2012-07-13

Evaluating the Performance of Computational Approaches for Identifying Critical Sites in Protein-coding DNA Sequences

Matthew Lewis Bendall
Brigham Young University - Provo

Evaluating the Performance of Computational Approaches for Identifying Critical Sites in Protein-coding DNA Sequences

Matthew L. Bendall

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Keith A. Crandall, Chair
Mark J. Clement
W. Evan Johnson

Department of Biology

Brigham Young University

August 2012

Abstract

Evaluating the Performance of Computational Approaches for Identifying Critical Sites in Protein-coding DNA Sequences

Matthew L. Bendall
Department of Biology, BYU
Master of Science

The ability to link a particular phenotype to its causative genotype is one of the most challenging objectives for biological research. Although the genetic code provides an explicit formula for determining the sequence of amino acid phenotypes produced by a given nucleotide sequence, identifying specific residues that are functionally important remains problematic. Many computational approaches have been developed that use patterns observed in DNA sequences to identify these critical sites. However, very few research studies have used empirical data to test whether these approaches are truly able to identify sites of interest.

In most empirical studies, the actual protein function and selective pressures are unknown; thus it is difficult to assess whether computational approaches are correctly identifying critical sites. Here I present two studies that utilize well-characterized empirical systems to evaluate and compare the performance of several computational approaches. In both cases, the proteins under study have specific amino acid substitutions that are confirmed to alter protein function and expected to be constrained by natural selection. In chapter 2, I examine functional variants in angiopoietin-like protein 4 (ANGPTL4), a protein involved in regulating plasma triglyceride levels; loss-of-function variants in this gene are believed to decrease the risk of cardiovascular disease. I apply several computational approaches to identify functional variants, including phylogenetic approaches for detecting positive selection. In chapter 3, I investigate the emergence of drug-resistance in HIV-1 during the course of antiretroviral drug therapy. I compare the performance of eight selection detection methods in identifying drug-resistant mutations in 109 intrapatient datasets with HIV-1 sequences isolated at multiple timepoints throughout drug treatment.

It is critical that we develop methods to detect positively selected sites. The ability to detect these sites *in silico*, without the need for expensive and time consuming assays, would be invaluable to researchers in evolutionary biology, human genetics, and medicine. Through the research presented in this thesis, I hope to provide insight into the strengths and weaknesses of current approaches, thereby facilitating future research towards the development and improvement of evolutionary models.

Keywords: positive selection, evolutionary models, HIV-1 drug resistance, ANGPTL4

Acknowledgments

I sincerely appreciate the guidance, support, and patience of my advisor, Dr. Keith Crandall. His continual confidence in me has truly enhanced my development as a scientist, and his example has inspired me to set the highest goals in all my future endeavors.

I would like to express my gratitude to members of my thesis committee, Dr. Mark Clement and Dr. Evan Johnson. I appreciate Dr. Clement's mentoring, both as a professor and committee member. His enthusiasm has greatly enhanced my interest in the field. I am grateful to Dr. Johnson for contributing his expertise, and for going the extra 2,000 miles to be present at my defense.

My research would not have been possible without the assistance and encouragement of my colleagues, Dr. Heather Bracken-Grissom, Jesse Breinholt, Maegan Leary, Eduardo Castro, Dr. Seth Bybee, and Dr. Madelaine Bartlett. I would especially like to thank Dr. Nicole Lewis-Rogers for her mentorship as I entered the world of academic research.

I would like to thank my grandparents, Walter and Louella Chiou, who have always emphasized the importance of education in our family. I am grateful to my father, Dan Bendall, who has taught me the value of hard work and perseverance. Thank you to my mother, Debbie Bendall, for always challenging me to excel and to be my best self. I appreciate the prayers and support offered by my family and friends.

This thesis would not have been possible without the love and support of my wife, Michele. No matter how frustrated or discouraged I was feeling, she was always there to encourage me. Whenever I was hungry because I forgot to eat, she made sure that I had a delicious meal. She has supported me every step of the way, even when it meant that she would have to double her course load so we could be finished at the same time. Her beautiful example helps me to always remember what is most important and to trust in the Lord.

Contents

List of Tables

Chapter 2

Chapter 3

Appendix B

List of Figures

Chapter 2

Chapter 3

Chapter 1

Introduction

The ability to link a particular phenotype to its causative genotype is one of the most challenging objectives for biological research. At the molecular level, the genotype can be readily identified through DNA sequencing technology, while the phenotype is often more difficult to discern. Although the genetic code provides an explicit formula for determining the sequence of amino acid phenotypes produced by a given nucleotide sequence, identifying specific residues that are functionally important remains problematic. Many computational approaches have been developed that use patterns observed in DNA sequences to identify these critical sites.

The statistical models used by many approaches are based directly on the evolutionary principles of mutation, genetic drift, and natural selection. These processes are intertwined, and the data we observe is the result of complex interactions among these forces. The processes of mutation and gene flow introduce heritable variation in a population. Genetic drift and natural selection regulate the frequencies of a given variation in the next generation. Genetic drift acts randomly; variation may become fixed or lost due to random chance. On the other hand, natural selection is deterministic; variants that confer a reproductive advantage will increase in frequency, while deleterious variation will be removed from the population.

Natural selection is the primary process that acts on the phenotype of the individual; therefore, identifying sites under natural selection is crucial for identifying important amino acid sites. In practice, sites are predicted to be under natural selection when inferred evolutionary changes are significantly different than the random changes that are expected under neutral evolution. Computational approaches for detecting positive selection differ in the techniques used for inferring evolutionary change and in the null hypothesis assumed.

In this thesis, I present two studies comparing computational approaches for identifying critical sites in protein-coding DNA sequences. In the first study, I investigate whether phylogenetic methods for identifying positively selected sites are suitable for identifying functional variants in angiopoietin-like protein 4 (ANGPTL4). The second study examines the performance of eight different site-prediction methods when applied to emerging drug-resistance in HIV-1.

1.1 Identifying adaptive evolution

Experimental verification

The "gold standard" for identifying adaptive substitutions is through experimental verification. This technique uses three-dimensional structures or functional assays to associate observed amino acid changes with phenotypic differences. Although this provides convincing evidence of adaptive change, in most cases experimental verification is prohibitively expensive and time-consuming. Investigators who wish to verify adaptive substitutions in the laboratory often rely on computational approaches to generate hypotheses, thus reducing the sites or substitutions that must be considered.

In the studies presented here, the amino acid sites and residues that are responsible for functional changes in the protein have been previously determined through experimental verification. With this information, I am able to conclude whether a particular method is successfully identifying known adaptive sites.

Predicting functional variants in ANGPTL4

In this study, published in the International Journal of Molecular Sciences, I analyzed nucleotide sequences from ANGPTL4, a protein implicated in cardiovascular disease (Maxwell et al. 2010). This data was obtained from the Dallas Heart Study, a population-based random study of ethnic differences in cardiovascular health (Victor et al. 2004). I use several computational approaches, including phylogenetic site-prediction methods, to identify variants that cause functional changes in the protein. I compared the results of these analyses to the actual functional status of each variant confirmed by *in vitro* assay.

Detecting drug-resistant mutations in HIV-1

I compare eight different computational methods for detecting positive selection using empirical data. The evolution of drug resistance in HIV-1 is one of the few biological study systems where positive selection can be observed in real time. I designed 109 intra-patient datasets that demonstrate the evolution of resistance to antiretroviral drugs during the course of treatment. The specific drug-resistant mutations that are under positive selection are well-documented in the scientific literature. Each of the eight approaches was used to predict positively selected sites; performance and other statistical properties were compared among methods.

1.2 Conclusion

It is critical that we develop methods to detect positively selected sites. The ability to detect these sites *in silico*, without the need for expensive and time consuming assays, would be invaluable to researchers in evolutionary biology, human genetics, and medicine. However, current methods for identifying positively selected sites are clearly missing crucial aspects of the evolutionary process. By comparing existing methods for detecting selection, I am hoping to provide new insights into the strengths and weaknesses of current models, thereby providing information towards the improvement and development of evolutionary models.

References

T. J. Maxwell, M. L. Bendall, J. Staples, T. Jarvis, and K. A. Crandall. Phylogenetics applied to Genotype/Phenotype association and selection analyses with sequence data from ANGPTL4 in humans. *International Journal of Molecular Sciences*, 11(1):370385, 2010.

R. G. Victor, R. W. Haley, D. L. Willett, R. M. Peshock, P. C. Vaeth, D. Leonard, M. Basit, R. S. Cooper, V. G. Iannacchione, W. A. Visscher, J. M. Staab, and H. H. Hobbs. The dallas heart study: a population-based probability sample for the multidisciplinary study of ethnic differences in cardiovascular health. *The American Journal of Cardiology*, 93 (12):1473–1480, June 2004.

Chapter 2

Predicting functional variants in ANGPTL4

Preface

In this research study, published in the International Journal of Molecular Sciences, I analyzed nucleotide sequences encoding angiopoietin-like protein 4 (ANGPTL4), a protein implicated in cardiovascular disease (Maxwell et al. 2010). A major objective of this study was to investigate whether phylogenetic-based selection detection methods are useful for identifying functional variants. To test this hypothesis, I used ANGPTL4 sequence data sampled from 3,551 individuals in the Dallas Heart Study (Victor et al. 2004). ANGPTL4 inhibits the activity of lipoprotein lipase (LPL), resulting in an increase in plasma triglyceride levels (Romeo et al. 2007). Functional studies have confirmed several nonsynonymous substitutions that interfere with protein function or secretion (Romeo et al. 2009). It has been shown that loss-of-function mutations are associated with low plasma triglyceride levels and it is believed that such variants may decrease the risk of cardiovascular disease (Smart-Halajko et al. 2011).

I use three computational approaches to identify variants that influence protein function. The first approach, implemented in PolyPhen, predicts whether variants are "benign" or "damaging" using homologous protein alignments (Ramensky et al. 2002). Prediction is based on position-specific amino acid profiles and position in the protein structure. The second approach, implemented in TreeSAAP, infers substitutions that produce radical or conservative shifts in amino acid properties (Woolley et al. 2003). Substitutions are inferred using a phylogenetic approach and examined with respect to 31 physicochemical properties. Finally, I use two likelihood-based site-prediction methods to identify sites under positive selection (Pond et al. 2005; Yang 2007).

Individual contribution

My individual contributions to this research study include:

Phylogenetic analysis Performed multiple sequence alignment, model fitting, maximum likelihood phylogenetic reconstruction, and statistical parsimony network reconstruction.

TreeSAAP analysis Executed analysis, parsed output files, interpreted and summarized results.

