acid sequences, as the median estimate of $\omega = d_N/d_S$ measured across each complete ORF is 0.068, much lower than the 0.138 for other genes, despite containing regions with greatly reduced d_S.

Next, we analyzed Gene Ontology (GO) annotations for the genes containing SCEs. While 36% of CCDS genes contain a 15-codon SCE, they include 70% of genes annotated with the term "chromatin modification," a twofold enrichment (Bonferroni-corrected hypergeometric $P < 7.9 \times 10^{-13}$). Additionally, they include most of the genes in these and related categories: "'ubiquitin-protein ligase activity" (1.8-fold, $P < 2.9 \times 10^{-6}$), "ion channel complex" (1.7-fold, $P < 5.6 \times 10^{-5}$), "nervous system development" (1.6-fold, $P < 6.8 \times 10^{-9}$), "transcription factor activity" (1.6-fold, $P < 3.5 \times 10^{-6}$), and "RNA splicing" (1.5-fold, $P < 8.4 \times 10^{-4}$). These enrichments remain strongly significant when controlling for the varying ORF lengths, and suggest a few interesting hypotheses about genes that encode overlapping functional elements. For example, the enrichment for genes encoding chromatin modification and RNA splicing functions could suggest the existence of autoregulatory circuits for many such genes, similar to known examples such as ADARB1, which edits its own pre-mRNA and causes a change in its splicing (Rueter et al., 1999), and DGCR8, which binds its own mRNA and causes it to be cleaved by Drosha (also know as RNASEN) (Han et al., 2009). Also, the enrichment for ion channel genes suggests a connection with A-to-I editing, since several such genes are known targets of this recoding mechanism (Bass, 2002); we explore this further below.

6.3 More than one-third of short SCEs can be provisionally assigned roles in transcript splicing or translation initiation

As expected, based on the aforementioned general trends in synonymous rate variation (above; Figure 5-1B), many SCEs can be provisionally classified as possible splicing regulatory elements based on their location within gene models. In particular, 34.7% of the nine-codon SCEs span an exon-exon junction, compared to only 20.3% of a set of random control regions placed uniformly throughout CCDS ORFs, with matching length distribution and total number.

Interestingly, the introns flanked by these SCEs tend to have weaker 3' splice acceptor

sites (nine-codon resolution; $P = 3.9 \times 10^{-7}$, Mann-Whitney U test), based on analysis of the sequence information content of the surrounding nucleotides (Yeo and Burge, 2004). This is consistent with the hypothesis that while "strong" splice sites are constitutively recognized by the splicing machinery, the activity of weaker splice sites is more reliant on additional nearby cis-regulatory sequences, increasing the possibilities for their combinatorial and condition-specific regulation (Fairbrother et al., 2002; Chen and Manley, 2009).

We also examined individual exons with multiple alternative acceptor or donor sites, for which the coding sequence of the longest exon isoform additionally encodes splice sites for shorter isoforms and perhaps additional splicing regulatory elements (e.g., Figure 6-1A). Of 551 such alternative donor sites in RefSeq transcripts encoded within the CCDS exons we analyzed, 84 (15.2%) fall within SCEs, compared to only 1% in the random regions ($P < 10^{-21}$). Similarly, 57 of 576 (9.9%) alternative acceptor sites lie within SCEs (1.7% random; $P < 10^{-21}$).

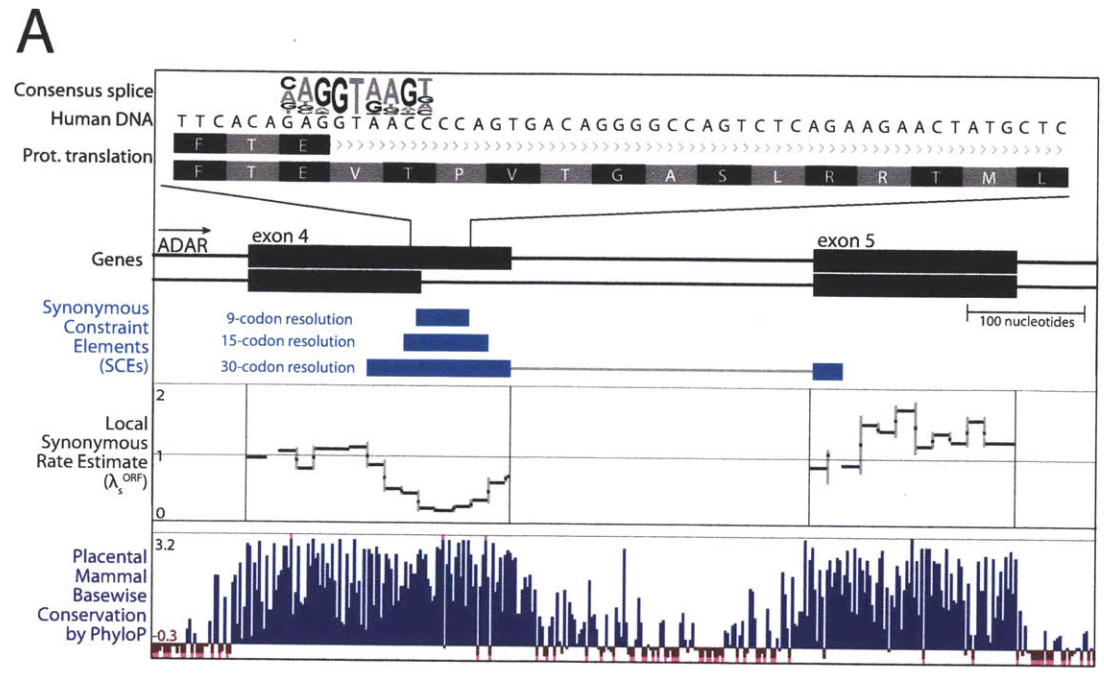
SCEs are also enriched for elements potentially involved in translation initiation. The first nine-codon window (beginning at the first site following the start codon) in 3.9% of CCDS ORFs is found to be under synonymous constraint, a strong enrichment compared to 1.04% of all windows. Additionally, there are 744 RefSeq-annotated internal translation initiation sites encoded within longer coding exons of CCDS ORFs, of which 5.4% fall within SCEs (e.g., Figure 6-1B) compared to 1.6% in the random regions ($P < 10^{-9}$).

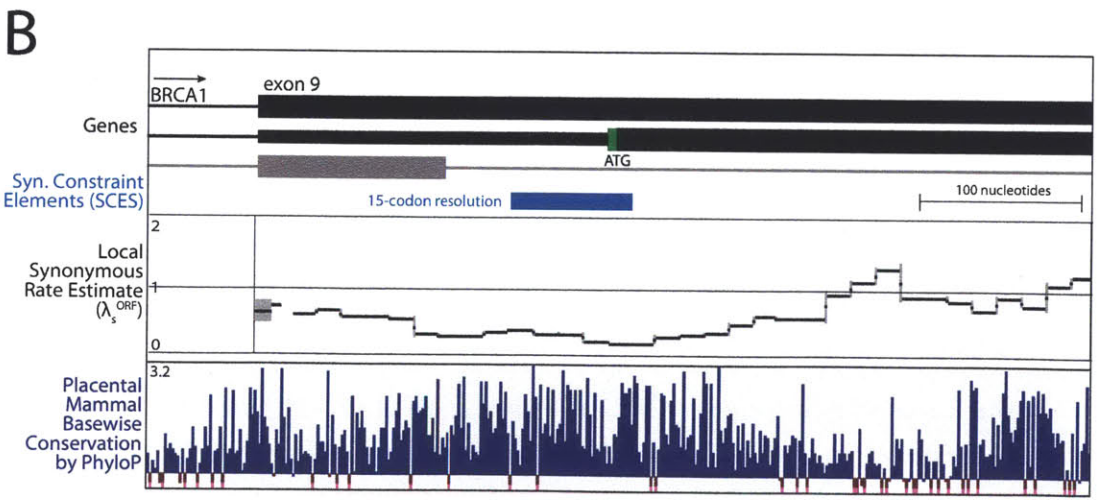
Taken together, these provisional classifications of possible splicing and translation regulatory elements account for slightly more than one-third of the SCEs (at nine-codon resolution). Conversely, nearly two-thirds are not found in locations that directly suggest such roles, although it should be noted that some splicing regulatory sequences may act at considerable distances (Parmley and Hurst, 2007a; Parmley et al., 2007).

Synonymous constraint at an alternate translation start site in BRCA1

One noteworthy example of an SCE with a possible role in translation initiation is found within the tumor suppressor gene *BRCA1*. Hurst and Pal (2001) first observed that an extended region within this ORF shows unusually low synonymous substitution rates, based on sliding windows of 100 codons in pairwise comparisons among the human, mouse, and rat orthologs. Recently, however, Schmid and Yang (2008) raised certain issues with their statistical methods and argued that their result was artifactual.

Our analysis is based on much more data than both previous studies and strongly supports the original conclusion of Hurst and Pal (2001). Within the most significant


A
Consensus splice
Human DNA
T T C A C A G A G G T A A C C C C A G T G A C A G G G G C C A G T C T C A G A A G A A C T A T G C T C
Prot. translation
F T E
F T E V T P V T G A S L R R T M L
Genes
ADAR
exon 4
exon 5
Synonymous Constraint Elements (SCEs)
9-codon resolution
15-codon resolution
30-codon resolution
100 nucleotides
Local Synonymous Rate Estimate (λ_s^{ORF})
2
1
0
Placental Mammal Basewise Conservation by PhyloP
3.2
-0.3


B
BRCA1
exon 9
Genes
ATG
Syn. Constraint Elements (SCES)
15-codon resolution
100 nucleotides
Local Synonymous Rate Estimate (λ_s^{ORF})
2
1
0
Placental Mammal Basewise Conservation by PhyloP
3.2

15-codon window in *BRCA1*, the estimated rate of synonymous substitution in placental mammals is reduced by 80% ($\lambda_s^{\text{ome}} = 0.20$, $P < 9.9 \times 10^{-8}$), ranking among the slowest 1% of windows in human ORFs, and slower than average for SCEs at this resolution. Like all SCEs, the window also has a very low synonymous rate estimate compared to the *BRCA1* ORF specifically ($\lambda_s^{\text{ORF}} = 0.23$, $P < 2.3 \times 10^{-5}$). Due to the controversy over this region in particular, we also performed auxiliary permutation tests that further confirmed its statistical significance.

Furthermore, the much higher resolution of our analysis precisely localizes the significant region of synonymous constraint to the annotated translation initiation site of a CCDS-supported alternative splice form of *BRCA1* (Figure 6-1B), immediately suggesting a hypothesis for an overlapping biological function – namely, a role in regulating translation initiation in this splice form. This highly suggestive positional association may have been much less clear at the 100-codon resolution used by both previous studies. Indeed, the synonymous constraint is not significant at the 30-codon resolution according to our analysis, possibly corroborating the statistical concerns raised by Schmid and Yang (2008), while nonetheless confirming and extending the main conclusion of Hurst and Pal (2001).

Figure 6-1 *(facing page)*: Examples of candidate synonymous constraint elements (SCEs) with likely roles in splicing and translation regulation. **(A)** Predicted SCEs (light blue) overlapping two isoforms of ADAR exon 4 (black) arising from an alternative splice donor site encoded within the longer exon variant. With increasing resolution, the SCE is more precisely localized to the region of overlap with the alternative splice site (motif logo for human donor sites rendered by WebLogo) (Crooks et al. 2004). The localization of the synonymous constraint to the splice site is also seen in the local synonymous rate estimate λ_s^{ORF} (relative to the ORF average). Note that the significant reduction in the synonymous rate is not obvious from the nucleotide-level conservation measure (dark blue, bottom panel). The extent of the predicted SCE may suggest the presence of additional splicing regulatory elements downstream from the alternative splice site. **(B)** Predicted SCE (light blue) overlapping an alternate translation initiation site (green) in BRCA1 encoded within exon 9 of a longer isoform. Synonymous constraint ranges from shortly upstream to immediately downstream of the alternate start codon, suggesting this region may be involved in regulating translation initiation at the alternate site. The region just upstream of the predicted SCE also shows a reduced synonymous rate (black curve) overlapping an alternative splice donor site for a third BRCA1 isoform (gray), although this reduction is not statistically significant. Annotation visualizations in Figures 6-1 and 6-2 are based on the UCSC Genome Browser (Kent et al., 2002).

6.4 SCEs in known and novel dual-coding genes

Genomic sequences can simultaneously encode different amino acid sequences in multiple reading frames, a common phenomenon in viral genomes but rare in animal genomes. Such "dual-coding" regions can involve ORFs on the same strand, but in an alternate "shifted" reading frame, which can be mediated by alternative splicing, internal translation initiation, or ribosomal frameshifting. Of the six long human dual-coding gene structures for which likely biological functions have been demonstrated (Sharpless and DePinho, 1999; Klemke, 2001; Yoshida et al., 2001; Hameed et al., 2003; Poulin, 2003; Ahmed et al., 2008), all show at least some evidence of overlapping evolutionary constraint in our analysis: *XBP1*, *GNAS*, and the *ANKHD1/EIF4EBP3* fusion transcript contain SCEs in their dual-coding regions; the dual-coding 3' end of *IGF1* narrowly missed our threshold with a 47% reduced synonymous rate; and the dual-coding regions of *CDKN2A* and *LRTOMT* have very low synonymous rates, but were excluded from SCEs because of elevated non-synonymous rates. Aside from these individually studied examples, CCDS annotates 237 other exons as protein-coding in multiple reading frames of the same strand, 24 (10.1%) of which contain SCEs, compared to six (2.5%) containing random control regions. This lower overlap might suggest dual-coding regions not under selection across placental mammals.