Likelihood analysis Identified homologous protein domains, partitioned sequences, and performed partitioned selection analyses in PAML and HyPhy.

Software utilities Created the following perl scripts (see Appendix A):

- TCSparser.pl - manipulates phylogenetic network
- PPparser.pl - annotates network using PolyPhen output
- TSAAPparser.pl - annotates network using TreeSAAP output

Manuscript sections Authored or contributed to the manuscript text.

Primary author of the following sections:

- 2.2.2 (PAML and HyPhy results)
- 3.4.2 (TreeSAAP analysis)
- 3.4.3 (Likelihood Selection Analysis)

Significant contributions for the following sections:

- 2.1.1 (Variants, Haplotypes, Networks, and Phylogenetic Trees)
- 2.2.1 (PolyPhen and TreeSAAP Results)
- 2.3 (Comparison of PolyPhen and TreeSAAP)

- 3.2 (Haplotype Networks and Phylogenetic Trees)
- 3.5 (Comparison of TreeSAAP and PolyPhen)
- 4 (Conclusions)

Figures Created all figures and wrote all figure legends.

- Phylogenetic network of sampled haplotypes (Figure1)
- Schematic representation of ANGPTL4 coding region (Figure2)

Tables Contributed to Table 4 (Frequencies of sampled haplotypes)

Review process Provided comments and revisions throughout the review process

Int. J. Mol. Sci. **2010**, *11*, 370-385; doi:10.3390/ijms11010370

OPEN ACCESS

International Journal of
Molecular Sciences
ISSN 1422-0067
www.mdpi.com/journal/ijms

Article

Phylogenetics Applied to Genotype/Phenotype Association and Selection Analyses with Sequence Data from Angptl4 in Humans

Taylor J. Maxwell [1],*, Matthew L. Bendall [2], Jeffrey Staples [2], Todd Jarvis [2] and Keith A. Crandall [2]

[1] Human Genetics Center, University of Texas School of Public Health, Houston, TX 77030, USA
[2] Department of Biology, Brigham Young University, Provo, UT 84602, USA;
E-Mails: matthew.bendall@gmail.com (M.L.B.); grapas2@gmail.com (J.S.);
todd.jarvis@gmail.com (T.J.); keith_crandall@byu.edu (K.A.C.)

* Author to whom correspondence should be addressed; E-Mail: Taylor.J.Maxwell@uth.tmc.edu;
Tel.: +1-713-500-9896; Fax: +1-713-500-0900.

Received: 13 November 2009; in revised form: 6 January 2010 / Accepted: 17 January 2010 / Published: 25 January 2010

Abstract: Genotype/phenotype association analyses (Treescan) with plasma lipid levels and functional site prediction methods (TreeSAAP and PolyPhen) were performed using sequence data for ANGPTL4 from 3,551 patients in the Dallas Heart Study. Biological assays of rare variants in phenotypic tails and results from a Treescan analysis were used as "known" variants to assess the site prediction abilities of PolyPhen and TreeSAAP. The E40K variant in European Americans and the R278Q variant in African Americans were significantly associated with multiple lipid phenotypes. Combining TreeSAAP and PolyPhen performed well to predict "known" functional variants while reducing noise from false positives.

Keywords: ANGPTL4; TreeSAAP; treescan; phylogenetics; association studies; selection

1. Introduction

No single method of analysis is sufficient to uncover all the information that can come from sequence data. What we can strive for is a set of methods that complement each other. For example, the fields of molecular evolution, phylogenetics, and population genetics have a long history of sequence analysis [1,2]; however these methods do not typically use phenotype information. Many of these methods use knowledge about gene structure, amino acids, protein structure, and phylogenetics. We can borrow methods from these fields to identify polymorphic sites that may show evidence for selection or are likely to cause significant changes in expression or the nature of a protein.

Romeo *et al*. [3,4] sequenced the exonic regions and boundaries for the ANGPTL4 (angiopoietin-like protein 4) gene in patients from the Dallas Heart Study [5]. Results from analysis of these data [3] and subsequently in other ANGPTL genes [4] found that rare variants substantially contribute to variation in triglyceride levels. These groundbreaking papers substantiated these claims with biological assays showing that most rare variants in individuals in the tails of the triglyceride phenotypic distribution were functionally important by affecting secretion, expression, LDL inhibition, or loss of function. These results and data give us a rare opportunity to use "known" functional variants to assess the relative abilities of some site prediction methods such as PolyPhen [6] and TreeSAAP [7].

Using this data, we performed a series of analyses using phylogenetic approaches. We used Treescanning [8] to identify variants associated with lipid phenotypes. We used PAML [9] and HyPhy [10] to describe selection patterns across the sequence. Finally, we used the known rare functional variants from Romeo *et al.* [3,4] and results for common variants from the Treescanning analyses to compare the relative specificity and sensitivity of PolyPhen, TreeSAAP, and various combinations of the two.

2. Results and Discussion

2.1. Phylogenetic and Treescanning Results

2.1.1. Variants, Haplotypes, Networks, and Phylogenetic Trees

Including a human reference sequence, there were 39 variants (27 missense, 11 synonymous) that produced 45 unique haplotypes. One missense mutation (G77R) from the previous study [3,4] was not included because the individual that harbored it had too much missing data to reliably infer its haplotypes. Four other variants (IVS3+1, K217X, FsK245, and FsS302) that were nonsense, frame shifts, or splicing mutations were not included in the selection analyses because the selection detection methods only consider amino acid replacements.

The haplotype inferences were relatively easy because most individuals were heterozygous for only one site. Technically, the singleton variants cannot be definitively placed on a haplotype unless it is heterozygous for only that site. However, because Treescanning uses genotypes, these individuals will always be grouped in the heterozygous class when the two possible haplotype backgrounds are defined as different allele classes making the test invariant to the phasing of the singleton. Regardless, the treescanning results were the same when all singletons were excluded. As for the phylogenetic analyses, the short branch lengths suggest that they will have little impact on analysis. For TreeSAAP, the same substitution event will always be inferred as long they are seen as tips.

The bootstrap analysis for the maximum likelihood (ML) tree revealed low resolution throughout the tree. The reason for such low resolution is the shortness of the branches. The haplotype network (Figure 1) illustrates this. Every single branch in the tree is only one step long, meaning that no haplotype is more than one site different than its nearest neighbor in the network. Bootstrapping works by sampling sites with replacement, which means that a site on a particular branch will be excluded in some of the replicates. Only branches with many sites will show any confidence in a bootstrap analysis. However, in coalescent theory, these short connections are considered more likely. Statistical Parsimony [11] was designed to incorporate these criteria when the haplotypes are sampled within a population. Another feature of this network is that two haploypes (H1 and H2) represent 70% to 80% of each population (see Table 1). Almost every other haplotype is a single step from either of these two haplotypes.

Figure 1. Phylogenetic network showing relationships among sampled haplotypes. Edges are labeled with the base or amino acid change and colored based on results from significant.PolyPhen and TreeSAAP results. The nodes are colored according to the known effect of the variant on the protein as determined by *in vitro* assays [3,4]. Yellow nodes indicate variants that prevent secretion; orange nodes indicate variants that cause a non-functional protein to be secreted; white nodes were not tested *in vitro*; and gray is a synonymous substitution.

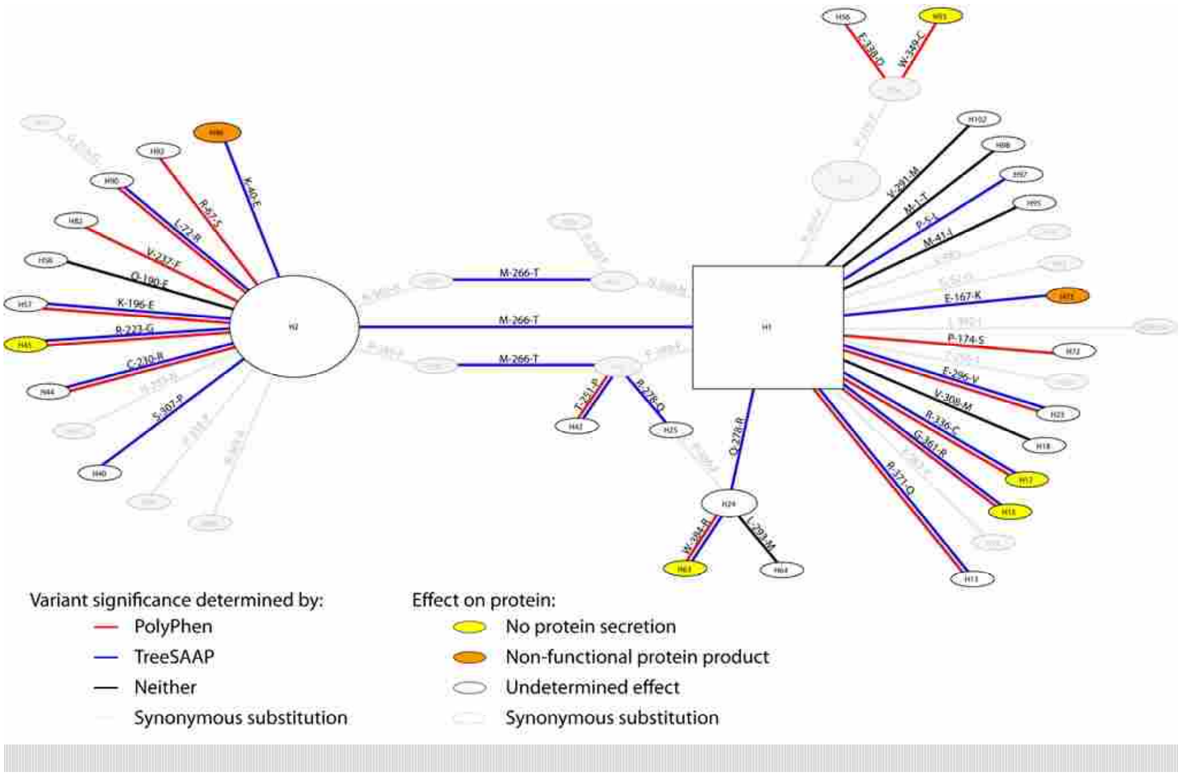


Table 1. Haplotype frequencies for the haplotypes in Figure 1 overall and each population. All = combined; EA = European American, AA = African American, MA = Mexican American.

Haplotype	All	EA	AA	MA	Other
h1	0.51242	0.54341	0.47735	0.54052	0.68493
h2	0.26020	0.28447	0.21706	0.34828	0.23288
h12	0.06475	0.14402	0.02102	0.05862	0.06164
h13	0.00015	0	0.00030	0	0
h14	0.00015	0.00051	0	0	0
h15	0.00030	0.00102	0	0	0
h16	0.04186	0.00153	0.07936	0.00517	0.00685
h17	0.00105	0.00255	0.00030	0.00086	0
h18	0.00045	0.00102	0	0.00086	0
h19	0.01220	0	0.02399	0	0
h23	0.00015	0.00051	0	0	0
h24	0.02997	0.00051	0.05774	0.00259	0
h25	0.00015	0	0.00030	0	0
h26	0.00030	0	0.00059	0	0
h35	0.00045	0	0.00089	0	0
h36	0.00015	0	0.00030	0	0
h37	0.00211	0	0.00415	0	0
h39	0.00015	0	0.00030	0	0
h40	0.00015	0.00051	0	0	0
h41	0.00015	0	0.00030	0	0
h42	0.00015	0.00051	0	0	0
h44	0.00015	0	0	0.00086	0
h45	0.00015	0.00051	0	0	0
h53	0.00030	0	0.00059	0	0
h54	0.05285	0.00204	0.10127	0.00431	0
h55	0.00015	0	0.00030	0	0
h56	0.00015	0	0.00030	0	0
h57	0.00015	0.00051	0	0	0
h58	0.00467	0.00153	0.00651	0.00517	0
h63	0.00015	0	0.00030	0	0
h64	0.00015	0	0.00030	0	0
h72	0.00015	0	0.00030	0	0
h73	0.00015	0.00051	0	0	0
h82	0.00015	0.00051	0	0	0
h90	0.00015	0	0.00030	0	0
h91	0.00030	0	0.00059	0	0
h92	0.00030	0.00051	0.00000	0	0.00685
h93	0.00060	0	0.00118	0	0
h94	0.00015	0	0.00030	0	0
h95	0.00407	0	0	0.02328	0
h96	0.00708	0.01277	0.00296	0.00948	0.00685
h97	0.00030	0	0.00059	0	0
h98	0.00015	0	0.00030	0	0
h102	0.00015	0.00051	0	0	0

2.1.2. Treescanning Results

As found by Romeo *et al.* [3], the branch carrying the E40K variant in the European American population was associated with various phenotypes. It was significant after correcting for multiple tests for triglycerides (multiple p = 0.0277), LDL (multiple p = 0.0141), and nominally for VLDL (nominal p = 0.0147, multiple p = 0.065). The full multivariate model for E40K is significant after multiple tests (multiple p = 0.0064). The univariate p-values are significant for triglycerides, LDL, and VLDL; however, the partial Wilk's tests are only significant for LDL (p = 0.001) and nearly for triglycerides (p = 0.0556). This suggests that LDL probably contributes most to the association in the presence of the other variables followed by triglycerides while the univariate association of VLDL is probably accounted for correlations between the phenotypes. E40K was also nominally significant for triglycerides in the Mexican American populations even with a very small count of nine heterozygotes carrying the K variant (haplotype H96). It displayed the same protective effect of lower triglyceride levels as that found in the European Americans. No other variants within the European American population were significant in the second round of Treescanning.

The branch carrying the R278Q variant was significant for HDL (multiple p = 0.0122) in the African American population. The Q allele (carried by haplotypes H23, H24, H63, and H64) is fairly common in African Americans (see Table 1) at about 5.8% but is very rare in all other populations. The full multivariate model was not significant after multiple test corrections but was nominally significant (nominal p = 0.0394). Triglycerides were nominally significant for the univariate test (nominal p = 0.0359) but this effect went away using the multivariate context of a partial Wilk's test (p = 0.3814); however, HDL-c remained significant using the partial Wilk's test (p = 0.0098). The QQ homozygote shows much higher adjusted HDL-c levels (67.88 mg/dL; n = 7) *versus* the RQ heterozygote (55.13 mg/dL; n = 156) and RR homozygote (52.06 mg/dL; n = 1263). This finding has not previously been reported, however the sample of the QQ homozygotes is relatively small (n = 7). The significance of R278Q still held up, even after conditioning for E40K in the African American population (multiple p = 0.0108).

Talmud *et al*. [12] found mild evidence for an association with triglycerides with the T266M variant which is also the variant that separates the two major haplotypes (H1 and H2) in the network (see Figure 1). They found that this effect went away after conditioning upon the E40K variant. It is easy to see in Figure 1 how this variant could show an effect due to correlation with the E40K variant. Historically, the 40K mutation occurred on a 266M background resulting in linkage disequilibrium (LD) between the two variants. However, Talmud *et al.* [12] also found that T266M (but not E40K) was associated with postprandial triglyceride and glucose levels in a case/control study for individuals with a paternal history of myocardial infarction. In our study, T266M shows no association with any of the phenotypes and in any of our populations. At present, we do not have any data on postprandial stress to follow up their significant association with T266M.

2.2. Bioinformatics and Site Prediction Analysis Results

2.2.1. PolyPhen and TreeSAAP Results

PolyPhen identified eight residues as “probably damaging” and seven as “possibly damaging” for a total of 15, leaving 12 as benign (see Table 2). Of the eight residues that were either functional or significant, PolyPhen identified five as “probably damaging” and three as benign. TreeSAAP identified 10 at category 8, 5 with category 6 or 7, and 12 as nothing. Of the eight residues that were either functional or significant, TreeSAAP identified five at category 8, two with category 6 or 7, and one as nothing. While both methods predicted similar numbers of sites under selection, only nine sites were found in common between the methods while 10 sites were unique to a particular method.

Table 2. PolyPhen and TreeSAAP results for each missense polymorphisms used in the study. Each rare variant is defined by which part of the triglyceride phenotype distribution it was found (H = high, M = Middle, L= Low) according to Romeo *et al.* [3]. For the five common missense variants, Significant and NonSig (Nonsignificant) refers to phenotypic associations from the Treescan analyses. The Biological Assay column refers to assays in Table 3 of Romeo *et al*. [4]. A “-“ means no tests were performed. All significant PolyPhen predictions are in bold. All TreeSAAP properties considered significant with a score of 6 or more [13] are reported, all with an extreme value of 8 are in bold. The TreeSAAP property symbol key is provided below.

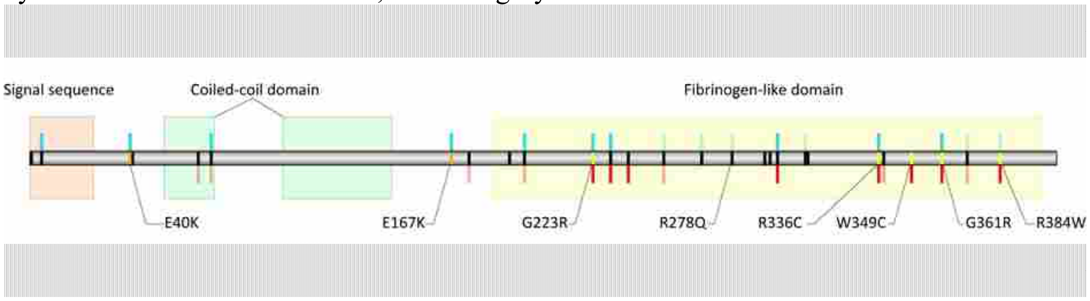
Missense Variant	Phenotype Distribution	Biological Assay	PolyPhen Score	PolyPhen Prediction	TreeSAAP Property
M-1-T	M	-	NA	benign	
P-5-L	M	-	NA	benign	**αc**, **αn**, K0, Hp
E-40-K	Significant	-	1.424	benign	**pHi**
M-41-I	NonSig	-	1.16	benign	
S-67-R	M	-	1.563	**possibly damaging**	
R-72-L	M	-	1.958	**possibly damaging**	**H**, **Hnc**, αn
E-167-K	L	LPL Inhib	0.194	benign	**pHi**
P-174-S	M	-	1.715	**possibly damaging**	
E-190-Q	NonSig	-	0.243	benign	
E-196-K	M	-	1.541	**possibly damaging**	**pHi**, El
G-223-R	L	Secretion	2.065	**probably damaging**	**E’sm**
R-230-C	M	-	2.792	**probably damaging**	**pHi**, **E’sm**, **Et**, **Br**, Ns, C
F-237-V	M	-	2.51	**probably damaging**	
P-251-T	H	Nothing	1.781	**possibly damaging**	
T-266-M	NonSig	-	0.783	benign	K0, Ht
R-278-Q	Significant	-	0.644	benign	pHi
V-291-M	M	-	1.012	benign	
L-293-M	M	-	1.236	benign	
E-296-V	M	-	2.057	**probably damaging**	**Ns**, **Pβ**, Br, H, Ra
P-307-S	M	-	0.955	benign	αc

Table 2. *Cont.*

Missense Variant	Phenotype Distribution	Biological Assay	PolyPhen Score	PolyPhen Prediction	TreeSAAP Property
V-308-M	M	-	1.199	benign	
R-336-C	L	Secretion	2.255	**probably damaging**	**Br**, **pHi**, **Et**, Ns, C, Ca, Hnc
D-338-E	M	-	1.626	**possibly damaging**	
W-349-C	L	Secretion	3.677	**probably damaging**	
G-361-R	L	Secretion	2.274	**probably damaging**	**Ca**, **E'sm**, Mv, Mw, Hnc, V0, μ
R-371-Q	H	Nothing	1.558	**possibly damaging**	pHi
R-384-W	L	Secretion	2.304	**probably damaging**	Br, Ht

TreeSAAP Property Key			
Alpha-helical tendency	Pα	Molecular weight	Mw
Average # of surrounding residues	Ns	Normalized hydrophobicity	Hnc
Beta-structure tendency	Pβ	Partial specific volume	V0
Buriedness	Br	Power to be at the C-terminal	αc
Composition	C	Power to be at the N-terminal	αn
Compressibility	K0	Refractive index	μ
Helical contact	Ca	Short-range & medium-range nonbonded energy	E'sm
Hydropathy	H	Solvent accessible reduction ratio	Ra
Isoelectric point	pHi	Surrounding hydrophobicity	Hp
Long-range nonbonded energy	El	Thermodyn. transfer hydrophobicity	Ht
Molecular volume	Mv	Total non-bonded energy	Et

Figure 2. A schematic representation of ANGPTL4 coding region. The locations of variant sites are colored according to their affect on protein functionality as previously described [4]. Yellow sites prevent protein secretion, orange sites cause a non-functional protein to be secreted, and black sites were not tested *in vitro*. Amino acid sites identified by PolyPhen as "possibly damaging" are indicated in light red; "probably damaging" sites are shown in dark red. Radically changing (categories 6, 7, and 8) amino acid sites identified by TreeSAAP are shown in blue, with category 8 sites in dark blue.