Alternatively, both strands of the genomic DNA can encode different protein sequences, expressed in "sense" and "antisense" transcription units. At least one such case has been thoroughly studied: an ~200-nt sense/antisense dual-coding sequence of the convergent transcription units for *THRA* and *NR1D1* (Hastings, 2000). Indeed, we find synonymous constraints in both ORFs precisely coinciding with the known dual-coding region (Figure 6-2A). We similarly detect synonymous constraints in 10 of 44 individual exons that CCDS annotates on both strands (five in random regions).

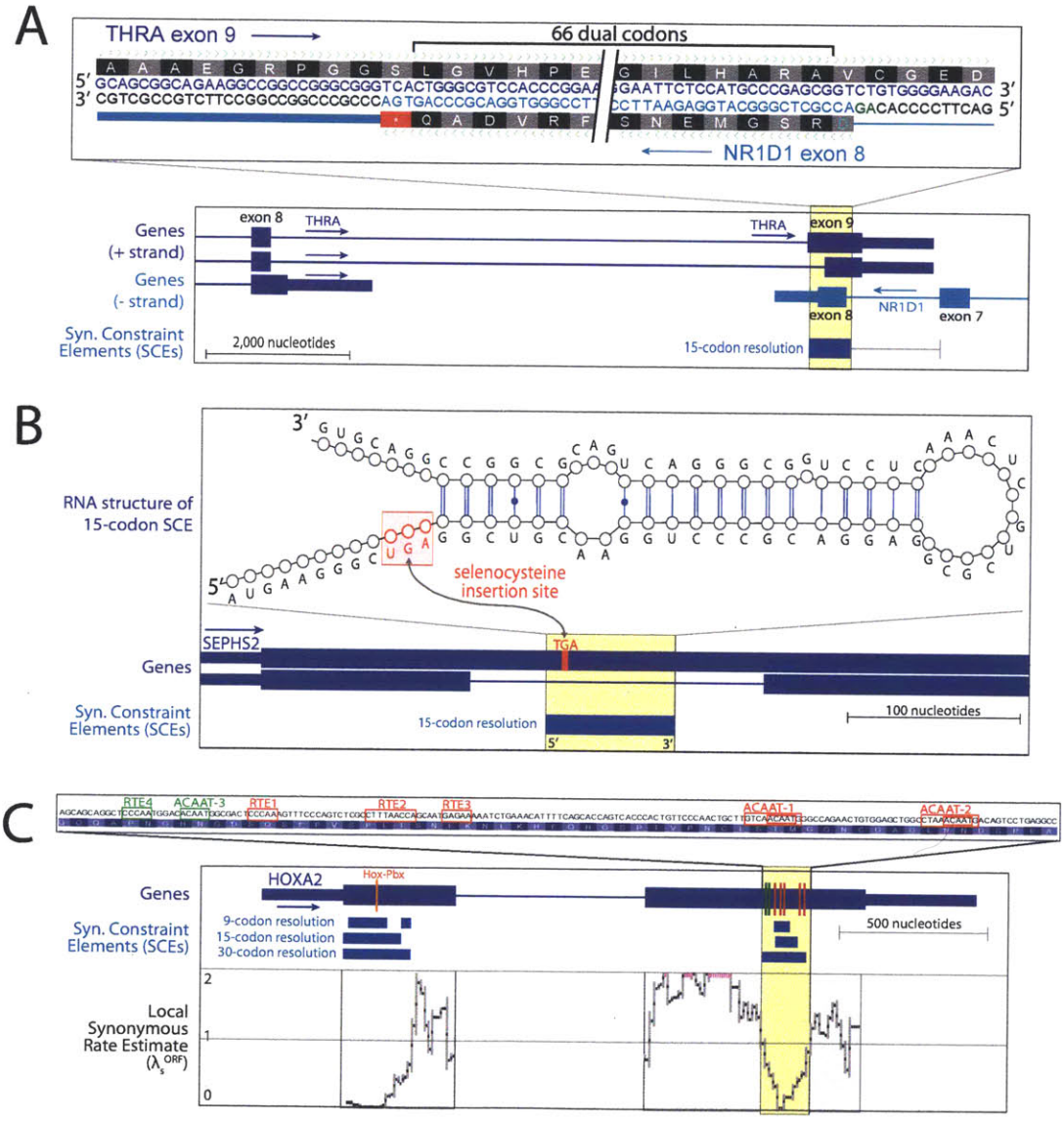
In addition to these known examples, we found 19 candidate novel dual-coding ORFs in alternate reading frames of known CCDS ORFs, which contain one or more of our 30-codon SCEs, are longer than expected by chance, and also appear to be depleted of stop codons in the other mammals. Twelve of these 19 are encoded on the same strand as the CCDS gene structure, some spanning multiple exons, and the remaining seven are found antisense to individual CCDS exons. A few are further supported by transcript cDNA evidence or similarity to known proteins. This preliminary assessment suggests that the SCEs probably capture several additional long dual-coding mammalian gene structures, although specialized methods for detecting the evolutionary signatures unique to dual-coding regions (Chung et al., 2007; Ribrioux et al., 2008) would probably have more power applied to the new set of 29 species.

6.5 SCEs capture most known A-to-I recoding sites

A-to-I editing is a recoding mechanism in which certain adenosine (A) bases in RNA transcripts are edited to inosine (I), which is read as a guanine (G) (Bass, 2002). This mechanism is essential for normal development of the mammalian nervous system, and, because the enzymes that mediate the reaction target double-stranded RNA, known A-to-I recoding sites conserved between human and mouse transcripts typically show extensive conservation of flanking sequence, presumably reflecting the interlocking constraints of encoding amino acids and pairing with another portion of the transcript (Aruscavage and Bass, 2000; Hoopengardner et al., 2003).

Indeed, 10 of 14 known human A-to-I recoding sites in CCDS ORFs lie within SCEs (15-codon; none within random regions), although this is not surprising since exceptional conservation was one signature originally used to identify many of the known sites. A recent human-specific study used a high-throughput sequencing approach not biased for highly conserved regions (Li et al., 2009) to identify 40 new edited sites within CCDS ORFs, only three of which also lie within our SCEs (two amino acid changing sites in *CADPS*, *FLNB*, and a synonymously edited site in *GRIA2*). Therefore, consistent with that study's report that the 37 remaining experimentally identified sites lack extensive nucleotide-level conservation, we do not find specific evidence for synonymous constraint in other mammals.

Figure 6-2 *(facing page)*: Synonymous constraint elements (SCEs) corresponding to dual-coding, selenocysteine insertion, and expression enhancer functions. **(A)** A large SCE (blue) fully encompasses a 66-codon sense/antisense dual-coding region in the convergent transcripts of *THRA* and *NR1D1*. The SCE is specifically localized to the overlapping exons, while upstream exons of each gene are excluded. **(B)** A predicted SCE in the selenoprotein-encoding gene *SEPHS2* encompasses the selenocysteine insertion site (red) and a predicted RNA hairpin structure (minimum free energy fold rendered by VARNA) (Darty et al. 2009) immediately downstream from the selenocysteine codon. Inferred structure is similar to a hairpin known to stimulate selenocysteine recoding in *SEPN1* (Howard et al. 2005). **(C)** Two SCEs are found within the *HOXA2* ORF, each corresponding to a different enhancer element regulating expression of the mouse ortholog in distinct segments of the developing hindbrain. The 5' element encodes a Hox-Pbx responsive element and drives expression in rhombomere 4 (Lampe et al. 2008), and the 3' element encodes *SOX2* binding sites and drives expression in rhombomere 2 (Tumpel et al. 2008). The 3' element includes several RTE and ACAAT motif instances that were investigated by site-directed mutagenesis in the latter study (red), as well as two additional upstream instances (green). SCEs are also found within most other HOX genes.


A
THRA exon 9
66 dual codons
NR1D1 exon 8
Genes (+ strand)
Genes (- strand)
Syn. Constraint Elements (SCEs)
exon 8
THRA
exon 9
NR1D1
exon 7
2,000 nucleotides
15-codon resolution
B
RNA structure of 15-codon SCE
selenocysteine insertion site
SEPHS2
TGA
Genes
Syn. Constraint Elements (SCEs)
15-codon resolution
100 nucleotides
C
RTE4
ACAAT-3
RTE1
RTE2
RTE3
ACAAT-1
ACAAT-2
HOXA2
Hox-Pbx
Genes
Syn. Constraint Elements (SCEs)
9-codon resolution
15-codon resolution
30-codon resolution
500 nucleotides
Local Synonymous Rate Estimate (λ_s^{ORF})

6.6 SCEs lie within most Hox genes and include two known developmental enhancers

Many of the lengthiest and most strikingly conserved SCEs are found within 27 of the 40 genes in the four Hox clusters. For example, the first 60 codons of *HOXB5* exhibit absolutely no synonymous substitutions in any of the species in our alignment of its ORF. More generally, we identify SCEs in eight of the 11 genes in the HoxA cluster, seven of nine HoxB genes, seven of nine HoxC genes, and three of nine HoxD genes, as well as the *EVX1* and *EVX2* homeobox-encoding genes adjacent to the HoxA and HoxD clusters. Lin et al. (2008b) also noted striking regions of synonymous constraint in Hox genes, which they defined as stretches of at least 40 codons without any synonymous substitutions in pairwise comparisons. Our results confirm their findings while also providing much greater power and resolution, locating several additional shorter and/or less extremely conserved SCEs.

Remarkably, the two SCEs found within *HOXA2* correspond to known tissue-specific enhancers that regulate expression in distinct segments of the developing mouse hindbrain (Figure 6-2C). A lengthier region (~200 bp) in the upstream exon encodes a Hox-Pbx responsive element and drives *Hoxa2* expression in rhombomere 4 (Lampe et al., 2008), and a shorter region (~75 bp) in the downstream exon encodes *SOX2* binding sites and drives expression in rhombomere 2 (Tumpel et al., 2008). Considering these two examples, the SCEs could suggest the existence of a largely unknown regulatory network relying on nucleotide sequence elements embedded within the ORFs of most of these key developmental genes (Woltering and Duboule, 2009).

6.7 Discussion

In this chapter, we presented preliminary results implicating many of our ~10,000 SCEs in a variety of biological roles, including tissue-specific developmental enhancers, splicing and translation regulatory elements, and RNA structures involved in A-to-I editing or selenocysteine incorporation. Altogether however, these provisional explanations do not even account for half of all the SCEs, suggesting that there may be many other overlapping biological roles yet to be elucidated. Therefore, just as our view of nucleotide-level conserved elements throughout the genome has been greatly refined since the first human/mouse comparisons nearly a decade ago, we expect that this initial survey of SCEs can motivate many additional computational and experimental studies to reveal their biological functions.

In addition to identifying candidate overlapping functional elements, our extensive annotation of synonymous sites under selection in placental mammals can help refine many types of evolutionary and functional analyses that typically assume they are neutral. For example, widely used methods for detecting positive and negative selection on amino acid sites are based on the ratio of non-synonymous and synonymous substitution rates ($\omega = d_N/d_S$). Since the extreme drop in d_S observed in SCEs would naturally tend to elevate local estimates of ω, any claims of positive selection on the amino acid sites encoded within SCEs ought to be regarded with caution (Xing and Lee, 2006; Parmley and Hurst, 2007b). Our results can also inform future disease association studies and other types of population genetics analyses, which frequently ignore synonymous SNPs (Chamary et al., 2006). The SCEs we identified represent specific regions in which synonymous SNPs may well have significant consequences and should therefore be included in such analyses.

Chapter 7

Lineage-specific and high-resolution analysis of synonymous constraint elements

7.1 Introduction

In this chapter, we build on the work described in the previous two chapters to ask additional biological questions and increase the resolution even further. First, we extend our previous methodology to detect SCEs that evolved within a specific subtree of the phylogeny under analysis. We apply this new lineage-specific test to show that a substantial fraction of the SCEs detected in our original study appear to have evolved within placental mammals when compared to other tetrapod species, including several SCEs found within Hox genes. Second, we develop a new Bayesian statistical framework for estimating synonymous substitution rates, designed to operate reliably on just a few codon sites and even at individual-site resolution. We apply this new high-resolution method to produce new genome-wide synonymous constraint tracks, which we also use to further study a controversial SCE found within the tumor suppressor gene *BRCA1*, and to predict potential miRNA target sites embedded within human coding genes. Third, we present and apply a hidden Markov model for delineating discrete SCEs based on the new high-resolution analysis, which complements the sliding-window approach used in our previous study. We present evidence that these SCEs have continued to evolve under selection in human populations. The genome-wide annotation datasets produced by our new methods can inform future studies on mammalian gene structures, human disease associations, and personal genome interpretation.