2.2.2. PAML and HyPhy results

The PAML [9] M8 site prediction analysis did not find any sites under positive selection. A likelihood ratio test between the null model (M7) and the positive selection model (M8) did not support a class of sites under positive selection (ω > 1). The dual-rate random effects analysis implemented in HyPhy [10] did not detect positive selection at the absolute threshold of 0.95. Purifying selection was only detected on sites where synonymous substitutions had occurred. We conclude that likelihood-based site prediction methods were ineffective at identifying functional variants for our data. Our findings correspond to other studies concluding that TreeSAAP is more sensitive than likelihood-based site prediction methods for identifying sites under adaptive selection [13,14].

However, likelihood-based methods proved to be useful in characterizing the selective constraint over distinct functional regions of ANGPTL4. We used the one ratio method (M0) implemented in PAML [9] to estimate the variation in functional constraint across the protein. The value of ω for the coiled-coil and fibrinogen-like domains is 1.057 and 0.386, respectively (see Figure 2). Using $\omega = 1$ as the threshold between positive and negative selection, these results indicate that the coiled-coil domain is under nearly neutral selection, while the fibrinogen-like domain is under strong purifying selection. However, the metric of $\omega = 1$ has been shown to underestimate the true amount of selective pressure on a protein region [15]. We estimated the value of ω to be 0.480 for the entire coding region. By using $\omega = 0.480$ as a baseline, the coiled-coil domain appears to be under positive selection with respect to the rest of the gene, while the fibrinogen-like domain is under slightly negative (purifying) selection. These results suggest that the fibrinogen-like domains are under stronger functional constraint than the coiled-coil domain.

The two domains of ANGPTL4 each have unique selective pressures that are driving the evolution of these domains. Post-translational processing cleaves the coiled-coil and fibrinogen-like domains. The coiled-coil domain is involved in the inhibition of LPL, which results in high triglyceride levels. The exact function of the fibrinogen-like domain is not well known. However, it is clear that the functional role performed by each domain is vastly different, and these differences in function would imply a specific set of evolutionary constraints. This is affirmed by the discrepancy between nonsynonymous and synonymous substitution rate ratios. It is interesting to observe that the five variants found to affect secretion from *in vitro* assays are all found in the fibrinogen-like domain (Figure 2).

2.3. Comparison of PolyPhen and TreeSAAP

While both PolyPhen and TreeSAAP identified similar numbers of mutations under selection, they differed considerably in terms of which mutations each identified (Table 3). Only two criteria show a significant difference between the Functional column and the “Middle or Not Sig” column: The TreeSAAP alone criteria or the Strict PolyPhen and Strict TreeSAAP criteria. Both share a very high sensitivity (87.5%) however the TreeSAAP alone criterion has a slightly higher false positive rate (41.2%) and also misclassifies the two high-tail nonfunctional variants. As expected, there is a trend of lower false positive (alpha) rates as we move to the stricter criteria, which is also accompanied by lower sensitivity (power). The Strict PolyPhen & Strict TreeSAAP criteria for significance have the highest specificity but also the lowest sensitivity. The Strict PolyPhen criteria may have the best combination of specificity and sensitivity.

Table 3. A comparison of results between PolyPhen, TreeSAAP, and their combinations with "known" data. Strict PolyPhen only counts "probably damaging" as significant while Strict TreeSAAP only counts category 8 as significant. P-values are from a two-tailed Fisher's exact test of a 2 by 2 table comparing the "Functional or Significant column to the "Middle or Not Sig" column. Sensitivity, specificity, alpha, and beta levels are from this comparison.

Significance Criteria		**Functional or Significant**	**Tested Not Functional**	**Middle or Not Sig**	**p-val**	**Odds Ratio**	**Lower 95 CI**	**Upper 95 CI**
PolyPhen	Significant	5	2	8	0.673	1.828	0.254	15.766
	Not Significant	3	0	9				
	Sensitivity	0.625	**Specificity**	0.529	**alpha**	0.471	**beta**	0.375
TreeSAAP	Significant	7	2	7	**0.042**	9.130	0.859	493.088
	Not Significant	1	0	10				
	Sensitivity	0.875	**Specificity**	0.588	**alpha**	0.412	**beta**	0.125
Strict PolyPhen	Significant	5	0	3	0.061	7.012	0.846	77.356
	Not Significant	3	2	14				
	Sensitivity	0.625	**Specificity**	0.824	**alpha**	0.176	**beta**	0.375
Strict TreeSAAP	Significant	5	0	5	0.194	3.762	0.505	34.675
	Not Significant	3	2	12				
	Sensitivity	0.625	**Specificity**	0.706	**alpha**	0.294	**beta**	0.375
PolyPhen & TreeSAAP	Significant	4	0	4	0.359	3.084	0.385	27.020
	Not Significant	4	2	13				
	Sensitivity	0.5	**Specificity**	0.765	**alpha**	0.235	**beta**	0.5
Strict PolyPhen & Strict TreeSAAP	Significant	3	0	2	0.283	4.192	0.369	64.438
	Not Significant	5	2	15				
	Sensitivity	0.375	**Specificity**	0.882	**alpha**	0.118	**beta**	0.625
Strict PolyPhen OR Strict TreeSAAP	Significant	7	0	6	**0.030**	11.526	1.077	626.871
	Not Significant	1	2	11				
	Sensitivity	0.875	**Specificity**	0.647	**alpha**	0.353	**beta**	0.125

The purpose of these comparisons is to determine what is the best way to use these methods to define a subset of variants for biological assays and/or association analyses. For rare variants, individual association tests are meaningless; however, the phenotype data can be used in conjunction with these methods to narrow the likely candidates. In this case, most of the rare variants in the tails of triglyceride were functionally relevant according to biological assays. If these are a subset of variants sent for testing, both PolyPhen and TreeSAAP perform very well. TreeSAAP was able to identify five of the six rare functional low-tail functional variants plus the two phenotypically associated variants while wrongly finding the two high-tail nonfunctional variants as significant. The two misclassifications disappear when moving to more strict criteria where both methods are identical.

Both methods can be complementary because they give different information and have different aims. PolyPhen attempts to determine if a variant will damage a protein. TreeSAAP tries to identify mutations that are extremely out of the norm relative to the substitution patterns observed in the data for a specific biochemical property. The Strict PolyPhen and Strict TreeSAAP criteria suggest a variant has a high likelihood of importance by a least one method. In many cases, both methods give significant results because a variant is both damaging and it is a very extreme mutation according to the empirical data. It is not surprising that these two methods, which differ in their criteria for determining selection, differ in their outcomes. What is more surprising is that these methods that explore functional differences perform much better than the approaches (PAML and HyPhy) that simply look at dn/ds ratios. Clearly with these population genetic data, examining functional differences seems to provide greater insights into sites under natural selection.

Based on these limited results, we recommend a combination of the two methods that look at functional variants in a population to be most desirable for choosing variants to create *a priori* tests. If the investment in following up with biological assays is very high then the Strict PolyPhen and Strict TreeSAAP criteria are a very strict filter that together have the lowest false positive rate. However, if the goal is to be inclusive, the Strict PolyPhen OR Strict TreeSAAP criterion was very sensitive while still lowering the false positive rate.

3. Materials and Methods

3.1. Study Description and Genetic Data

The Dallas Heart Study is based on a population sample restricted to the Dallas area [5]. That is, individuals in the sample were ascertained randomly without reference to their phenotypic values or disease status. The samples sequenced for the ANGPTL4 gene contains 3,551 individuals (1,830 African Americans, 601 Hispanics, 1045 European Americans, and 75 other ethnicities). All exons from each gene were sequenced along with each intron/exon boundary. All sequencing was done at the Joint Genome Institute. Base calling, quality assessment and assembly were carried out using the Phred, Phrap, Polyphred, Consed software suite. All sequence variants identified were verified by manual inspection of the chromatograms, and missense mutations were confirmed by independent resequencing [3,16]. Five quantitative lipid measures related to heart disease were analyzed: Triglyceride, HDL, VLDL, LDL, and total cholesterol levels.

3.2. Haplotype Networks and Phylogenetic Trees

All exonic regions were aligned and haplotypes were statistically inferred from the genotype data, using PHASE 2.2 [17,18]. A haplotypes network was inferred using a modified version of TCS [19]. The haplotype tree showed no evidence for recombination [20]. Coalescent criteria [21,22] allowed for resolution of each loop by breaking the H16-H37, H12-H35, and H24-H25 branches in Figure 1.

Likelihood scores were calculated from the sequences for the unique haplotypes for 56 models of nucleotide evolution using PAUP* [23]. We determined the best-fit model of nucleotide evolution using a maximum likelihood ratio test implemented in Modeltest [24]. The HKY model [25] with a gamma distribution shape parameter of 0.0104 and a ti/tv ratio of 2.1982 was determined to be the best model given the data. A phylogenetic tree was estimated using the maximum likelihood criterion as implemented in the application PhyML [26]. Branch support for the tree was estimated using non-parametric bootstrap sampling with 1,000 replicates. The ML tree was used for all analyses with TreeSAAP, PAML, and HyPhy.

3.3. Genotype/Phenotype Association via Treescan

After being adjusted for age, sex, and BMI, separate analyses for cholesterol, triglyceride, VLDL, LDL, and HDL levels were performed separately for African-Americans, Mexican-Americans, and European-Americans. Romeo *et al.* [3,4] found a number of rare variants that were functionally significant through biological assays. These known effects may group in ways that may affect associations at other common polymorphisms and branches in the network. Therefore, analyses were performed with and without the individuals harboring these variants. The estimated haplotype network was used for all Treescan [8,27,28] analyses. All treescanning analyses used genotypes as factors and only included genotypic classes with counts of five or more. All nominal and multiple-test corrected significance levels were obtained with 10,000 permutations. A permutation analog of the sequential step-down Bonferroni [29] was used for multiple test correction because it takes into account the correlation between tests.