7.2 Likelihood ratio test for lineage-specific SCEs

We first extend our previous LRT methodology to detect lineage-specific synonymous constraint elements, which show a reduced synonymous rate only within a certain subtree of the phylogeny under analysis. We refer to this subtree as the "foreground" and the remainder of the phylogeny as the "background." Our lineage-specific test introduces separate scale factors for synonymous and non-synonymous rates in the foreground, λ_s^{fg} and λ_n^{fg}, and the background, λ_s^{bg} and λ_n^{bg}. These parameters determine adjusted foreground and background rate matrices that are effective on the respective branches of the tree, similar to the effect of λ_s and λ_n on the single adjusted rate matrix effective throughout the tree in our original test (T is held fixed in both). This approach is similar to the "branch model" for detecting lineage-specific changes in the d_N/d_S ratio (Yang, 1998, 2007).

We also define a new likelihood ratio test to determine statistical significance for lineage-specific synonymous constraint. Specifically, suppose we estimate the parameters by maximum likelihood and find the estimated $\lambda_s^{\text{fg}} < \lambda_s^{\text{bg}}$. We would then like to determine whether reducing the synonymous rate specifically in the foreground subtree leads to a significantly better model fit than assuming a uniformly reduced synonymous rate throughout the tree. We define an appropriate likelihood ratio statistic,

$$\Lambda = \frac{\max_{a,b,c,d \geq 0} \Pr(A|M, \lambda_s^{\text{fg}} = a, \lambda_s^{\text{bg}} = b, \lambda_n^{\text{fg}} = c, \lambda_n^{\text{bg}} = d)}{\max_{b,c,d \geq 0} \Pr(A|M, \lambda_s^{\text{fg}} = \lambda_s^{\text{bg}} = b, \lambda_n^{\text{fg}} = c, \lambda_n^{\text{bg}} = d)}$$

This allows us to estimate a P value for lineage-specific synonymous rate reduction based on an assumed χ_1^2 distribution of $2 \log \Lambda$ under the null hypothesis that $\lambda_s^{\text{fg}} = \lambda_s^{\text{bg}}$. (The power of the test can be increased using a slightly different null distribution corresponding to a "one-sided" test (Ota et al., 2000).) Benchmarks on simulated alignments confirmed that this test controls false positives at the nominal rate.

A desirable property of this rate LRT approach is that it would not consider lineage-specific synonymous constraint more likely based on a simple lack of alignment in the background species, which could be due to numerous other explanations, such as incomplete genome sequence coverage or complex duplication/paralogy relationships. The likelihood calculations in our approach remove unaligned informant species from consideration through marginalization. Thus, our test requires the presence of putatively orthologous alignments with at least some background species, showing increased synonymous rates, in order to infer lineage-specific synonymous constraint.

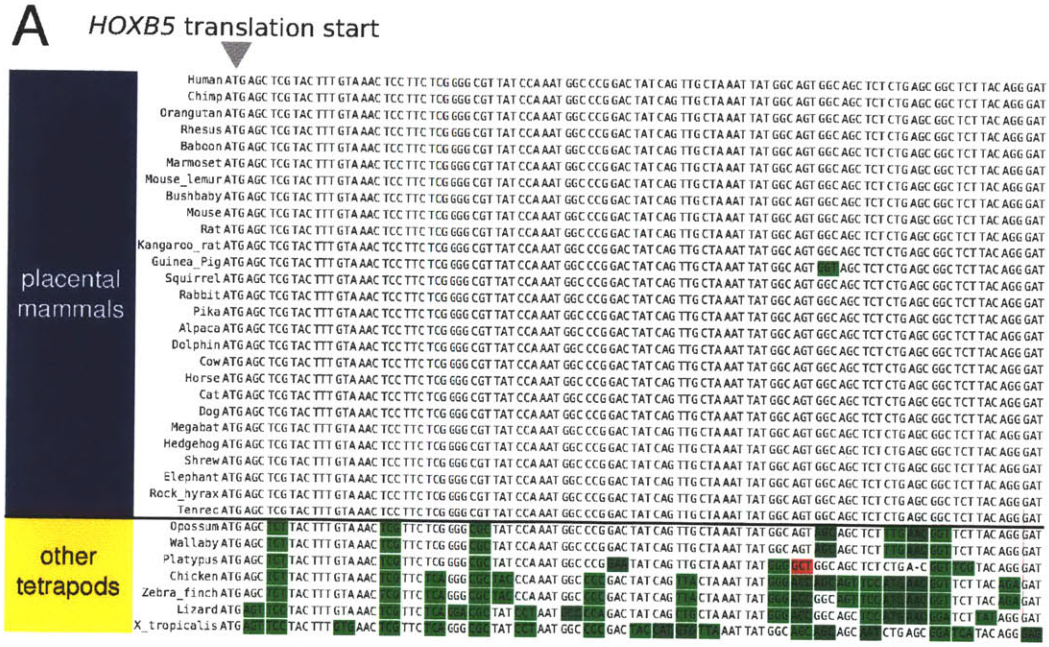
7.2.1 Many previously defined SCEs appear specific to placental mammals, including several in Hox genes

Our previous work identified ~10,000 SCEs in the human genome based on genome alignments of 29 placental mammals. We revisited these regions to test for synonymous constraint specific to the placental lineage, based on newer genome alignments including 7 other tetrapods (marsupials, avians and amphibians) as well as 33 placental mammals, obtained from the UCSC Genome Browser (Blanchette et al., 2004; Fujita et al., 2010). We applied our lineage-specific LRT to each of the original SCEs using an MG94–F3×4 null model estimated from a random sample of codon sites from all autosomal ORFs in the CCDS annotation (2009-09-02 build) (Anisimova and Kosiol, 2008; Delport et al., 2008; Pruitt et al., 2009).

The results provide evidence that a substantial fraction of our original SCEs evolved in placental mammals. Of the 10,757 original autosomal SCEs, 2547 (23.6%) pass the LRT with false discovery rate (FDR) < 0.001. These 2547 candidate lineage-specific SCEs typically show a greater than four-fold reduction in the estimated synonymous rate in placental mammals compared to the other tetrapods, in which they show approximately average rates (median MLE $\lambda_s^{\text{fg}} = 0.24$, $\lambda_s^{\text{bg}} = 1.07$). Furthermore, benchmarks on simulated alignments indicate that our test has ~90% sensitivity for detecting lineage-specific SCEs with $\lambda_s^{\text{fg}} = 0.25$ and $\lambda_s^{\text{bg}} = 1.0$ in the mammal/tetrapod alignment (under the assumptions of our codon models), suggesting that these constitute the bulk of our original SCEs that are highly specific to mammals.

A striking example of an apparent lineage-specific SCE is found at the 5' end of the *HOXB5* ORF, which is one of the lengthiest and most highly conserved mammalian SCEs. The first sixty codons of this ORF are almost perfectly conserved among placental mammals, but show numerous substitutions compared to, and within, the other tetrapods (Figure 7-1A). The synonymous rates estimated in our codon models indicate a seven-fold reduction in placental mammals compared to the other tetrapods ($P < 10^{-10}$), and the example ranks in the top 4% of all the original SCEs by statistical significance according to our lineage-specific LRT.

More generally, 25 of the 40 SCEs we originally found within Hox genes pass the lineage-specific test, including 10 of the 13 found within the Hox C cluster (Figure 7-1B). While the biological functions of these Hox SCEs remain largely unknown, there is evidence that some play roles as tissue-specific expression enhancers in the developing embryo (Lampe et al., 2008; Tumpel et al., 2008; Woltering and Duboule, 2009; Fredman et al., 2011). Along with a previous study of Hox SCEs (Lin et al., 2008b), our results suggest that several of these – as well as many other SCEs throughout the genome –


A
HOXB5 translation start
placental mammals
other tetrapods
Human ATG AGC TCG TAC TTT GTA AAC TCC TTC TCG GGG CGT TAT CCA AAT GGC CCG GAC TAT CAG TTG CTA AAT TAT GGC AGT GGC AGC TCT CTG AGC GGC TCT TAC AGG GAT
Chimp ATG AGC TCG TAC TTT GTA AAC TCC TTC TCG GGG CGT TAT CCA AAT GGC CCG GAC TAT CAG TTG CTA AAT TAT GGC AGT GGC AGC TCT CTG AGC GGC TCT TAC AGG GAT
Orangutan ATG AGC TCG TAC TTT GTA AAC TCC TTC TCG GGG CGT TAT CCA AAT GGC CCG GAC TAT CAG TTG CTA AAT TAT GGC AGT GGC AGC TCT CTG AGC GGC TCT TAC AGG GAT
Rhesus ATG AGC TCG TAC TTT GTA AAC TCC TTC TCG GGG CGT TAT CCA AAT GGC CCG GAC TAT CAG TTG CTA AAT TAT GGC AGT GGC AGC TCT CTG AGC GGC TCT TAC AGG GAT
Baboon ATG AGC TCG TAC TTT GTA AAC TCC TTC TCG GGG CGT TAT CCA AAT GGC CCG GAC TAT CAG TTG CTA AAT TAT GGC AGT GGC AGC TCT CTG AGC GGC TCT TAC AGG GAT
Marmoset ATG AGC TCG TAC TTT GTA AAC TCC TTC TCG GGG CGT TAT CCA AAT GGC CCG GAC TAT CAG TTG CTA AAT TAT GGC AGT GGC AGC TCT CTG AGC GGC TCT TAC AGG GAT
Mouse_lemur ATG AGC TCG TAC TTT GTA AAC TCC TTC TCG GGG CGT TAT CCA AAT GGC CCG GAC TAT CAG TTG CTA AAT TAT GGC AGT GGC AGC TCT CTG AGC GGC TCT TAC AGG GAT
Bushbaby ATG AGC TCG TAC TTT GTA AAC TCC TTC TCG GGG CGT TAT CCA AAT GGC CCG GAC TAT CAG TTG CTA AAT TAT GGC AGT GGC AGC TCT CTG AGC GGC TCT TAC AGG GAT
Mouse ATG AGC TCG TAC TTT GTA AAC TCC TTC TCG GGG CGT TAT CCA AAT GGC CCG GAC TAT CAG TTG CTA AAT TAT GGC AGT GGC AGC TCT CTG AGC GGC TCT TAC AGG GAT
Rat ATG AGC TCG TAC TTT GTA AAC TCC TTC TCG GGG CGT TAT CCA AAT GGC CCG GAC TAT CAG TTG CTA AAT TAT GGC AGT GGC AGC TCT CTG AGC GGC TCT TAC AGG GAT
Kangaroo_rat ATG AGC TCG TAC TTT GTA AAC TCC TTC TCG GGG CGT TAT CCA AAT GGC CCG GAC TAT CAG TTG CTA AAT TAT GGC AGT GGC AGC TCT CTG AGC GGC TCT TAC AGG GAT
Guinea_Pig ATG AGC TCG TAC TTT GTA AAC TCC TTC TCG GGG CGT TAT CCA AAT GGC CCG GAC TAT CAG TTG CTA AAT TAT GGC AGT GGC AGC TCT CTG AGC GGC TCT TAC AGG GAT
Squirrel ATG AGC TCG TAC TTT GTA AAC TCC TTC TCG GGG CGT TAT CCA AAT GGC CCG GAC TAT CAG TTG CTA AAT TAT GGC AGT GGC AGC TCT CTG AGC GGC TCT TAC AGG GAT
Rabbit ATG AGC TCG TAC TTT GTA AAC TCC TTC TCG GGG CGT TAT CCA AAT GGC CCG GAC TAT CAG TTG CTA AAT TAT GGC AGT GGC AGC TCT CTG AGC GGC TCT TAC AGG GAT
Pika ATG AGC TCG TAC TTT GTA AAC TCC TTC TCG GGG CGT TAT CCA AAT GGC CCG GAC TAT CAG TTG CTA AAT TAT GGC AGT GGC AGC TCT CTG AGC GGC TCT TAC AGG GAT
Alpaca ATG AGC TCG TAC TTT GTA AAC TCC TTC TCG GGG CGT TAT CCA AAT GGC CCG GAC TAT CAG TTG CTA AAT TAT GGC AGT GGC AGC TCT CTG AGC GGC TCT TAC AGG GAT
Dolphin ATG AGC TCG TAC TTT GTA AAC TCC TTC TCG GGG CGT TAT CCA AAT GGC CCG GAC TAT CAG TTG CTA AAT TAT GGC AGT GGC AGC TCT CTG AGC GGC TCT TAC AGG GAT
Cow ATG AGC TCG TAC TTT GTA AAC TCC TTC TCG GGG CGT TAT CCA AAT GGC CCG GAC TAT CAG TTG CTA AAT TAT GGC AGT GGC AGC TCT CTG AGC GGC TCT TAC AGG GAT
Horse ATG AGC TCG TAC TTT GTA AAC TCC TTC TCG GGG CGT TAT CCA AAT GGC CCG GAC TAT CAG TTG CTA AAT TAT GGC AGT GGC AGC TCT CTG AGC GGC TCT TAC AGG GAT
Cat ATG AGC TCG TAC TTT GTA AAC TCC TTC TCG GGG CGT TAT CCA AAT GGC CCG GAC TAT CAG TTG CTA AAT TAT GGC AGT GGC AGC TCT CTG AGC GGC TCT TAC AGG GAT
Dog ATG AGC TCG TAC TTT GTA AAC TCC TTC TCG GGG CGT TAT CCA AAT GGC CCG GAC TAT CAG TTG CTA AAT TAT GGC AGT GGC AGC TCT CTG AGC GGC TCT TAC AGG GAT
Megabat ATG AGC TCG TAC TTT GTA AAC TCC TTC TCG GGG CGT TAT CCA AAT GGC CCG GAC TAT CAG TTG CTA AAT TAT GGC AGT GGC AGC TCT CTG AGC GGC TCT TAC AGG GAT
Hedgehog ATG AGC TCG TAC TTT GTA AAC TCC TTC TCG GGG CGT TAT CCA AAT GGC CCG GAC TAT CAG TTG CTA AAT TAT GGC AGT GGC AGC TCT CTG AGC GGC TCT TAC AGG GAT
Shrew ATG AGC TCG TAC TTT GTA AAC TCC TTC TCG GGG CGT TAT CCA AAT GGC CCG GAC TAT CAG TTG CTA AAT TAT GGC AGT GGC AGC TCT CTG AGC GGC TCT TAC AGG GAT
Elephant ATG AGC TCG TAC TTT GTA AAC TCC TTC TCG GGG CGT TAT CCA AAT GGC CCG GAC TAT CAG TTG CTA AAT TAT GGC AGT GGC AGC TCT CTG AGC GGC TCT TAC AGG GAT
Rock_hyrax ATG AGC TCG TAC TTT GTA AAC TCC TTC TCG GGG CGT TAT CCA AAT GGC CCG GAC TAT CAG TTG CTA AAT TAT GGC AGT GGC AGC TCT CTG AGC GGC TCT TAC AGG GAT
Tenrec ATG AGC TCG TAC TTT GTA AAC TCC TTC TCG GGG CGT TAT CCA AAT GGC CCG GAC TAT CAG TTG CTA AAT TAT GGC AGT GGC AGC TCT CTG AGC GGC TCT TAC AGG GAT
Opossum ATG AGC
Wallaby ATG AGC
Platypus ATG AGC
Chicken ATG AGC
Zebra_finch ATG AGC
Lizard ATG
X_tropicalis ATG
B
P value
1
10-5
10-10
10-15
10-20
HoxA
HoxB
HoxC
HoxD
1
2
5
10
20
50
100
Fold-reduction in mammalian synonymous rate (vs. other tetrapods)

evolved specifically within placental mammals.