Because the five lipid phenotypes are biologically related to each other through hepatic and intestinal lipid metabolism, the results from the univariate Treescan analyses were tested in a multivariate one-way MANOVA model where each branch is jointly associated with triglyceride, HDL, VLDL, and LDL levels. Total cholesterol is excluded because it is a composite value of the other three. Significance levels will be derived in a similar fashion using the parametric p-value from the F transformation of the Wilk’s statistic. A partial Wilk's test can be used to test the effects of individual dependent (phenotypes) or independent variables while controlling for all the other variables in the model. The partial Wilk's test is a reduced *versus* full model approach. This conditional Wilk's statistic is calculated as follows [30]:

$$\Lambda(y_g \,|\, y_1, \ldots, y_{g-1}, y_{g+1}, \ldots, y_p) = \frac{\Lambda_p}{\Lambda_{p-1}} \tag{1}$$

where p is the number of phenotypes (dependent variables), y_g is the phenotype of interest, Λ_p is the Wilk's statistic for the full model, and Λ_{p-1} is the Wilk's statistic for a reduced model where y_g is excluded. The resulting partial Wilk's statistic has an exact transformation to a partial F-statistic [30]:

$$F_{v_H, v_{E-p+1}} = \frac{1-\Lambda}{\Lambda} \frac{v_E - p + 1}{v_H} \tag{2}$$

where Λ is the result of equation 1, p is the number of phenotypes, $v_E = N - k$, $v_H = k - 1$, N = number of individuals, and k = the number of factor levels in the one-way MANOVA. The partial Wilk's test measures the contribution of a single phenotype to the genotypic association in the presence of the other phenotypes across all eigenvectors of the $\mathbf{E}^{-1}\mathbf{H}$ matrix. Univariate F tests and partial Wilk's tests are calculated for each significant MANOVA result emerging from the initial Treescan.

3.4. Bioinformatics, Site Prediction, and Selection Analyses

3.4.1. PolyPhen Analysis

Some nonsynonymous variants are benign and have little to no effect on protein function, while others can be extremely harmful and cripple the protein. A variant's effect on protein function can be predicted using a multiple protein alignment to assign each variant a score of how harmful the variant will be to protein function *via* the software PolyPhen [6] (e.g., a PolyPhen score of 0 = benign and a score of 4 = probably damaging). Given the 3D protein structure for ANGPTL4 is unknown, PolyPhen predictions were based predominantly on an alignment of the homologous sequences obtained through a Blast search of the NRDB database. The 45 unique SNPs were submitted to and retrieved from the PolyPhen web server using batch submission and retrieval scripts.

3.4.2. TreeSAAP Analysis

Another approach to identifying sites that are subject to adaptive change is by analyzing the changes in physiochemical properties when a substitution occurs [31]. Substitutions are determined by reconstructing ancestral states given a phylogenetic tree. Operating under the assumption of completely random amino acid replacement, we can calculate the expected distribution of amino acid substitutions. The substitutions inferred from the ancestral states are then compared to the expected distribution to determine the significance of the observed changes *via* the software package TreeSAAP [7]. The ancestral character states used by TreeSAAP are estimated using BaseML, which is part of the PAML software package [9]. We analyzed 31 different physiochemical properties, with 8 magnitude categories. Substitutions with changes of magnitude 6, 7, and 8 are considered to be radically changing [13] and are used in this paper to indicate significant variants.

3.4.3. Likelihood Selection Analysis

We used several likelihood-based methods to estimate the influence of selection on ANGPTL4. Likelihood methods use a codon-based model [32] of evolution to estimate the nonsynonymous to synonymous rate ratio (ω). A value of $\omega > 1$ is commonly thought to be an indicator of positive selection, $\omega = 1$ is neutral evolution, and $\omega < 1$ indicates purifying selection. We implemented the M8 model in PAML [9], which allows the nonsynonymous rate to vary among sites, while the synonymous rate is assumed to be homogeneous. The dual-rate model, implemented in HyPhy [10], allows both the nonsynonymous and synonymous rates to vary between sites, which has been shown to have greater power when compared to models where only the nonsynonymous rate is allowed to vary [10]. Both methods were used to infer sites under selection across the entire coding sequence.

Maximum likelihood was also used to estimate overall levels of selection in each of the protein domains and across the entire coding sequence. The coiled-coil and fibrinogen-like domains were separated, and ω was computed independently for each region. We used the one ratio (M0) model implemented in PAML [9], which assumes that ω is constant across all sites.

3.5. Comparison of TreeSAAP and PolyPhen

Of the 27 nonsynonymous variants analyzed, the functional polymorphisms from the biological assays [3,4] and the significant variants from Treescan analyses will be treated as "known" functional variants from which to evaluate the results of TreeSAAP, PolyPhen and their combination. We will compare PolyPhen, Strict PolyPhen (only "probably damaging"), TreeSAAP, Strict TreeSAAP (only category 8), PolyPhen and TreeSAAP, and Strict PolyPhen and Strict TreeSAAP. The remaining variants of the 27 will be defined as nonfunctional. This is conservative because only variants in the tails of triglyceride were biologically tested. A two-tailed Fisher's exact test was performed on a 2 by 2 table with the rows being the results of the method and the columns being the "known" information on the variants. If a method performs well, we would expect that it should have a higher ratio in the functional column than nonfunctional column. Biological assays were performed on eight variants (2 in the high tail and 6 in the low). All six low-tail variants were shown to have some type of functional effect. The two high-tail variants did not show any functional evidence. These two variants were classified separately from the other variants.

4. Conclusions

From our study, we had three different types of analyses: Genotype/phenotype association (Treescan), overall selection analyses (PAML M0), and three site prediction methods (REL, PolyPhen and TreeSAAP). Besides PolyPhen, each type of analysis used some form of phylogenetic data, and each gave us additional insight. First, the Treescan analysis provided evidence for an association with HDL in African Americans with the R278Q variant. Second, the PAML M0 analysis demonstrated the coiled domain is under positive selection while the fibrinogen-like domain is under slightly negative selection. It is of interest that most of the rare functional variants are within the fibrinogen-like domain. Finally, the "known" functional variants were leveraged such that we could evaluate the relative merits of site prediction from PolyPhen and TreeSAAP. We concluded that a combination of both methods is likely the best approach to take.

While no sequence analysis method is going to reveal everything about genotype/phenotype relations, we do have tools that can work together to give us greater insight and lead us towards productive paths. For association studies, sequence data can give us greater ability to estimate the phylogenetic relationships between haplotypes. This in turn leads to a greater context for which to direct and interpret statistical tests. For TreeSAAP, phylogenetic estimation allows for an empirical estimate of the distributions of different types of amino acid changes. From these distributions, we can make predictions about which particular changes are out of the ordinary and are more likely to have an impact on gene function and subsequently on the phenotypes that we are interested in.

In future studies, these site prediction methods will be a first step to provide greater statistical power and impetus to invest in biological follow up. These methods create *a priori* hypotheses to be tested leading to greater statistical power with the reduced number of tests to correct for. These methods may also suggest the biological nature of sites predicted to have functional consequence. Many labs are currently embarking on whole exome sequencing. These methods will be useful as we try to comb through this mass of data to separate the functional from nonfunctional variants.

Acknowledgements

We would like to thank Jonathan C. Cohen, Len A. Pennacchio, and Eric Boerwinkle for access to the data, discussions, and comments on this manuscript. This work was partially supported by Brigham Young University Office of Research and Creative Activities. We are thankful for the opportunity to contribute an article in a volume edited by our colleague, mentor, and friend, Alan Templeton.

References and Notes

1. Templeton, A.R. *Population Genetics and Microevolutionary Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2006.
2. Page, R.D.M.; Holmes E.C. *Molecular Evolution: A Phylogenetic Approach*; Blackwell Science Ltd.: Osney Mead, Oxford, UK, 1998.
3. Romeo, S.; Pennancchio, L.A.; Fu, Y.-X.; Boerwinkle, E.; Tybjaerg-Hansen, A.; Hobbs, H.H.; Cohen, J.C. Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nat. Genet.* **2007**, *39*, 513–516.
4. Romeo, S.; Yin, W.; Kozlitina, J.; Pennacchio, L.A.; Boerwinkle, E.; Hobbs, H.H.; Cohen, J.C. Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. *J. Clin. Invest.* **2009**, *119*, 70–79.
5. Victor, R.G.; Haley, R.W.; Willett, D.L.; Peshock, M.D.; Vaith, P.C.; Leonard, D.; Basit, M.; Cooper, R.S.; Iannacchione, V.G.; Visscher, W.A.; Staab, J.M.; Hobbs, H.H.; Dallas Heart Study Investigators. The Dallas heart study: A population-based probability sample for the multidisciplinary study of ethnic differences in cardiovascular health. *Am. J. Cardiol.* **2004**, *93*, 1473–1480.
6. Ramensky, V.; Bork P.; Sunyaev S. Human non-synonymous SNPs: Server and survey. *Nucleic Acids Res.* **2002**, 30, 3894–4900.
7. Woolley, S.; Johnson, J.; Smith, M.J.; Crandall, K.A.; McClellan, D.A. TreeSAAP: Selection on amino acid properties using phylogenetic trees. *Bioinformatics* **2003**, *19*, 671–672.
8. Templeton, A.R.; Maxwell, T.; Posada, D.; Stengård, J.H.; Boerwinkle, E.; Sing, C.F. Tree scanning: A method for using haplotype trees in phenotype/genotype association studies. *Genetics* **2005**, *169*, 441–453.
9. Yang, Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Bio. Evol.* **2007**, *24*, 1586–1591.
10. Kosakovsky Pond, S.L.; Frost, S.D.; Muse, S.V. HyPhy: Hypothesis testing using phylogenies. *Bioinformatics* **2005**, *21*, 676–679.
11. Templeton, A.R.; Crandall, K.A.; Sing, C.F. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics* **1992**, *132*, 619–633.
12. Talmud, P.J.; Smart, M.; Presswood, E.; Cooper, J.A.; Nicaud, V.; Drenos, F.; Palmen, J.; Marmot, M.G.; Boekholdt, S.M.; Wareham, N.J.; Khaw, K.; Kumari, M.; Humphries, S.E.; On behalf of the EARSII Consortium and the HIFMECH Consortium. ANGPTL4 E40K and T266M: Effects on Plasma Triglyceride and HDL Levels; Postprandial Responses; and CHD Risk. *Arterioscler. Thromb. Vasc. Biol.* **2008**, *28*, 2319–2325.