7.3 High-resolution Bayesian estimates of synonymous rates

As more species become available for comparative genomics analysis, it should be possible to detect synonymous constraint at finer and finer resolutions – well below the 30-, 15-, and 9-codon window lengths used in our original study of 29 mammalian genomes. However, the LRT methodology used above and in our previous work faces two potential difficulties in "scaling down" to accomplish this. First, the underlying statistical test is based on an assumed χ^2 distribution of the likelihood ratio statistic – asymptotically valid for large sample sizes, but potentially unreliable when we wish to analyze just a few sites (Yang and dos Reis, 2010). Second, simulations indicated that the maximum likelihood estimates of λ_s and λ_n are less accurate and more biased with such small sample sizes, making it perilous to attempt direct interpretation of these point estimates.

To avoid these difficulties, we develop a new Bayesian approach for estimating codon evolutionary rates in very small windows and even at individual-site resolution, which does not rely on asymptotic assumptions and uses "weakly informative" prior distributions to regularize parameter estimates. Again, our approach draws inspiration from previous Bayesian approaches and sitewise-resolution methods for analyzing the d_N/d_S ratio (Massingham, 2004; Yang, 2005; Anisimova and Kosiol, 2008; Delport et al., 2008). In our framework, inferences about synonymous rates for a given alignment A are represented by a posterior probability distribution for λ_s, marginalizing over λ_n:

Figure 7-1 *(facing page)*: Lineage-specific synonymous constraint elements (SCEs). **(A)** Alignment of the 5' end of the *HOXB5* ORF across placental mammals and other tetrapod species. Non-colored codons are identical to the human sequence; bright green, synonymous codon substitution; dark green, non-synonymous substitution but conservative of amino acid properties; red, non-conservative substitution. The region is essentially perfectly conserved among placental mammals, but shows numerous substitutions compared to, and within, other tetrapod species. **(B)** Many SCEs appear specific to placental mammals. A likelihood ratio test for mammal-specific reduction in synonymous rate was applied to ~10,000 SCEs defined in our previous study (Lin et al., 2011b). Overall, about 25% of these pass the test with false discovery rate less than 0.1%. Highlighted are SCEs falling within Hox genes, many of which appear highly specific to placental mammals, especially those in the HoxC cluster.

$$\Pr(\lambda_s = x|A) = \frac{\int_0^\infty \Pr(\lambda_s = x, \lambda_n = y) \Pr(A|M, \lambda_s = x, \lambda_n = y)\, dy}{\int_0^\infty \int_0^\infty \Pr(\lambda_s = w, \lambda_n = z) \Pr(A|M, \lambda_s = w, \lambda_n = z)\, dwdz}$$

where $\Pr(\lambda_s = x, \lambda_n = y)$ is a prior distribution for the two parameters and the likelihood $\Pr(A|M, \lambda_s = x, \lambda_n = y)$ is computed by Felsenstein's algorithm, as in the previous maximum likelihood framework. A point estimate of λ_s can be obtained from the maximum *a posteriori* (MAP) value based on this distribution, and we also summarize the evidence for synonymous rate reduction using the posterior odds ratio that $\lambda_s < 1$,

$$K = \frac{\int_0^1 \Pr(\lambda_s = x|A)\, dx}{\int_1^\infty \Pr(\lambda_s = x|A)\, dx}$$

It is convenient to express K as a log-odds score in units of decibans (db); for example, $K = 20$db indicates that we should be willing to lay up to 100:1 odds that $\lambda_s < 1$ rather than $\lambda_s \geq 1$, while $K = 30$db indicates 1,000:1 odds. K may also be less than one (negative in decibans) to indicate that it appears less probable that $\lambda_s < 1$ than $\lambda_s \geq 1$.

Our goal in choosing a prior distribution $\Pr(\lambda_s = x, \lambda_n = y)$ was simply to ensure a proper posterior and provide regularization for the parameter estimates. We used independent Cauchy distributions, centered at one and truncated at zero, as prior distributions for λ_s and λ_n. Centering the distribution at one ensures that maximum a posteriori (MAP) estimates will tend toward this default in the absence of informative data, while the heavy tail of the Cauchy distribution leads to minimal distortion of the likelihood (Gelman, 2006).

In practice, the integrals in these definitions can be approximated by interpolating the joint posterior density of λ_s and λ_n evaluated in a "grid" of closely spaced values of these parameters (Yang, 2005). We continue to use maximum likelihood point estimates for all parameters in the null model M, which is estimated from a relatively copious amount of data. Although it would be straightforward in theory to extend this approach to treat M probabilistically, analyzing just two variables avoids the need for Monte Carlo methods to evaluate the desired posterior quantities.

7.3.1 High-resolution detection of synonymous constraint in mammalian alignments

We applied our new high-resolution Bayesian analysis of codon rates to all CCDS genes in the 33-mammal alignments. For each gene, we estimated a MG94–F3×4 codon model of the complete ORF by maximum likelihood, and used it as the null model M for the analysis of the individual sites in that ORF. The use of the average rates for each ORF

as the null model controls for region- and gene-specific biases in sequence composition, neutral evolutionary rates, and codon usage, and should lead to a somewhat conservative analysis, assuming that purifying selection is more common than positive selection in synonymous sites (Resch et al., 2007).

We used this approach to analyze all $\sim 10^7$ codon sites in CCDS ORFs using sliding windows of 3 codons (for each individual site, we recorded the score of the 3-codon window surrounding that site; for genes with multiple isoforms, we used the isoform with the lengthiest coding sequence). This identified 125,846 sites (1.3%) as being under synonymous constraint with log-odds scores of at least 30db, corresponding to $\geq$1000:1 posterior odds that $\lambda_s < 1$. To ensure the stringency of this cutoff given the multiplicity of sites tested, we estimated a "Bayesian false discovery rate" (Whittemore, 2007; Ji et al., 2008) by combining the log-odds scores with a prior assumption that 1% of all synonymous sites are under selection (slightly conservative compared to the findings of our previous study at 15-codon resolution), which indicates that less than 4% of these 125,846 sites are expected "false discoveries". These sites have a median MAP estimate $\lambda_s = 0.10$, corresponding to a 90% reduced synonymous rate in the surrounding 3-codon window compared to the ORF average.

Figure 7-2 shows the results of the new high-resolution analysis in a region of the tumor suppressor gene *BRCA1*. This was the first mammalian protein-coding region in which unusually low synonymous rates were noted (Hurst and Pal, 2001), but that finding has been called into question on statistical grounds (Schmid and Yang, 2008). Our previous study showed that the apparent synonymous constraint in the *BRCA1* ORF is not only statistically significant, but also can be strikingly localized to a region encoding an alternative splice donor site and an alternate translation initiation site. While we estimated reduced synonymous rates spanning both of these sites at 15-codon resolution, only the region surrounding the alternate translation initiation site was statistically significant under our previous methodology (Figure 3A of Lin et al. (2011b)). Our new analysis, conducted at 3-codon resolution, shows pronounced peaks in the log-odds scores localizing with *both* the splice donor site and the alternative translation initiation site (Figure 7-2), providing evidence that both evolved under purifying selection. This example illustrates how the higher resolution achieved with our new method can support increasingly-specific biological hypotheses.

7.3.2 Detecting miRNA target sites embedded in human ORFs

Animal miRNAs recognize their target transcripts via highly sequence-specific sites of just $\sim$8nt (although surrounding positions also influence targeting efficiency (Bartel, 2009)).

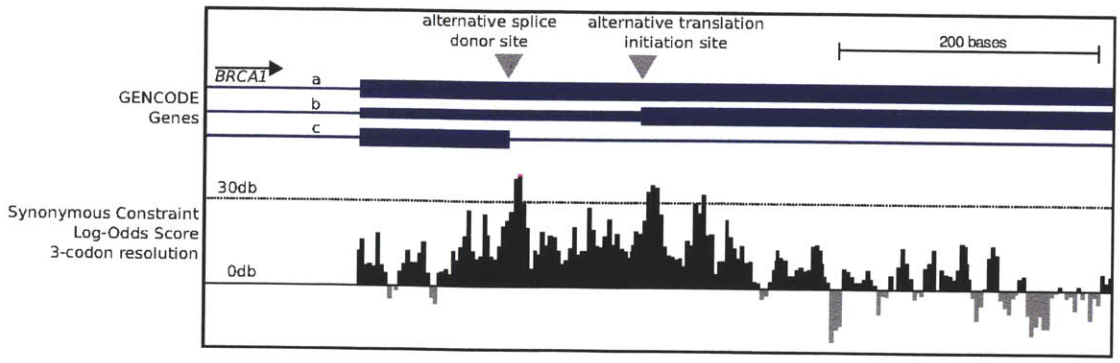


Figure 7-2: High-resolution detection of synonymous constraint in *BRCA1*. According to gene annotations, a coding exon of *BRCA1* (isoform a) additionally encodes an alternative splice donor site (b) and an alternative translation initiation site (c) within a ~100bp region. Analysis of synonymous rates at 3-codon resolution in 33 mammals reveals pronounced peaks of constraint coinciding with the splice site and the translation initiation site, indicating that both have evolved under purifying selection.

Several previous studies have observed trends toward preferential conservation of these miRNA target 8mer motifs within mammalian coding regions by averaging across potential targets genome-wide (Hurst, 2006; Kural et al., 2009; Lewis et al., 2005), suggesting miRNA targeting within ORFs. Some progress has also been made toward the more difficult problem of systematically locating individual miRNA target sites within coding regions (Forman et al., 2008; Forman and Coller, 2010; Schnall-Levin et al., 2011, 2010; Brest et al., 2011), so that the specific genes targeted by each miRNA can be reliably predicted. The 9-codon resolution considered in our previous study was still not fully satisfactory for this purpose, but our new high-resolution synonymous constraint analysis should enable further progress. As a case study, we analyzed potential target sites of *let-7*, a deeply conserved miRNA which has previously been shown to target human coding genes through their ORFs (Forman et al., 2008).

First, we tested our method on simulated alignments to estimate its theoretical power to locate ORF-embedded instances of the *let-7* target motif `CTACCTCA`. Under the assumptions of our codon models, these simulations indicate that the 30db cutoff and should permit detection of about half of embedded *let-7* target sites that have evolved under selection in mammals to such a degree that they tolerated synonymous substitutions at an ~80% reduced rate (Table 7.1). We can also expect to detect a large majority of sites with even lower synonymous rates; and, given the high sequence specificity of the "seed" region for miRNA targeting efficiency (Bartel, 2009), it is not unreasonable to expect sites truly under selection to show such very low divergence.

Next, we examined occurrences of the *let-7* target motif in human CCDS genes with

	Simulated λ_s				
species	1.00	0.40	0.20	0.10	0.05
33 mammals	$< 0.1\%$	21%	70%	91%	96%
18 mammals	$< 0.1\%$	8%	37%	69%	89%
11 mammals	$< 0.1\%$	$< 0.1\%$	$< 0.1\%$	$< 0.1\%$	$< 0.1\%$

Table 7.1: Estimated power to detect *let-7* miRNA target sites embedded in coding regions. Shown is the proportion of artificial *let-7* target 8mer sites receiving synonymous constraint log-odds scores of at least 30db (1000:1 posterior odds), in simulated alignments of varying numbers of species and at varying degrees of constraint on the target site. For example, 70% of simulated *let-7* target sites, aligned across 33 mammals with the synonymous rate reduced by 80% ($\lambda_s = 0.20$), are detected at this threshold. Alignments of 3- and 4-codon sites encoding the *let-7* target 8mer CTACCTCA were generated based on an MG94–F3×4 codon model estimated from autosomal CCDS sites, with simulated reduced synonymous rates. Reading frames and positions flanking the 8mer were sampled according to the frequency of the resulting codons implied in the model. Since the real alignments have widely-varying species coverage across individual sites, we performed the simulations using subsets of 11 and 18 as well as the full 33 mammals (the specific subsets of species correspond to those available in previous versions of the human whole-genome alignments from the UCSC Genome Browser). CCDS sites show coverage of median 24 species in the real alignments.

our previously-computed 33-mammals, 3-codon synonymous constraint scores. The 8mer appears a total of 948 times, of which 107 (11%) fall within one of the 3-codon windows assigned a synonymous constraint score 30db or above. Less than 2% of these are expected false discoveries under the 1% prior (a conservative assumption, given that they match the known miRNA seed). The 107 candidate target sites appear within 100 different genes. The six genes containing two or more sites include four genes encoding known or predicted DNA-binding transcription factors (*GLI3*, *RAD54L2*, *SALL1*, and *SALL4*), the predicted histone methyltransferase gene *MLL4*, and *DICER1*, which encodes a crucial factor in the miRNA biogenesis pathway and in which *let-7* targeting through the ORF was previously investigated (Forman et al., 2008).

In summary, a straightforward downstream analysis of our synonymous constraint scores identified 100 genes that *let-7* potentially regulates through target sites embedded in their ORFs, with a previously known example among the top candidates. These predictions each come with a high degree of statistical confidence, and furthermore, our simulation benchmarks provide an argument that they should constitute the majority of *let-7* targets that match the seed motif CTACCTCA and have evolved under strong selection within human CCDS ORFs. Previous efforts to locate embedded miRNA target sites used permutation arguments to estimate their specificity or FDR, but to our knowledge, did not propose model-based estimates of their sensitivity. As *let-7* illustrates, our data can be

used to investigate potential embedded target sites of other miRNAs and, more generally, other short regulatory motif instances.

7.3.3 Towards single-codon resolution of synonymous constraint

While the 3-codon resolution used above should be useful for detecting numerous types of short regulatory motif instances, we sought to increase the resolution even further, to the level of individual codon sites – similar to sitewise methods for detecting nucleotide-level constraint (Davydov et al., 2010; Garber et al., 2009; Pollard et al., 2009) and for estimating the d_N/d_S ratio (Massingham, 2004; Yang and Nielsen, 2002). Using the 33-mammal alignments, vanishingly few individual CCDS sites reach the 30db cutoff, and lower cutoffs led to high estimated FDRs. Additional mammalian species evidently will be needed to achieve sitewise resolution within this clade.

Until then, we can increase our power by extending the analysis outside of mammals to include the seven other tetrapods used in our previous lineage-specific analysis – with the obvious restriction that this will only allow detection of sitewise constraint conserved across these much larger evolutionary distances. Indeed, the 40-tetrapod alignments permit the identification of 25,900 CCDS sites (0.26%) at the 30db cutoff, again with estimated FDR below 4% (under the 1% prior). The sites passing this threshold typically show no synonymous substitutions in the aligned species (median MAP $\lambda_s = 0$). Sixfold-degenerate codons are strongly over-represented in these high-ranking sites, as expected since they can provide more evidence for a reduced synonymous rate; conversely, twofold-degenerate and non-degenerate codons are strongly under-represented.

These results indicate that genome-wide synonymous constraint can be resolved at tens of thousands of individual codon sites using the current 40-species alignments, as illustrated by an alternative splice donor site embedded within an exon of *ADAR* (Figure 7-3). The small proportion of sites identified, relative to our previous estimates, may reflect limited discovery power at sitewise resolution given the current alignments, as well as the more stringent requirement of conservation across much larger evolutionary distances.

7.4 Hidden Markov model for defining SCEs based on Bayesian rate estimates

The new method presented in the previous section is designed to provide very high-resolution measurements of evolutionary rates – even down to single codon sites, given enough aligned species. However, there appeared to be limited power at true sitewise

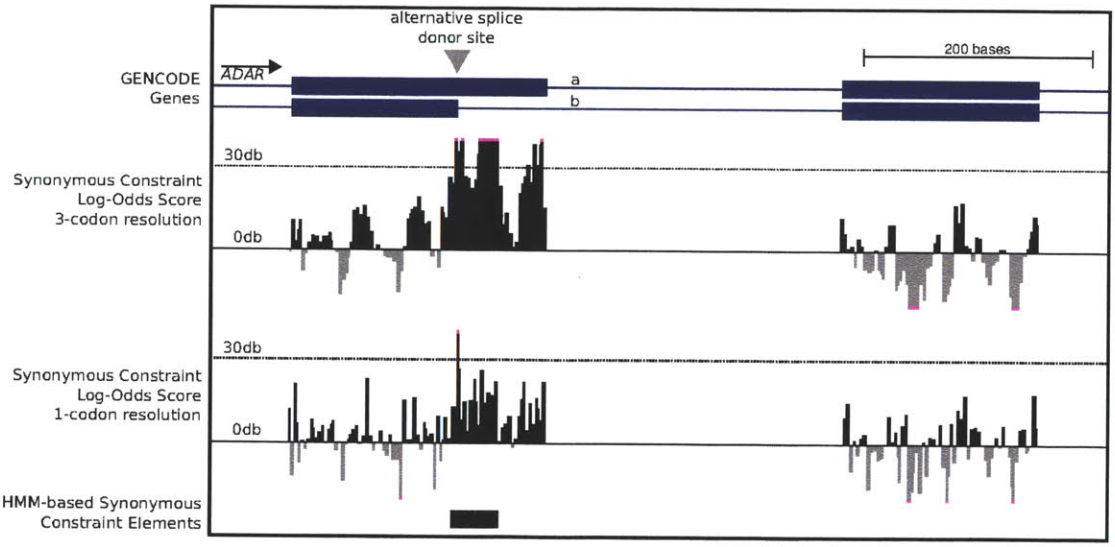


Figure 7-3: Sitewise resolution of synonymous constraint. The RNA editing gene *ADAR* includes an exon with two splice forms, with the coding sequence of the longer isoform (a) additionally encoding the splice donor site used in the shorter isoform (b). Shown are synonymous constraint scores at 3-codon resolution based on 33 mammals, and at single-codon resolution based on 40 tetrapods. The splice donor site can be localized precisely at single-codon resolution, while at 3-codon resolution there is evidence for synonymous constraint in the downstream region as well. Also shown is a synonymous constraint element spanning the splice site and the downstream region, predicted by an HMM operating on the single-codon log-odds scores.

resolution given the current alignments, and furthermore, many downstream analyses still call for a set of lengthier, contiguous "elements" throughout the genome, like those reported in our previous study. We therefore designed a hidden Markov model (HMM) to define synonymous constraint elements by generating a segmentation of the new high-resolution rate estimates.

The HMM has two states: one corresponding to codon sites under synonymous constraint, and the other corresponding to all other coding regions. The HMM emissions are the previously defined posterior log-odds scores for synonymous rate reduction. In the "constraint" state, the log-odds emission distribution is Gaussian. In the "other" state, the log-odds emissions are modeled as a mixture of a point mass corresponding to all non-positive scores, and an exponential distribution for positive scores. In notation,

$$\begin{aligned} \log K \mid \text{Constraint} &\sim \mathcal{N}(\mu_1, \sigma^2) \\ \max(0, \log K) \mid \text{Other} &\sim p_0\delta(0) + (1 - p_0)\text{Exponential}(\mu_2) \end{aligned}$$

Flooring the log-odds scores at zero for the "other" state avoids expending model design and estimation effort on the detailed distribution of negative scores, which is generally not of interest for detecting synonymous constraint. Using the log-odds scores computed across the genome, we can obtain maximum likelihood estimates of μ_1, σ^2, p_0, μ_2, and the HMM transition parameters in an unsupervised fashion from the Baum-Welch algorithm (expectation-maximization). We then use the Viterbi algorithm to produce a decoding of SCEs throughout the genome. Our approach here is simpler than a true "phylo-HMM", since the generative emissions are the log-odds scores rather than alignment columns, but the log-odds scores should provide an excellent summary of the alignment for the purpose of defining SCEs.

7.4.1 A new set of SCEs based on high-resolution estimates

We trained and applied our HMM to delineate SCEs based on the sitewise log-odds scores for the 40-tetrapod alignments. This should increase detection power by effectively combining evidence across multiple adjacent sites, in a way that complements our previous sliding-window approach (Lin et al., 2011b). In particular, while both methods involve different assumptions about the length of SCEs to be discovered – sliding windows requiring a choice of window size, and the HMM assuming a geometric distribution – the HMM's can be learned from data in a more principled way. On the other hand, the HMM is less transparent than the sliding-window approach.

The HMM expectation-maximization procedure, applied to all CCDS ORFs, resulted in a model in which SCEs average 6 codon sites in length, with sitewise synonymous constraint log-odds scores averaging 14.8 ± 8.0 db. The model also indicates that an average of 337 sites separate consecutive SCEs, in which the log-odds scores are non-positive 59.6% of the time, and the remaining exponentially-distributed positive scores average 8.2 db. Lastly, the model implies that 1.8% of codon sites arise from the SCE state.

The Viterbi decoding of CCDS ORFs based on this model predicted 11,010 SCEs, averaging 12 codons in length, with 50% of the elements nine codons or shorter. These predicted SCEs cover 131,687 of the $\sim 10^7$ codon sites (1.3%), which have an average synonymous constraint log-odds score of 15.3 ± 8.4 db and median MAP estimate $\lambda_s = 0.23$. Compared to the aforementioned theoretical model, these elements are relatively long and cover a smaller fraction of sites, reflecting limited current power to detect extremely short SCEs. Figure 7-3 illustrates the results for the *ADAR* example, in which an HMM-based SCE covers a small region surrounding the alternative splice site, even though only one of the sitewise scores rises above our 30db cutoff.

The HMM-based SCEs fall within 5127 of the 17,850 CCDS ORFs we analyzed (29%). These ORFs do not show increased codon usage bias, with a median Effective Number of Codons (Wright, 1990; Fuglsang, 2005) of 51.9, higher than the remaining genes (median 48.9). This confirms that the low synonymous divergence in these SCEs is not explained by ORF-wide codon usage biases arising from regional compositional effects or selection for translational efficiency – as expected, since we adjusted the "null model" used to evaluate the sites in each ORF in order to reflect any such biases.

Slightly less than one-half (46%) of the new HMM-based SCEs overlap those identified in our previous study, which was based on placental mammals only and a sliding-window methodology with a relatively large window size of 15 codons. Conversely, 40% of the original SCEs overlap the new HMM-based elements. This is clearly a strong enrichment since both sets cover just a few percent of CCDS sites, but the differences in the study designs – in the resolution, species used, and analytical methodology – evidently permit detection of substantially different sets of SCEs.

7.5 SCEs show reduced synonymous variation in human populations

Lastly, we used publicly-available population variation data to study the recent evolution of the new HMM-based SCEs in humans. First, we analyzed the 326 synonymous

single-nucleotide polymorphisms (SNPs) that lie within SCEs and were genotyped in the International HapMap Project (release 27) (Altshuler et al., 2010). Of these, 75% appeared monomorphic in the Yoruba population (all individuals genotyped were homozygous for the same allele). The same is true of only 57% of the 18,557 synonymous HapMap SNPs falling within the CCDS ORFs containing SCEs, but outside of the SCEs themselves – a highly significant difference (hypergeometric $P < 10^{-11}$). This trend also holds in other HapMap populations.

Second, we analyzed single-nucleotide variant calls based on whole-genome shotgun sequencing in the pilot phase of the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010). The SCEs contain 186 distinct synonymous variant calls in the 59 sequenced individuals of the Yoruba population, which is an average of one every 2124 nucleotides – a marked depletion compared to one every 857 nucleotides in the CCDS ORFs containing SCEs, but outside of the SCEs themselves (hypergeometric $P < 10^{-16}$). Furthermore, the 186 synonymous variants in SCEs show a reduced mean derived-allele frequency of 14%, compared to 19% outside SCEs (Mann-Whitney $P < 0.01$). Again, these trends also hold in the other populations sequenced in the pilot project.

The reduced synonymous variation found specifically within SCEs suggests that many have continued to evolve under selection in human populations, and thus may have significance to human physiology and disease (Cartegni et al., 2002; Fairbrother et al., 2004; Carlini and Genut, 2005; Chamary et al., 2006; Sauna and Kimchi-Sarfaty, 2011).

7.6 Data availability

We provide the genome-wide synonymous constraint scores and the HMM-based SCEs in formats suitable for display in genome browsers, as well as downstream analysis, at: `http://compbio.mit.edu/SCE2.0`

7.7 Discussion

In this chapter, we extended our previous methodology for detecting mammalian SCEs to greatly increase its resolution and address additional biological questions. We first developed a new likelihood ratio test for lineage-specific selection on synonymous sites. We found that a quarter of our previously defined SCEs appear to represent evolutionary innovations in placental mammals, including several in Hox genes. This is a significant step towards elucidating how SCEs may have contributed to eutherian evolution (Dong et al., 2009; Lin et al., 2008b; Fredman et al., 2011). We also developed a new analytical

methodology for measuring synonymous constraint at much higher resolution, which we applied to generate genome-wide measurements of synonymous constraint and to define a new set of 11,010 SCEs based on sitewise rate estimates.