13. McClellan, D.A.; Palfreyman, E.J.; Smith, M.J.; Moss, J.L.; Christensen, R.G., Sailsbery, J.K. Physicochemical evolution and molecular adaption of the cetacean and artiodactyls cytochrome b proteins. *Mol. Bio. Evol.* **2005**, *22*, 437–455.
14. Pérez-Losada M.; Viscidi R.P.; Demma J.C.; Zenilman J.; Crandall K.A. Population genetics of Neisseria gonorrhoeae in a high-prevalence community using a hypervariable outer membrane porB and 13 slowly evolving housekeeping genes. *Mol. Biol. Evol.* **2005**, *22*, 1887–1902.
15. Crandall, K.A.; Kelsey, C.R.; Imamichi, H.; Lane, H.C.; Salzman, N.P. Parallel evolution of drug resistance in HIV: Failure of nonsynonymous/synonymous substitution rate ratio to detect selection. *Mol. Bio. Evol.* **1999**, *16*, 372–382.
16. Tartaglia, M.; Pennacchio, L.A.; Zhao, C.; Yadav, K.K.; Fodale, V.; Sarkozy, A.; Pandit, B.; Oishi, K.; Martinelli, S.; Schackwitz, W.; Ustaszewska, A.; Martin, J.; Bristow, J.; Carta, C.; Lepri, F.; Neri, C.; Vasta, I.; Gibson, K.; Curry, C.J.; Siguero, J.P.L.; Digilio, M.C.; Zampino, G.; Dallapiccola, B.; Bar-Sagi, D.; Gelb, B.D. Gain-of-function SOS1 mutations cause a distinctive form of Noonan syndrome. *Nat. Genet.* **2007**, *39*, 75–79.
17. Stephens, M.; Donnelly, P. Inference in molecular population genetics. *J. R. Stat. Soc. Ser. B* **2000**, *62*, 605–655.
18. Stephens, M.; Donnelly, P. A comparison of Bayesian methods for haplotypes reconstruction. *Am. J. Hum. Genet.* **2003**, *73*, 1162–1169.
19. Clement, M.; Posada, D.; Crandall, K.A. TCS: A computer program to estimate gene genealogies. *Mol. Ecol.* **2000**, *9*, 1657–1659.
20. Crandall, K.A.; Templeton A.R. Statistical methods for detecting recombination. In *The Evolution of HIV*; Crandall, K.A., Ed.; The Johns Hopkins University Press: Baltimore, MD, USA, 1999; pp. 153–176.
21. Castelloe J.; Templeton A.R. Root probabilities for intraspecific gene trees under neutral coalescent theory. *Mol. Phylogenet. Evol.* **1994**, *3*, 102–113.
22. Crandall, K.A.; Templeton, A.R. Empirical tests of some predictions from coalescent theory with applications to intraspecific phylogeny reconstruction. *Genetics* **1993** *134*, 959–969.
23. Swofford, D.L. PAUP*. *Phylogenetic Analysis Using Parsimony (*and Other Methods)*, Version 4; Sinauer Associates: Sunderland, MA, USA, 2002.
24. Posada, D.; Crandall, K.A. Modeltest: Testing the model of DNA substitution. *Bioinformatics* **1998**, *14*, 817–818.
25. Hasegawa, M.; Kishino, H.; Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **1985**, *22*, 160–174.
26. Guindon, S.; Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **2003**, *52*, 696-704.
27. Nowotny, P.; Hinrichs, A.L.; Smemo, S.; Kauwe, J.S.K.; Maxwell, T.; Holmans, P.; Hamshere, M.; Turic, D.; Jehu, L.; Hollingsworth, P.; Moore, L.; Bryden, P.; Myers, A.; Doil, L.M; Tacey, K.M.; Gibson, A.M.; McKeith, I.G.; Perry, R.H.; Morris, C.M.; Thal, L.; Morris, J.C.; O'Donovan, M.C.; Lovestone, S.; Grupe, A.; Hardy, J.; Owen, M.J.; Williams, J.; Goate, A. Association studies between risk for late-onset alzheimer's disease (LOAD) and variants in Insulin Degrading Enzyme. *Am. J. Med. Genet. B* **2005**, *136B*, 62–68.

28. Grupe, A.; Li, Y.; Rowland, C.; Nowotny, P.; Hinrichs, A.L.; Smemo, S.; Kauwe, J.S.K.; Maxwell, T.J.; Cherny, S.; Doil, L.; Tacey, K.; van Luchene, R.; Myers, A.; Vriexe, F.W.; Kaleem, M.; Hollingworth, P.; Jehu, L.; Foy, C.; Archer, N.; Hamilton, G.; Homans, P.; Morris, C.M.; Catanese, J.; Sninsky, J.; White, T.J.; Powell, J.; Hardy, J.; O'Donovan, M.; Lovestone, S.; Jones, L.; Morris, J.C.; Thal, L.; Owen, M.; Williams, J.; Goate, A. A scan of chromosome 10 identifies a novel locus showing strong association with Late-Onset alzheimer disease. *Am. J. Hum. Genet.* **2006**, *78*, 78–88.
29. Westfall, P.; Young, S.S. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustments*; Wiley-Interscience: New York, NY, USA 1993.
30. Rencher, A.C. *Methods of Multivariate Analysis*; Wiley: New York, NY, USA, 1995; p. 316.
31. McClellan, D.A.; McCracken, K.G. Estimating the influence of selection on the variable amino acid sites of the cytochrome B protein functional domains. *Mol. Biol. Evol.* **2001**, 18, 917–925.
32. Goldman, N.; Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **1994**, *11*, 725–736.

Additional References

T. J. Maxwell, M. L. Bendall, J. Staples, T. Jarvis, and K. A. Crandall. Phylogenetics applied to genotype/phenotype association and selection analyses with sequence data from angptl4 in humans. *International journal of molecular sciences*, 11(1):370–85, Jan. 2010.

S. L. K. Pond, S. D. W. Frost, and S. V. Muse. HyPhy: hypothesis testing using phylogenies. *Bioinformatics (Oxford, England)*, 21(5):676–9, Mar. 2005.

V. Ramensky, P. Bork, and S. Sunyaev. Human non-synonymous SNPs: server and survey. *Nucleic Acids Research*, 30(17):3894–900, Sept. 2002.

S. Romeo, L. A. Pennacchio, Y. Fu, E. Boerwinkle, A. Tybjaerg-Hansen, H. H. Hobbs, and J. C. Cohen. Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nature genetics*, 39(4):513–6, Apr. 2007.

S. Romeo, W. Yin, J. Kozlitina, L. A. Pennacchio, E. Boerwinkle, H. H. Hobbs, and J. C. Cohen. Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. *The Journal of clinical investigation*, 119(1):70–9, Jan. 2009.

M. C. Smart-Halajko, A. Kelley-Hedgepeth, M. C. Montefusco, J. A. Cooper, A. Kopin, J. M. McCaffrey, A. Balasubramanyam, H. J. Pownall, D. M. Nathan, I. Peter, P. J. Talmud, and G. S. Huggins. ANGPTL4 variants E40K and T266M are associated with lower fasting triglyceride levels in Non-Hispanic White Americans from the Look AHEAD Clinical Trial. *BMC medical genetics*, 12:89, Jan. 2011.

R. G. Victor, R. W. Haley, D. L. Willett, R. M. Peshock, P. C. Vaeth, D. Leonard, M. Basit, R. S. Cooper, V. G. Iannacchione, W. A. Visscher, J. M. Staab, and H. H. Hobbs. The Dallas Heart Study: a population-based probability sample for the multidisciplinary study of ethnic differences in cardiovascular health. *The American journal of cardiology*, 93(12): 1473–80, June 2004.

S. Woolley, J. Johnson, M. J. Smith, K. A. Crandall, and D. A. McClellan. TreeSAAP: Selection on Amino Acid Properties using phylogenetic trees. *Bioinformatics*, 19(5):671–672, Mar. 2003.

Z. Yang. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8):1586–91, Aug. 2007.

Chapter 3

Assessing the performance of computational methods for identifying positively selected sites

3.1 Background

Identifying the selective pressures acting on specific amino acid sites is essential for understanding adaptive evolutionary process (Stewart et al. 1987). For protein-coding DNA sequences, the fundamental measure of selective pressure at the protein level is the ratio of nonsynonymous and synonymous substitution rates, $d_N/d_S = \omega$. Assuming selectively neutral evolution, the synonymous and nonsynonymous substitution rates are equal, and $\omega = 1$. When selective pressure is present at the amino acid level, the acceptance rate of nonsynonymous mutations is influenced by these constraints, while the fixation of synonymous mutations is largely unaffected. Under the current paradigm, estimates of $\omega > 1$ indicate that the nonsynonymous substitution rate is greater than the synonymous rate, providing strong evidence for positive Darwinian selection (Sharp 1997).

For most proteins, the majority of sites are highly conserved, while only a small proportion of sites may be affected by positive selection. Several site-prediction methods have been proposed for identifying specific sites under adaptive evolution. These methods account for heterogeneity in selective pressure among sites, allowing some sites to have $\omega > 1$, while other sites are under purifying selection. Statistical tests are used to determine whether estimates of ω are significantly different than the neutral expectation.

Recently, there has been much debate regarding the performance and relative merit of various site-prediction methods (Yokoyama et al. 2008; Nozawa et al. 2009a; Yang et al. 2009; Nozawa et al. 2009b; Yang and dos Reis 2011). Statistical properties of site-prediction methods, such as sensitivity and type I error rates, have been studied extensively using computer simulation (Suzuki and Nei 2002; Anisimova et al. 2001, 2002; Wong et al. 2004;

Kosakovsky Pond and Frost 2005b; Nozawa et al. 2009a). However, the true concern for many biologists is not the relative performance of various approaches in ideal or contrived circumstances, but whether site-prediction methods can correctly identify adaptive sites *in real data*. This proves to be a difficult question, since adaptive sites are typically unknown in real data. A handful of studies have sought to evaluate performance with experimentally determined adaptive sites (Sawyer et al. 2005; Yokoyama et al. 2008). In these studies, only a few adaptive sites could be confirmed experimentally, and findings may be specific to the evolutionary history of the gene studied. Although experimental studies are invaluable for identifying positive selection, the narrow applications of such techniques limit broader conclusions about site-prediction methods.

The emergence of drug resistant HIV presents an excellent system for examining the process of adaptive evolution (Crandall et al. 1999; Shafer et al. 1999). Several characteristics of intrapatient HIV populations are well suited for evaluating site-prediction methods. First, HIV evolution occurs within an observable timeframe due to its short generation time and high *in vivo* mutation rates. Second, we can control (to some extent) for variation in selective pressure by sampling repeatedly from the same patient over time. Also, we can often identify when changes in selective pressure occur, e.g. at the start of drug therapy. Third, specific mutations responsible for drug resistance are known through confirming *in vitro* assay (Petropoulos et al. 2000; Zhang et al. 2005) and are well documented in the published literature (Shafer et al. 2007). Finally, the availability of multiple HIV-1 sequences collected from several different patients provides a variety of evolutionary histories for analysis; studies based on such data are more widely applicable than studies based on a single history.