As illustrated by our case studies of splicing and translation regulation in *BRCA1*, miRNA targeting by *let-7*, and sitewise resolution in *ADAR*, the data and annotations produced by this work can support many future insights into the function, regulation and evolution of mammalian gene structures. Furthermore, our finding that the synonymous sites in SCEs remain under selection in human population indicates that our results can also inform disease association studies and personal genome interpretation, by identifying synonymous variants that, though currently ignored in many such analyses, could have significant physiological consequences (Cartegni et al., 2002; Fairbrother et al., 2004; Carlini and Genut, 2005; Chamary et al., 2006; Sauna and Kimchi-Sarfaty, 2011). Our framework for detecting synonymous constraint elements should continue to grow more powerful as additional species become available for comparative genomics analysis.

Chapter 8

Conclusion and future directions

8.1 Contributions

This thesis has developed novel methodologies for comparative genomics analysis of protein-coding genes based on phylogenetic codon models. By applying these methods in the human genome and those of other scientifically important species, we improved their gene annotations, revealed novel biological insights, and produced genome-wide datasets that can inform many future studies on mammalian gene structures, human disease associations, and personal genome interpretation.

First, we developed PhyloCSF, an algorithm to distinguish multi-species sequence alignments of protein-coding and non-coding regions, based on a formal statistical comparison of empirical codon models. We showed that PhyloCSF outperforms previous methods, and we have provided a software implementation for use by the community. We applied PhyloCSF to improve the gene annotations of several species, including human, and to reveal unusual gene structures – including a surprisingly widespread mechanism of stop codon readthrough in fruitfly, with additional examples found in mammals.

Second, we proposed methods to detect SCEs and to study their evolutionary history. We used these methods to produce the first high-resolution, genome-wide annotations of SCEs in the human genome, locating more than 10,000 elements within over one-quarter of all human genes. We showed evidence that they encode numerous overlapping biological functions – though many remain yet to be elucidated – and illustrated several examples of how they can support a variety of specific biological insights. We also found that a substantial fraction appear to be evolutionary innovations in mammals, and that they have continued to evolve under selection in human populations – indicating that they could have relevance to human physiology and disease.

8.2 Convergence of models of molecular evolution and methods for phylogenetically-informed genome annotation

The comparative genomics methods we developed in this thesis are innovative in the setting of whole-genome annotation, whether improving existing gene catalogs or detecting new classes of overlapping functional elements. Notably however, they utilize a well-established framework for phylogenetic models of molecular evolution, which has developed over decades – long predating the availability of complete genome sequences. Many of the basic concepts underlying phylogenetic codon models, such as the d_N/d_S ratio and the Markov process model, were in fact foreseen with hardly any nucleotide sequence data at all, and long used for the study of individual genes without reference to their genomic context (Anisimova and Kosiol, 2008; Delport et al., 2008). Naturally however, it was more difficult to anticipate the challenges and opportunities of using these methods with genome-wide data until the past decade, which correspondingly has seen the development of richer model parameterizations and techniques to estimate them from vast amounts of data (Holmes and Rubin, 2002; Siepel and Haussler, 2004; Hobolth and Jensen, 2005; Klosterman et al., 2006; Kiryu, 2011).

Our methods for phylogenetically informed annotation of protein-coding genes can be seen as part of a larger effort to adapt relatively mature phylogenetic models for practical application in genome-wide settings. For example, our previous CSF metric (Lin et al., 2007, 2008a) took advantage of substitution frequencies empirically estimated from genome-wide data, and proved useful in many annotation applications, but its *ad hoc* nature was a salient deficiency. The development of fully-parameterized empirical codon models by Kosiol et al. (2007), who presented them as a more-accurate model of molecular evolution, allowed us to reformulate CSF in a completely rigorous framework, while preserving its essential ideas and applications. Similarly, our use of phylogenetic codon models to discover SCEs and study their evolutionary history, drawing inspiration directly from similar models used to study molecular evolution, is a significant advance over previous, less-rigorous methods developed by ourselves and others (Schattner, 2006; Parmley and Hurst, 2007b; Lin et al., 2008b; Suzuki and Saitou, 2011). We expect that applied methods for phylogenetically-informed genome annotation will continue to converge with ongoing research on models of molecular evolution.

8.3 Comparative genomics of thousands of species

We designed the methodologies in this thesis to operate on dozens of genomes – twelve flies and up to forty tetrapods. Our algorithms can accommodate more, with linear runtime in the number of species, and will grow more powerful as they do. But the number of closely-related genome sequences available for comparative analysis is shortly set to not just increase, but explode, with the Genome 10K project to sequence the genomes of 10,000 vertebrates, and i5k planning to sequence 5000 arthropods. Comparative genomics on this scale will present many new challenges and may demand fundamentally different design decisions at each step of the process, including genome assembly and alignment as well as subsequent phylogenetic analyses such as ours. Some inspiration will undoubtedly be drawn from developments in microbial genomics and metagenomics, which are already dealing with similar numbers of species (Kuczynski et al., 2011; Ribeca and Valiente, 2011) – albeit with far smaller genomes, containing less-complex gene structures and regulatory mechanisms.

A new generation of our methods, rising to these challenges, should be able to achieve dramatically greater power and resolution, and the ability to ask much more detailed evolutionary questions. Ideas from existing sparsely-parameterized codon models that analyze lineage-specific variation in the d_N/d_S ratio at sitewise resolution (Yang and Nielsen, 2002) will be generalizable to richly parameterized methods such as PhyloCSF, leading to ever more power to trace selective pressures on the encoded protein sequences. Similarly, lineage-specific synonymous rates will be measurable at sitewise resolution, further revealing evolutionary innovation of overlapping functions.

8.4 Population and personal genomics

In addition to the sequencing of many additional species, thousands – and perhaps millions – of individual people will have their genomes sequenced in coming years, with the goals of better understanding genetic variation and population structure, identifying causative variants in diseases with a genomic basis, and ultimately personalizing the treatment of individual patients to match their complete genetic profile (The 1000 Genomes Project Consortium, 2010; Cooper and Shendure, 2011; Chan and Ginsburg, 2011; Li, 2011).

Our high-resolution annotation of likely-functional synonymous sites provides a specific opportunity to follow up on a class of possible causative variants that has previously been refractory to systematic discovery, due to the lack of obvious functional consequences of synonymous variants. More generally, our results contribute to a complete and accurate catalog of functional elements in the human genome, which will be a key requirement to

properly interpret these genomes.

Beyond these immediate downstream uses of our current results, population and personal genomics will enable the development of rigorous but practical genome-wide methods for detecting evolutionary signatures in contemporary human populations (Stearns et al., 2010) – including in protein-coding sequences and their synonymous sites. As with phylogenetic models for cross-species analysis, many of the essential concepts that will underlie such methods have already been established over decades of research in population genetics, but the actual availability of the whole-genome data will present many unforeseen challenges and opportunities.

8.5 Evolutionary signatures: a central tool for understanding genomes

At the most general level, this thesis has illustrated comparative genomics and the detection of evolutionary signatures as a powerful approach to annotate functional sequence elements within genomes. We have focused on protein-coding genes, but evolutionary signatures can also be found for many other classes of functional elements (Stark et al., 2007). While powerful in their own right, they are just one piece of an integrative effort to annotate the human and other genomes, which also depends on many other computational approaches, and increasingly on high-throughput experimental datasets (The ENCODE Project Consortium, 2011; The modENCODE Consortium et al., 2010).

In this effort, evolutionary signatures complement modern high-throughput experiments in important ways. Because of the stochastic nature of biochemical activities such as transcription, regulatory factor binding and chromatin modification, experimental measurements of these genomic phenomena are inherently noisy – likely to capture many events that are biochemically reproducible but of little functional consequence. Evolutionary signatures provide a powerful signal to distinguish the sequence elements that actually perform important biological functions, to the extent that disruptions to them tend to be excluded by natural selection.

Indeed, beyond annotation, some of our most satisfying findings (from the author's perspective) were those that influenced or directly motivated hypothesis-driven followup work by experimental biologists, including PhyloCSF's role in recognizing novel non-coding RNA genes, and our investigation of stop codon readthrough and translational frameshifting in flies. As our SCEs also provide a wealth of future opportunities for hypothesis generation and experimental investigation, we hope this is only the beginning.

Bibliography

Ahmed, Z. M., Masmoudi, S., Kalay, E., Belyantseva, I. A., Mosrati, M. A., Collin, R. W. J., Riazuddin, S., Hmani-Aifa, M., Venselaar, H., Kawar, M. N., *et al.*, 2008. Mutations of LRTOMT, a fusion gene with alternative reading frames, cause nonsyndromic deafness in humans. *Nature Genetics*, **40**(11):1335–1340.

Alioto, T. and Guig, R., 2009. State of the art in eukaryotic gene prediction. In *Modern genome annotation : the BioSapiens Network*, pages 7–40. Springer, New York.

Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F., Peltonen, L., *et al.*, 2010. Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**(7311):52–58.

Anisimova, M., Bielawski, J. P., and Yang, Z., 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Molecular Biology and Evolution*, **18**(8):1585–1592. PMID: 11470850.

Anisimova, M. and Kosiol, C., 2008. Investigating Protein-Coding sequence evolution with probabilistic codon substitution models. *Molecular Biology and Evolution*, **26**(2):255–271.

Aruscavage, P. J. and Bass, B. L., 2000. A phylogenetic analysis reveals an unusual sequence conservation within introns involved in RNA editing. *RNA (New York, N.Y.)*, **6**(2):257–269. PMID: 10688364.

Arvestad, L. and Bruno, W. J., 1997. Estimation of reversible substitution matrices from multiple pairs of sequences. *Journal of Molecular Evolution*, **45**(6):696–703.

Baek, D., 2005. Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. *Proceedings of the National Academy of Sciences*, **102**(36):12813–12818.

Baranov, P. V., Wills, N. M., Barriscale, K. A., Firth, A. E., Jud, M. C., Letsou, A., Manning, G., and Atkins, J. F., 2011. Programmed ribosomal frameshifting in the expression of the regulator of intestinal stem cell proliferation, adenomatous polyposis coli (APC). *RNA Biology*, **8**(4):637–647.

Bartel, D. P., 2009. MicroRNAs: target recognition and regulatory functions. *Cell*, **136**:215–233.

Bass, B. L., 2002. RNA EDITING BY ADENOSINE DEAMINASES THAT ACT ON RNA. *Annual Review of Biochemistry*, **71**(1):817–846.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., and Haussler, D., 2004. Ultraconserved elements in the human genome. *Science (New York, N.Y.)*, **304**(5675):1321–1325. PMID: 15131266.

Benjamini, Y. and Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, **57**(1):289–300.

Benjamini, Y. and Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, **29**(4):1165–1188.

Blanchette, M., 2007. Computation and analysis of genomic Multi-Sequence alignments. *Annual Review of Genomics and Human Genetics*, **8**(1):193–213.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D., *et al.*, 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research*, **14**(4):708 –715.

Brent, M. R., 2008. Steady progress and recent breakthroughs in the accuracy of automated genome annotation. *Nat Rev Genet*, **9**(1):62–73.

Brest, P., Lapaquette, P., Souidi, M., Lebrigand, K., Cesaro, A., Vouret-Craviari, V., Mari, B., Barbry, P., Mosnier, J., Hbuterne, X., *et al.*, 2011. A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in crohn's disease. *Nature Genetics*, **43**(3):242–245.

Butler, G., Rasmussen, M. D., Lin, M. F., Santos, M. A. S., Sakthikumar, S., Munro, C. A., Rheinbay, E., Grabherr, M., Forche, A., Reedy, J. L., *et al.*, 2009. Evolution of pathogenicity and sexual reproduction in eight candida genomes. *Nature*, **459**(7247):657–662.

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J. L., 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development*, **25**(18):1915–1927.

Carlini, D. B. and Genut, J. E., 2005. Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. *Journal of Molecular Evolution*, **62**(1):89–98.

Cartegni, L., Chew, S. L., and Krainer, A. R., 2002. LISTENING TO SILENCE AND UNDERSTANDING NONSENSE: EXONIC MUTATIONS THAT AFFECT SPLICING. *Nature Reviews Genetics*, **3**(4):285–298.

Castillo-Davis, C. I., Mekhedov, S. L., Hartl, D. L., Koonin, E. V., and Kondrashov, F. A., 2002. Selection for short introns in highly expressed genes. *Nature Genetics*, .

Castresana, J., 2002. Genes on human chromosome 19 show extreme divergence from the mouse orthologs and a high GC content. *Nucleic Acids Research*, **30**(8):1751–1756.

Chamary, J. V., Parmley, J. L., and Hurst, L. D., 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature Reviews Genetics*, **7**:98–108.

Chan, I. S. and Ginsburg, G. S., 2011. Personalized medicine: Progress and promise. *Annual Review of Genomics and Human Genetics*, **12**(1):217–244.

Chen, F., 2005. Alternatively and constitutively spliced exons are subject to different evolutionary forces. *Molecular Biology and Evolution*, **23**(3):675–682.

Chen, H. and Blanchette, M., 2007. Detecting non-coding selective pressure in coding regions. *BMC Evolutionary Biology*, **7**(Suppl 1):S9.

Chen, M. and Manley, J. L., 2009. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nature Reviews. Molecular Cell Biology*, **10**(11):741–754. PMID: 19773805.

Chung, W., Wadhawan, S., Szklarczyk, R., Pond, S. K., and Nekrutenko, A., 2007. A first look at ARFome: Dual-Coding genes in mammalian genomes. *PLoS Computational Biology*, **3**(5):e91.

Clamp, M., Fry, B., Kamal, M., Xie, X., Cuff, J., Lin, M. F., Kellis, M., Lindblad-Toh, K., and Lander, E. S., 2007. Distinguishing protein-coding and noncoding genes in the human genome. *Proceedings of the National Academy of Sciences*, **104**(49):19428 –19433.

Cooper, G. M., 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research*, **15**(7):901–913.

Cooper, G. M. and Shendure, J., 2011. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Reviews Genetics*, **12**(9):628–640.

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., and Batzoglou, S., 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Computational Biology*, **6**:e1001025.

DeCaprio, D., Vinson, J. P., Pearson, M. D., Montgomery, P., Doherty, M., and Galagan, J. E., 2007. Conrad: Gene prediction using conditional random fields. *Genome Research*, **17**(9):1389–1398.

Delcher, A. L., Harmon, D., Kasif, S., White, O., and Salzberg, S. L., 1999. Improved microbial gene identification with GLIMMER. *Nucleic Acids Research*, **27**(23):4636 –4641.

Delport, W., Scheffler, K., and Seoighe, C., 2008. Models of coding sequence evolution. *Briefings in Bioinformatics*, **10**(1):97–109.

Dong, X., Navratilova, P., Fredman, D., Drivenes, O., Becker, T. S., and Lenhard, B., 2009. Exonic remnants of whole-genome duplication reveal cis-regulatory function of coding exons. *Nucleic Acids Research*, **38**:1071–1085.

Down, T., Leong, B., and Hubbard, T. J. P., 2006. A machine learning strategy to identify candidate binding sites in human protein-coding sequence. *BMC Bioinformatics*, **7**:419. PMID: 17002805.

Drosophila 12 Genomes Consortium, 2007. Evolution of genes and genomes on the drosophila phylogeny. *Nature*, **450**(7167):203–218.

Fairbrother, W. G., Holste, D., Burge, C. B., and Sharp, P. A., 2004. Single nucleotide PolymorphismBased validation of exonic splicing enhancers. *PLoS Biology*, **2**(9):e268.

Fairbrother, W. G., Yeh, R., Sharp, P. A., and Burge, C. B., 2002. Predictive identification of exonic splicing enhancers in human genes. *Science (New York, N.Y.)*, **297**(5583):1007–1013. PMID: 12114529.

Felsenstein, J., 2004. *Inferring phylogenies*. Sinauer Associates, Sunderland Mass.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., *et al.*, 2010. Ensembl 2011. *Nucleic Acids Research*, **39**(Database):D800–D806.

Forman, J. J. and Coller, H. A., 2010. The code within the code: microRNAs target coding regions. *Cell Cycle*, **9**:1533–1541.

Forman, J. J., Legesse-Miller, A., and Coller, H. A., 2008. A search for conserved sequences in coding regions reveals that the let-7 microRNA targets dicer within its coding sequence. *Proceedings of the National Academy of Sciences*, **105**:14879–14884.

Fox, A. K., Tuch, B. B., and Chuang, J. H., 2008. Measuring the prevalence of regional mutation rates: an analysis of silent substitutions in mammals, fungi, and insects. *BMC Evolutionary Biology*, **8**(1):186.

Fredman, D., Dong, X., and Lenhard, B., 2011. Making enhancers from spare parts of the genome. *Genome Biology*, **12**(12):138.

Fuglsang, A., 2005. Estimating the "Effective number of codons": The wright way of determining codon homozygosity leads to superior estimates. *Genetics*, **172**(2):1301–1307.

Fujita, P. A., Rhead, B., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Cline, M. S., Goldman, M., Barber, G. P., Clawson, H., Coelho, A., *et al.*, 2010. The UCSC genome browser database: update 2011. *Nucleic Acids Research*, **39**(Database):D876–D882.

Galtier, N. and Duret, L., 2007. Adaptation or biased gene conversion? extending the null hypothesis of molecular evolution. *Trends in Genetics*, **23**(6):273–277.

Gao, F. and Zhang, C., 2004. Comparison of various algorithms for recognizing short coding sequences of human genes. *Bioinformatics*, **20**(5):673 –681.

Garber, M., Guttman, M., Clamp, M., Zody, M. C., Friedman, N., and Xie, X., 2009. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics*, **25**:i54–i62.

Gelman, A., 2006. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1**:119.

Goldman, N. and Yang, Z., 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution*, **11**(5):725–736. PMID: 7968486.

Goren, A., Ram, O., Amit, M., Keren, H., Lev-Maor, G., Vig, I., Pupko, T., and Ast, G., 2006. Comparative analysis identifies exonic splicing regulatory SequencesThe complex definition of enhancers and silencers. *Molecular Cell*, **22**(6):769–781.

Green, E. D., Guyer, M. S., Green, E. D., Guyer, M. S., Manolio, T. A., and Peterson, J. L., 2011. Charting a course for genomic medicine from base pairs to bedside. *Nature*, **470**(7333):204–213.

Grimwood, J., Gordon, L. A., Olsen, A., Terry, A., Schmutz, J., Lamerdin, J., Hellsten, U., Goodstein, D., Couronne, O., Tran-Gyamfi, M., *et al.*, 2004. The DNA sequence and biology of human chromosome 19. *Nature*, **428**(6982):529–535.

Gross, S. S., Do, C. B., Sirota, M., and Batzoglou, S., 2007. CONTRAST: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction. *Genome Biology*, **8**(12):R269.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., Huarte, M., Zuk, O., Carey, B. W., Cassady, J. P., *et al.*, 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, **458**(7235):223–227.

Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M. J., Gnirke, A., Nusbaum, C., *et al.*, 2010. Ab initio reconstruction of cell typespecific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology*, **28**(5):503–510.

Hameed, M., Orrell, R. W., Cobbold, M., Goldspink, G., and Harridge, S. D. R., 2003. Expression of IGF-I splice variants in young and old human skeletal muscle after high resistance exercise. *The Journal of Physiology*, **547**(Pt 1):247–254. PMID: 12562960.

Han, J., Pedersen, J. S., Kwon, S. C., Belair, C. D., Kim, Y., Yeom, K., Yang, W., Haussler, D., Blelloch, R., and Kim, V. N., *et al.*, 2009. Posttranscriptional crossregulation between drosha and DGCR8. *Cell*, **136**(1):75–84.

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C., Chrast, J., Lagarde, J., Gilbert, J. G. R., Storey, R., Swarbreck, D., *et al.*, 2006. GENCODE: producing a reference annotation for ENCODE. *Genome Biology*, **7 Suppl 1**:S4.1–9. PMID: 16925838.

Hastings, M. L., 2000. Post-transcriptional regulation of thyroid hormone receptor expression by cis-Acting sequences and a naturally occurring antisense RNA. *Journal of Biological Chemistry*, **275**(15):11507–11513.

Hobolth, A. and Jensen, J. L., 2005. Statistical inference in evolutionary models of DNA sequences via the EM algorithm. *Statistical Applications in Genetics and Molecular Biology*, **4**:Article18. PMID: 16646835.

Holmes, I. and Rubin, G., 2002. An expectation maximization algorithm for training hidden substitution models. *Journal of Molecular Biology*, **317**(5):753–764.

Hoopengardner, B., Bhalla, T., Staber, C., and Reenan, R., 2003. Nervous system targets of RNA editing identified by comparative genomics. *Science (New York, N.Y.)*, **301**(5634):832–836. PMID: 12907802.

Hsu, F., Kent, W. J., Clawson, H., Kuhn, R. M., Diekhans, M., and Haussler, D., 2006. The UCSC known genes. *Bioinformatics*, **22**(9):1036–1046.

Hughes, J. F., Skaletsky, H., Pyntikova, T., Graves, T. A., van Daalen, S. K. M., Minx, P. J., Fulton, R. S., McGrath, S. D., Locke, D. P., Friedman, C., *et al.*, 2010. Chimpanzee and human y chromosomes are remarkably divergent in structure and gene content. *Nature*, **463**(7280):536–539.

Hughes, J. F., Skaletsky, H., Pyntikova, T., Minx, P. J., Graves, T., Rozen, S., Wilson, R. K., and Page, D. C., 2005. Conservation of y-linked genes during human evolution revealed by comparative sequencing in chimpanzee. *Nature*, **437**(7055):100–103.

Hung, T., Wang, Y., Lin, M. F., Koegel, A. K., Kotake, Y., Grant, G. D., Horlings, H. M., Shah, N., Umbricht, C., Wang, P., *et al.*, 2011. Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. *Nature Genetics*, **43**(7):621–629.

Hurst, L. D., 2006. Preliminary assessment of the impact of microRNA-mediated regulation on coding sequence evolution in mammals. *Journal of Molecular Evolution*, **63**(2):174–182. PMID: 16786435.

Hurst, L. D. and Pal, C., 2001. Evidence for purifying selection acting on silent sites in BRCA1. *Trends in Genetics*, **17**:62–65.

Inglis, D. O., Arnaud, M. B., Binkley, J., Shah, P., Skrzypek, M. S., Wymore, F., Binkley, G., Miyasato, S. R., Simison, M., and Sherlock, G., *et al.*, 2011. The candida genome database incorporates multiple candida species: multispecies search and analysis tools with curated gene and protein information for candida albicans and candida glabrata. *Nucleic Acids Research*, **40**(D1):D667–D674.

Itzkovitz, S. and Alon, U., 2007. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome Research*, **17**(4):405–412.

Ji, Y., Lu, Y., and Mills, G. B., 2008. Bayesian models based on test statistics for multiple hypothesis testing problems. *Bioinformatics*, **24**:943–949.

Jungreis, I., Lin, M. F., Spokony, R., Chan, C. S., Negre, N., Victorsen, A., White, K. P., and Kellis, M., 2011. Evidence of abundant stop codon readthrough in drosophila and other metazoa. *Genome Research*, **21**(12):2096–2113.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E. S., 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**(6937):241–254.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, a. D., 2002. The human genome browser at UCSC. *Genome Research*, **12**(6):996–1006.

Kiryu, H., 2011. Sufficient statistics and expectation maximization algorithms in phylogenetic tree models. *Bioinformatics*, **27**(17):2346–2353.

Klemke, M., 2001. Two overlapping reading frames in a single exon encode interacting proteins–a novel way of gene usage. *The EMBO Journal*, **20**(14):3849–3860.

Klosterman, P. S., Uzilov, A. V., Bendaa, Y. R., Bradley, R. K., Chao, S., Kosiol, C., Goldman, N., and Holmes, I., 2006. XRate: a fast prototyping, training and annotation tool for phylo-grammars. *BMC Bioinformatics*, **7**:428. PMID: 17018148.

Kosiol, C., Holmes, I., and Goldman, N., 2007. An empirical codon model for protein sequence evolution. *Molecular Biology and Evolution*, **24**(7):1464–1479.

Kuczynski, J., Lauber, C. L., Walters, W. A., Parfrey, L. W., Clemente, J. C., Gevers, D., and Knight, R., 2011. Experimental and analytical tools for studying the human microbiome. *Nature Reviews Genetics*, **13**(1):47–58.

Kural, D., Ding, Y., Wu, J., Korpi, A. M., and Chuang, J. H., 2009. COMIT: identification of noncoding motifs under selection in coding sequences. *Genome Biology*, **10**:R133.

Kuroki, Y., Toyoda, A., Noguchi, H., Taylor, T. D., Itoh, T., Kim, D., Kim, D., Choi, S., Kim, I., Choi, H. H., *et al.*, 2006. Comparative analysis of chimpanzee and human y chromosomes unveils complex evolutionary pathway. *Nature Genetics*, **38**(2):158–167.

Lampe, X., Samad, O. A., Guiguen, A., Matis, C., Remacle, S., Picard, J. J., Rijli, F. M., and Rezsohazy, R., 2008. An ultraconserved Hox-Pbx responsive element resides in the coding sequence of hoxa2 and is active in rhombomere 4. *Nucleic Acids Research*, **36**:3214–3225.

Lander, E. S., 2011. Initial impact of the sequencing of the human genome. *Nature*, **470**(7333):187–197.

Lang, G., Gombert, W. M., and Gould, H. J., 2005. A transcriptional regulatory element in the coding sequence of the human bcl-2 gene. *Immunology*, **114**(1):25–36. PMID: 15606792.

Lercher, M. J., Williams, E. J., and Hurst, L. D., 2001. Local similarity in evolutionary rates extends over whole chromosomes in human-rodent and mouse-rat comparisons: implications for understanding the mechanistic basis of the male mutation bias. *Molecular Biology and Evolution*, **18**(11):2032–2039. PMID: 11606699.

Lewis, B. P., Burge, C. B., and Bartel, D. P., 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are MicroRNA targets. *Cell*, **120**:15–20.

Li, C., 2011. Personalized medicine - the promised land: are we there yet? *Clinical Genetics*, **79**(5):403–412.

Li, J. B., Levanon, E. Y., Yoon, J., Aach, J., Xie, B., Leproust, E., Zhang, K., Gao, Y., and Church, G. M., 2009. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science (New York, N.Y.)*, **324**(5931):1210–1213. PMID: 19478186.