In this paper, we evaluate the performance of eight site-prediction methods using empirical data with known positively selected sites. We analyze 109 intrapatient HIV-1 datasets collected during phase I and II clinical studies of antiretroviral drug therapy (Condra et al. 1996; Zhang et al. 1997; Vaillancourt et al. 1999; Bacheler et al. 2000). Site-prediction methods are used to identify assumed adaptive sites, which are compared to sites known to be under positive Darwinian selection. We consider the sensitivity of each method and examine similarities and differences among methods. We also investigate various properties that may affect the performance of site-prediction methods.

3.2 Results and Discussion

Power in empirical data

We analyzed 3804 HIV-1 *pol* nucleotide sequences isolated from 109 subjects; a total of 31,172 amino acid (codon) sites were examined for the presence of positive selection. Among these sites, 49 contained missing or ambiguous data, 4657 sites had only synonymous differences, and 5165 sites were found to have nonsynonymous differences. We identified 971 nonsynonymous sites that have differences known to decrease drug susceptibility. We classify these differences as drug resistant mutations (DRMs) and the sites are considered to be evolving under positive selection. The remaining 4194 nonsynonymous sites have unknown selective pressure; however, there is evidence that they have little effect on drug susceptibility and may occur frequently in untreated persons. For the purposes of this study, we refer to these sites as other nonsynonymous mutations (ONMs). DRMs were identified using the Sierra webservice provided by the HIV Drug Resistance Database (HIVdb: http://hivdb.stanford.edu/), algorithm version 6.0.11F (Liu and Shafer 2006).

All sites were examined for the presence of positive selection using each of eight methods using nominal cutoff values (CV) (table 1). CVs were chosen based on default settings; this is meant to reflect the typical user experience with these software packages. SAAP is found to be the most powerful method, correctly identifying 369 out of 971 sites. NSR and DUAL also exhibit relatively high power, with 226 and 184 correct classifications, respectively. The five remaining methods have relatively low power using the default CVs, each identifying fewer than 10% of the positively selected sites. We also examined the number of ONMs identified as positively selected by each method. Since the actual selective pressure acting on these sites is unknown, it is possible that a site may indeed be under positive selection, unrelated to its classification with respect to drug resistance. Thus, we cannot conclude that a significant result at one of these sites is a “false positive”. However, the ONM detection rate does provide an upper bound on the false positive rate. SG and SLAC have ONM detection rates less than 1%, and FEL has an ONM detection rate of 2.2%. These values are much lower than the expected type I error rates, indicating that the tests are conservative for the given significance level. SAAP, on the other hand, classified 1719 out of 4194 (41%) ONM sites as positively selected. Although these cannot be considered to be false positives, this suggests that the type I error rate may exceed the nominal significance level of the test. For the Bayesian methods, the expected type I error rate is not given by the cutoff $P_b = 0.95$ or $BF = 50$. However, the number of ONMs identified by M2a (51) and M8 (79) suggest that these tests may be conservative at this cutoff. NSR and DUAL have ONM detection rates of 10.6% and 12.4%, respectively. Using the reciprocal of the Bayes factor as an approximate measure of the false positive rate (as suggested by Kosakovsky Pond and Frost (2005b)), these ONM detection rates could indicate poor specificity at the default CV ($1/BF = 0.02$).

Given the results at the default CVs, we attempted to improve the performance of each test using alternate CVs. The CVs for SG, SLAC, FEL, M2a and M8, are relaxed to increase power. The CVs for SAAP, NSR, and DUAL are chosen to increase the specificity of the test. The alternate CVs were chosen from published literature (table 1). DRMs identified by SG, M2a and M8 increased by approximately 50%, while DRMs identified by SLAC and FEL increased by nearly 300%. Note that the CVs for SG, M2a and M8 were only relaxed 5%, while the CVs for SLAC and FEL were relaxed 20%. ONMs identified by SAAP decreased by 39%, and ONMs identified by NSR and DUAL decreased by 14% and 37%, respectively.

Table 1: **DRMs and ONMs identified by each method.** The number of DRMs and ONMs identified are shown. The proportion of sites identified to all DRMs (971) or ONMs (4194) is in parentheses. The default and alternate CVs used for each method appears in columns 6 and 11, respectively.

	Default CV[1]			Alternate CV[2]		
	DRM	ONM	CV[3]	DRM	ONM	CV
SG	49 (0.050)	21 (0.005)	0.05	75 (0.077)	35 (0.008)	0.10
SLAC	18 (0.019)	21 (0.005)	0.10	72 (0.074)	96 (0.023)	0.25
SAAP	369 (0.380)	1719 (0.410)	0.05	215 (0.221)	1056 (0.252)	0.001
FEL	61 (0.063)	91 (0.022)	0.10	220 (0.227)	452 (0.108)	0.25
M2a	41 (0.042)	51 (0.012)	0.95	67 (0.069)	73 (0.017)	0.90
M8	76 (0.078)	79 (0.019)	0.95	104 (0.107)	122 (0.029)	0.90
NSR	226 (0.233)	444 (0.106)	50	201 (0.207)	381 (0.091)	100
DUAL	184 (0.189)	521 (0.124)	50	137 (0.141)	329 (0.078)	100

[1] The default CVs used for SLAC, FEL, NSR and DUAL are implemented in the Datamonkey webservice (Kosakovsky Pond and Frost 2005a).

[2] Alternate CVs for SLAC and FEL are based on reccomendation by Kosakovsky Pond and Frost (2005b). Alternate CV for SAAP was chosen based on observations by Porter et al. (2007).

[3] For frequentist hypothesis tests, the cutoff value (CV) corresponds to the significance level α of the test; Bayesian tests use a posterior probability cutoff P_b (M2a and M8) or a minimum Bayes factor (NSR and DUAL) as evidence supporting the hypothesis.

Random-sites model selection

Under random-sites models, the distribution of nonsynonymous and synonymous rates can be tested using the likelihood ratio (Anisimova et al. 2001) and the AIC. Yang et al. (2005) recommend using the likelihood ratio to test a null model, where $\omega \leq 1$ for all sites, against a model that allows a proportion of sites with $\omega > 1$, e.g., M1a vs. M2a, M7 vs. M7. If neutrality is rejected, empirical Bayes can be applied to identify positively selected sites. A similar test can be used to test for synonymous rate variation. Kosakovsky Pond and Muse (2005) suggest using the LRT and AIC to test NSR against DUAL. Although this is not a test for selection, it could indicate whether the assumption of a constant synonymous rate is violated.

We used the LRT and AIC to test for the presence of positively selected sites with random-sites models M2a and M8. Testing model M1a against M2a, the neutral hypothesis was rejected in 43 cases using the LRT ($\alpha = 0.01$, $d.f. = 2$). AIC scores favored M2 in 55 cases. M8 was selected in 36 datasets by LRT and 52 datasets using AIC. We recalculated the number of DRMs and ONMs identified by each method. Surprisingly, despite discarding over half of the datasets considered, the overall counts were not largely affected (table 2).

The presence of synonymous rate variation was also tested using the LRT and AIC. NSR was rejected in 83 datasets using the LRT ($\alpha = 0.01$, $d.f. = 4$) and 91 datasets by AIC. To test for variable rates, we used a hierarchical approach to choose between the Constant rates model (Kosakovsky Pond and Muse 2005), NSR and DUAL. The distribution of the test statistic for each set of hypotheses is assumed to be χ^2_4, and we test each hypothesis at $\alpha = 0.005$ to achieve an overall significance of $\alpha = 0.01$. The LRT failed to reject the Constant rates model in 41 datasets. In the remaining datasets, the LRT failed to reject NSR in 8 cases; DUAL was chosen for 68 datasets. Using the AIC, DUAL was chosen in 89 datasets, NSR was chosen six times, and Constant rates was in 14 datasets. We used these results to recalculate the number of DRMs and ONMs identified. Using results from testing NSR against DUAL, the N+D method counts only those sites identified by the selected model. Similarly, C+N+D counts sites when NSR or DUAL are chosen, but does not count any sites when the Constant rates model is chosen. The use of such tests greatly decreases the number of DRMs and ONMs identified. The ONM detection rate for C+N+D drops to 10.4%, while still recovering 16.2% of DRMs (table 2).

Agreement among methods

We determined the set of sites identified by each method at the default CVs, and applied simple set operations to determine whether there is agreement among methods. Of the 971 DRMs known to be under positive selection, 544 sites (56.0%) were identified by at least one method. 2127 out of 4194 ONMs were identified (50.7%). None of the sites were identified by all eight methods. We found that particular methods tend to agree more often; in some cases, all sites identified by a particular method are also detected using another method. All

Table 2: **Model selection using random-sites models.** The number of sites identified using the default CVs after using the LRT (left) and AIC (right) are shown. The number of sites eliminated is shown in parentheses. For M2a and M8, the number eliminated is equal to the number identified before the LRT or AIC (table 1 minus the number identified after the tests are applied. For N+D and C+N+D, the number eliminated is equal to the number of sites identified by NSR or DUAL before the tests are applied (see figure 1A) minus the post-test counts. The significance level of the LRT is $\alpha = 0.01$; C+N+R use $\alpha = 0.05$ for each test.

	LRT ($\alpha = 0.01$)		AIC	
	DRM (Δ)	ONM(Δ)	DRM(Δ)	ONM(Δ)
M2a	40 (1)	40 (11)	40 (1)	44 (7)
M8	63 (13)	64 (15)	67 (9)	68 (11)
N+D	189 (93)	536 (206)	186 (96)	526 (216)
C+N+D	158 (124)	437 (305)	182 (100)	518 (224)

41 DRMs chosen by M2a were also detected by M8, and 75 out of 76 DRMs found by M8 were also chosen by NSR. The 18 DRMs found by SLAC are a subset of the DRMs detected by FEL. We also examined whether particular methods tend to identify sites that are unique relative to other methods. SAAP discovered 246 DRMs that were not detected by other methods, which is nearly half of all the DRMs identified. NSR and DUAL detected 41 and 33 unique DRMs respectively. The remaining methods found fewer than 10 unique DRMs. We note that SAAP also found 1367 unique ONMs, nearly 65% of all ONMs identified; NSR found 104, and DUAL found 168. Impressively, 427 DRMs (44.0%) remained unidentified by any method suggesting there is significant room for improvement on methods to detect positive selection in protein coding genes.