Lin, M. F., Carlson, J. W., Crosby, M. A., Matthews, B. B., Yu, C., Park, S., Wan, K. H., Schroeder, A. J., Gramates, L. S., St. Pierre, S. E., *et al.*, 2007. Revisiting the protein-coding gene catalog of drosophila melanogaster using 12 fly genomes. *Genome Research*, **17**(12):000.

Lin, M. F., Deoras, A. N., Rasmussen, M. D., and Kellis, M., 2008a. Performance and scalability of discriminative metrics for comparative gene identification in 12 drosophila genomes. *PLoS Computational Biology*, **4**(4):e1000067.

Lin, M. F., Jungreis, I., and Kellis, M., 2011a. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, **27**(13):i275–i282.

Lin, M. F., Kheradpour, P., Washietl, S., Parker, B. J., Pedersen, J. S., and Kellis, M., 2011b. Locating protein-coding sequences under selection for additional, overlapping functions in 29 mammalian genomes. *Genome Research*, **21**(11):1916–1928.

Lin, Z., Ma, H., and Nei, M., 2008b. Ultraconserved coding regions outside the homeobox of mammalian hox genes. *BMC Evolutionary Biology*, **8**:260.

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., *et al.*, 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, **478**(7370):476–482.

Margulies, E. H., Cooper, G. M., Asimenos, G., Thomas, D. J., Dewey, C. N., Siepel, A., Birney, E., Keefe, D., Schwartz, A. S., Hou, M., *et al.*, 2007. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Research*, **17**(6):760–774.

Massingham, T., 2004. Detecting amino acid sites under positive selection and purifying selection. *Genetics*, **169**:1753–1762.

Mayrose, I., Doron-Faigenboim, A., Bacharach, E., and Pupko, T., 2007. Towards realistic codon models: among site variability and dependency of synonymous and non-synonymous rates. *Bioinformatics*, **23**(13):i319–i327.

Nguyen, M. Q., Zhou, Z., Marks, C. A., Ryba, N. J., and Belluscio, L., 2007. Prominent roles for odorant receptor coding sequences in allelic exclusion. *Cell*, **131**(5):1009–1017.

Nozawa, M., Suzuki, Y., and Nei, M., 2009. Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proceedings of the National Academy of Sciences*, **106**(16):6700–6705.

Ota, R., Waddell, P. J., Hasegawa, M., Shimodaira, H., and Kishino, H., 2000. Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. *Molecular Biology and Evolution*, **17**(5):798 –803.

Ozsolak, F. and Milos, P. M., 2011. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet*, **12**(2):87–98.

Parmley, J. L., 2005. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Molecular Biology and Evolution*, **23**(2):301–309.

Parmley, J. L. and Hurst, L. D., 2007a. Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals. *Molecular Biology and Evolution*, **24**(8):1600–1603.

Parmley, J. L. and Hurst, L. D., 2007b. How common are intragene windows with k a $>$ k s owing to purifying selection on synonymous mutations? *Journal of Molecular Evolution*, **64**:646–655.

Parmley, J. L., Urrutia, A. O., Potrzebowski, L., Kaessmann, H., and Hurst, L. D., 2007. Splicing and the evolution of proteins in mammals. *PLoS Biology*, **5**(2):e14.

Pauli, A., Valen, E., Lin, M. F., Garber, M., Vastenhouw, N. L., Levin, J. Z., Fan, L., Sandelin, A., Rinn, J. L., Regev, A., *et al.*, 2011. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Research*, .

Pedersen, J. S., 2004a. A comparative method for finding and folding RNA secondary structures within protein-coding regions. *Nucleic Acids Research*, **32**(16):4925–4936.

Pedersen, J. S., 2004b. An evolutionary model for Protein-Coding regions with conserved RNA structure. *Molecular Biology and Evolution*, **21**(10):1913–1922.

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., and Siepel, A., 2009. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, **20**:110–121.

Pond, S. K., 2005. Site-to-Site variation of synonymous substitution rates. *Molecular Biology and Evolution*, **22**(12):2375–2385.

Poulin, F., 2003. Gene fusion and overlapping reading frames in the mammalian genes for 4E-BP3 and MASK. *Journal of Biological Chemistry*, **278**(52):52290–52297.

Pozzoli, U., Menozzi, G., Comi, G. P., Cagliani, R., Bresolin, N., and Sironi, M., 2007. Intron size in mammals: complexity comes to terms with economy. *Trends in Genetics*, **23**(1):20–24.

Pruitt, K., Brown, G., Tatusova, T., and Maglott, D., 2002. The reference sequence (RefSeq) database. In *The NCBI Handbook*. National Center for Biotechnology Information (US), Bethesda (MD).

Pruitt, K. D., Harrow, J., Harte, R. A., Wallin, C., Diekhans, M., Maglott, D. R., Searle, S., Farrell, C. M., Loveland, J. E., Ruef, B. J., *et al.*, 2009. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Research*, **19**(7):1316 –1323.

Resch, A. M., Carmel, L., Marino-Ramirez, L., Ogurtsov, A. Y., Shabalina, S. A., Rogozin, I. B., and Koonin, E. V., 2007. Widespread positive selection in synonymous sites of mammalian genes. *Molecular Biology and Evolution*, **24**:1821–1831.

Rhind, N., Chen, Z., Yassour, M., Thompson, D. A., Haas, B. J., Habib, N., Wapinski, I., Roy, S., Lin, M. F., Heiman, D. I., *et al.*, 2011. Comparative functional genomics of the fission yeasts. *Science*, **332**(6032):930–936.

Ribeca, P. and Valiente, G., 2011. Computational challenges of sequence classification in microbiomic data. *Briefings in Bioinformatics*, **12**(6):614–625.

Ribrioux, S., Brngger, A., Baumgarten, B., Seuwen, K., and John, M. R., 2008. Bioinformatics prediction of overlapping frameshifted translation products in mammalian transcripts. *BMC Genomics*, **9**(1):122.

Richard, G., Kerrest, A., and Dujon, B., 2008. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiology and Molecular Biology Reviews: MMBR*, **72**(4):686–727. PMID: 19052325.

Robins, H., Krasnitz, M., and Levine, A. J., 2008. The computational detection of functional nucleotide sequence motifs in the coding regions of organisms. *Experimental Biology and Medicine*, **233**(6):665–673.

Rodrigue, N., Lartillot, N., and Philippe, H., 2008. Bayesian comparisons of codon substitution models. *Genetics*, **180**(3):1579–1591.

Rubinstein, N. D., Doron-Faigenboim, A., Mayrose, I., and Pupko, T., 2011. Evolutionary models accounting for layers of selection in Protein-Coding genes and their impact on the inference of positive selection. *Molecular Biology and Evolution*, **28**(12):3297–3308.

Rueter, S. M., Dawson, T. R., and Emeson, R. B., 1999. Regulation of alternative splicing by RNA editing. *Nature*, **399**(6731):75–80. PMID: 10331393.

Sauna, Z. E. and Kimchi-Sarfaty, C., 2011. Understanding the contribution of synonymous mutations to human disease. *Nature Reviews Genetics*, **12**:683–691.

Schattner, P., 2006. Regions of extreme synonymous codon selection in mammalian genes. *Nucleic Acids Research*, **34**:1700–1710.

Schmid, K. and Yang, Z., 2008. The trouble with sliding windows and the selective pressure in BRCA1. *PLoS ONE*, **3**:e3746.

Schnall-Levin, M., Rissland, O. S., Johnston, W. K., Perrimon, N., Bartel, D. P., and Berger, B., 2011. Unusually effective microRNA targeting within repeat-rich coding regions of mammalian mRNAs. *Genome Research*, **21**:1395–1403.

Schnall-Levin, M., Zhao, Y., Perrimon, N., and Berger, B., 2010. Conserved microRNA targeting in drosophila is as widespread in coding regions as in 3'UTRs. *Proceedings of the National Academy of Sciences*, **107**:15751–15756.

Sharpless, N. E. and DePinho, R. A., 1999. The INK4A/ARF locus and its two gene products. *Current Opinion in Genetics & Development*, **9**(1):22–30.

Siepel, A., 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, **15**(8):1034–1050.

Siepel, A. and Haussler, D., 2004. Phylogenetic estimation of Context-Dependent substitution rates by maximum likelihood. *Molecular Biology and Evolution*, **21**(3):468 –488.

Singh, N. D., Bauer DuMont, V. L., Hubisz, M. J., Nielsen, R., and Aquadro, C. F., 2007. Patterns of mutation and selection at synonymous sites in drosophila. *Molecular Biology and Evolution*, **24**(12):2687–2697.

Smith, N. G. and Hurst, L. D., 1999. The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents. *Genetics*, **153**(3):1395–1402. PMID: 10545467.

Stanley, S. M., Bailey, T. L., and Mattick, J. S., 2006. GONOME: measuring correlations between GO terms and genomic positions. *BMC Bioinformatics*, **7**:94. PMID: 16504139.

Stark, A., Lin, M. F., Kheradpour, P., Pedersen, J. S., Parts, L., Carlson, J. W., Crosby, M. A., Rasmussen, M. D., Roy, S., Deoras, A. N., *et al.*, 2007. Discovery of functional elements in 12 drosophila genomes using evolutionary signatures. *Nature*, **450**(7167):219–232.

Stearns, S. C., Byars, S. G., Govindaraju, D. R., and Ewbank, D., 2010. Measuring selection in contemporary human populations. *Nature Reviews Genetics*, .

Storey, J. D., 2003. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, **100**(16):9440–9445.

Suzuki, R. and Saitou, N., 2011. Exploration for functional nucleotide sequence candidates within coding regions of mammalian genes. *DNA Research*, **18**:177–187.

Suzuki, Y. and Nei, M., 2002. Simulation study of the reliability and robustness of the statistical methods for detecting positive selection at single amino acid sites. *Molecular Biology and Evolution*, **19**(11):1865–1869. PMID: 12411595.

The 1000 Genomes Project Consortium, 2010. A map of human genome variation from population-scale sequencing. *Nature*, **467**(7319):1061–1073. PMID: 20981092.

The ENCODE Project Consortium, 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biology*, **9**(4):e1001046.

The modENCODE Consortium, Meyer, P. E., Robine, N., Washington, N. L., Di Stefano, L., Berezikov, E., Brown, C. D., Candeias, R., Carlson, J. W., Carr, A., *et al.*, 2010. Identification of functional elements and regulatory circuits by drosophila modENCODE. *Science*, **330**(6012):1787 –1797.

Tumpel, S., Cambronero, F., Sims, C., Krumlauf, R., and Wiedemann, L. M., 2008. A regulatory module embedded in the coding region of hoxa2 controls expression in rhombomere 2. *Proceedings of the National Academy of Sciences*, **105**:20077–20082.

Urrutia, A. O., 2003. The signature of selection mediated by expression on human genes. *Genome Research*, **13**(10):2260–2264.

Vuong, Q. H., 1989. Likelihood ratio tests for model selection and Non-Nested hypotheses. *Econometrica*, **57**(2):307–333. ArticleType: research-article / Full publication date: Mar., 1989 / Copyright 1989 The Econometric Society.

Whelan, S. and Goldman, N., 1999. Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. *Molecular Biology and Evolution*, **16**(9):1292.

Whittemore, A. S., 2007. A bayesian false discovery rate for multiple testing. *Journal of Applied Statistics*, **34**:1–9.

Williams, E. J. B. and Hurst, L. D., 2002. Is the synonymous substitution rate in mammals gene-specific? *Molecular Biology and Evolution*, **19**(8):1395–1398. PMID: 12140253.

Woltering, J. M. and Duboule, D., 2009. Conserved elements within open reading frames of mammalian hox genes. *Journal of Biology*, **8**:17.

Wright, F., 1990. The effective number of codons used in a gene. *Gene*, **87**(1):23–29.

Xing, Y. and Lee, C., 2006. Can RNA selection pressure distort the measurement of Ka/Ks? *Gene*, **370**:1–5.

Yang, Z., 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular Biology and Evolution*, **15**(5):568–573. PMID: 9580986.

Yang, Z., 2005. Bayes empirical bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution*, **22**:1107–1118.

Yang, Z., 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**(8):1586 –1591.

Yang, Z. and Bielawski, J. P., 2000. Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution*, **15**(12):496–503.

Yang, Z. and dos Reis, M., 2010. Statistical properties of the Branch-Site test of positive selection. *Molecular Biology and Evolution*, **28**:1217–1228.

Yang, Z. and Nielsen, R., 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of Molecular Evolution*, **46**(4):409–418.

Yang, Z. and Nielsen, R., 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution*, **19**(6):908–917. PMID: 12032247.

Yang, Z. and Nielsen, R., 2008. Mutation-Selection models of codon substitution and their use to estimate selective strengths on codon usage. *Molecular Biology and Evolution*, **25**(3):568–579.

Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A. M., 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, **155**(1):431–449. PMID: 10790415.

Yeo, G. and Burge, C. B., 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of Computational Biology*, **11**(2-3):377–394.

Yoshida, H., Matsui, T., Yamamoto, A., Okada, T., and Mori, K., 2001. XBP1 mRNA is induced by ATF6 and spliced by IRE1 in response to ER stress to produce a highly active transcription factor. *Cell*, **107**(7):881–891.

Zhou, T., Gu, W., and Wilke, C. O., 2010. Detecting positive and purifying selection at synonymous sites in yeast and worm. *Molecular Biology and Evolution*, **27**(8):1912–1922.