We used Euler diagrams to depict the relationships among methods (fig. 1 and 2). Specifically, we are interested in whether methods agree among random-sites models, among independent-sites models, and between independent- and random-sites models. Since sites identified by SAAP did not appear similar to other independent-sites methods (fig. 1B), SAAP is considered separately from the other methods. Independent-sites methods detected 85 DRMs and 99 ONMs; random-sites methods identified 283 DRMs and 743 ONMs. Nearly

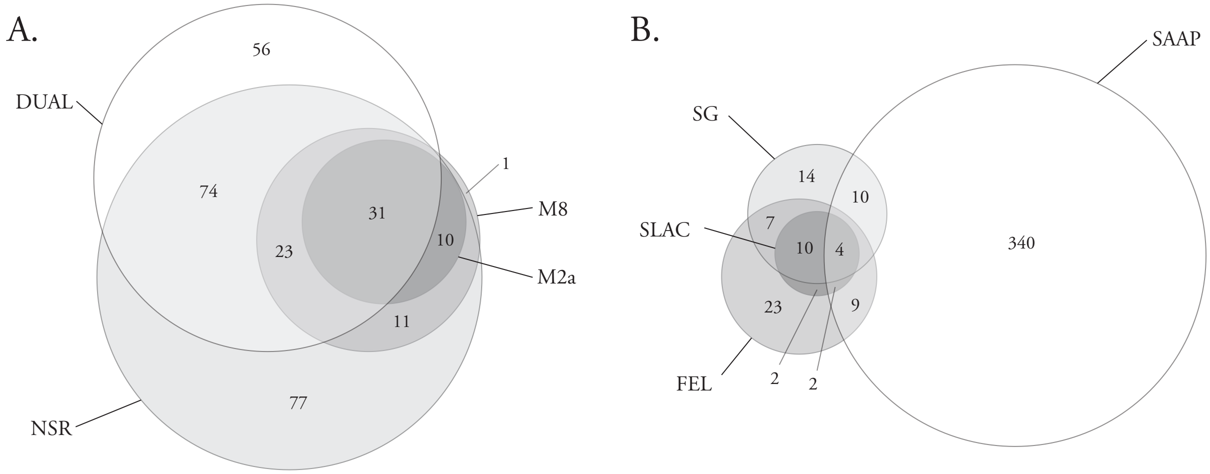


Figure 1: **Euler diagrams comparing eight site-prediction methods.** Figure A compares DRMs identified using the four random-sites models. Figure B shows DRMs predicted using independent-sites approaches. For each figure, the number of sites identified is labeled in the proper region. Circles within each figure are proportional to the number of sites identified; however, the intersection areas are approximate. Efforts were made to make the intersection areas proportional to the the number of sites represented by the intersection. Note that circles between figures A and B are not drawn to scale.

all sites found by independent-sites methods were also found using random-sites models; 15 DRMs and 17 ONMs were identified by independent-sites that were not detected by random sites. Agreement with SAAP is consistent ($\approx 40\%$) regardless of the method or site category considered.

Influence of data properties on method performance

Site-specific strength of selection Ideally, the performance of site-prediction methods should be tightly correlated with the actual strength of selective pressure at a site. In real data, the actual strength is unknown and impossible to quantify; however, the extensive literature correlating genotypic and phenotypic changes in drug-resistant HIV-1 offers some insight into the magnitude of these pressures. To investigate how site-specific selective pressures may influence the performance of these methods, we categorized the set of known DRMs according to location and substitution type. The set of 971 known DRMs included 141 drug-resistant

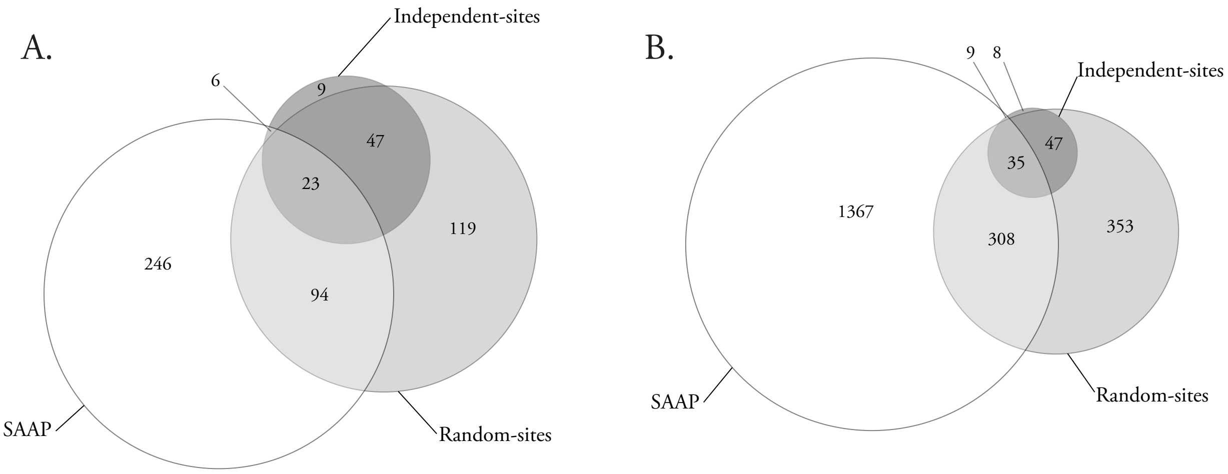


Figure 2: **Euler diagrams comparing independent- and random-sites approaches.** Figure A compares DRMs identified using SAAP, independent-sites, and random-sites approaches. Figure B shows ONMs predicted using the three approaches. The labels, circle sizes and intersections are the same as in figure 1. Note that circles between figures are not drawn to scale.

genotypes at 52 positions. For each genotype, we determined the total number of datasets where the genotype was found and the number of times this genotype is correctly identified to be under positive selection by one or more methods. By cross-referencing this information with known amino acid properties and protein positions, we can easily summarize the rate of positive identification for a given genotype, site, residue, or property.

The performance of site-prediction methods varied widely among positions. Table 3 compares the proportion of DRMs identified at various positions along the protein; positions with fewer than 10 observations were excluded. Site P-30 was found to be the most poorly predicted position, with only 2 of the 19 drug-resistant substitutions identified (10.5%). On the other hand, all 37 drug-resistant substitutions observed at site R-225 were detected by one or more methods. Performance was also shown to vary widely among particular "target" amino acids. Isoleucine, the most frequently observed drug-resistant genotype, was detected 170 times out of 181 possible instances (93.9%). This success, however, was not typical; other frequently observed target residues, such as valine (59.8% out of 169) and asparagine (44.4% out of 117), were much more difficult to detect. We also considered the amino acid properties involved in the drug-resistant substitution. Performance also varied widely among different categories for target residue polarity, acid-base properties, and chemical distance between initial and target residues. We did not observe any clear pattern indicating which properties were important for successful site-prediction.

Finally, we considered performance based on the magnitude of change in drug susceptibility. Drug-resistant genotypes were classified as “major” or “minor” DRMs based on the criteria used by the HIVdb algorithm version 6.0.11F (Liu and Shafer 2006). Positions shown in table 3 are categorized according to the types of DRMs found at each position. We identified four “major-only” positions (where only major DRMs were found), nine “minor-only” positions, and 19 “major+minor” positions (where both major and minor DRMs were found). We found that 82 out of 148 instances (55.4%) were correctly identified at the major-only positions, while 197 out of 263 (74.9%) were correctly identified at the minor-only positions. 33.9% of the minor and 64.8% of the major DRMs were detected at the major+minor positions (65/192 and 245/378, respectively). Across all 32 positions, 62.2% of the major and 57.6% of the minor DRMs were correctly identified by at least one method.

Dataset properties and taxon sampling We are interested in determining which aspects of a dataset have the most significant impact on the performance of site-prediction methods. We examined various properties of each dataset, including number of sequences, timepoints sampled, length of study, genetic diversity, mutation rate, summary statistics (D, D^* and F^*), and sequence divergence. Number of sequences and sequence divergence (measured by the tree length) were found to correlate well with the ability to detect selection. Mutation rate (μ) and D^* also appeared to correlate with detection; however, this correlation was not significant when the number of sequences is taken into account. For five of the methods, the percentage of DRMs identified increased as we increased the number of sequences. SG, SLAC, and FEL show steep increase in power when dataset size exceeds 60. Power for M2a and M8 increases sharply with 25 to 30 sequences, then continues to increase gradually. The percentage of ONMs identified remains low for these methods. For SAAP, NSR, and DUAL,

Table 3: **Site-specific effects.** The number of correct identifications for any method at each drug-resistant site; sites with substitutions in fewer than 10 datasets are excluded. DRMs are classified as "major" or "minor" based on the HIVdb criteria (algorithm version 6.0.11F). The codon position according to the reference protease (P) or reverse transcriptase (R) gene is shown alongside the total proportion of DRMs identified. Columns *Major DRMs* and *Minor DRMs* list the residues associated with each class of DRM, the number of DRMs correctly identified, and the number of observed DRMs in the data.

Codon	%	Major DRMs		Minor DRMs	
Major-only positions					
P-90	0.615	M	16/26		
R-41	0.412	L	7/17		
R-103	0.518	N,S,T	43/83		
R-190	0.727	A,E,S,V	16/22		
Minor-only positions					
P-10	0.810			R,I,F,Y,V	47/58
P-24	1.000			I	24/24
P-35	0.818			G	9/11
P-53	0.231			L	3/13
P-71	0.547			I,T,V	29/53
P-73	0.273			C,S	3/11
R-108	1.000			I	35/35
R-219	0.476			R,N,Q,E	10/21
R-225	1.000			H	37/37
Major+minor positions					
P-30	0.105	N	0/1	E,G,V	2/18
P-32	0.812	I	8/8	A,E,G	5/8
P-46	0.761	I,L	46/54	R,T,V	5/13
P-47	0.742	V	17/19	L,M,T	6/12
P-50	0.524	V	8/10	N,M,T	3/11
P-54	0.703	A,L,T,V	24/35	N	2/2
P-76	0.357	V	3/7	S	2/7
P-82	0.462	A,F,T	23/51	G	1/1
P-84	0.667	V	10/13	L,K,M,T	6/11
P-88	0.167	S	2/9	D,G,T	2/15
R-67	0.278	N	5/18	G,Y,V	5/18
R-69	0.360	D	4/6	A,N,E,I,S	5/19
R-70	0.654	R,E,G	15/24	A,Q	2/2
R-100	0.750	I	12/12	S	0/4
R-101	0.595	E,P	14/14	N,Q,H	8/23
R-181	0.333	C	4/4	H	0/8
R-184	0.423	I,V	20/45	A,L,T	2/7
R-188	0.591	C,H,L	12/21	P	1/1
R-215	0.667	F,Y	18/27	A,D,C,H,I,S	8/12

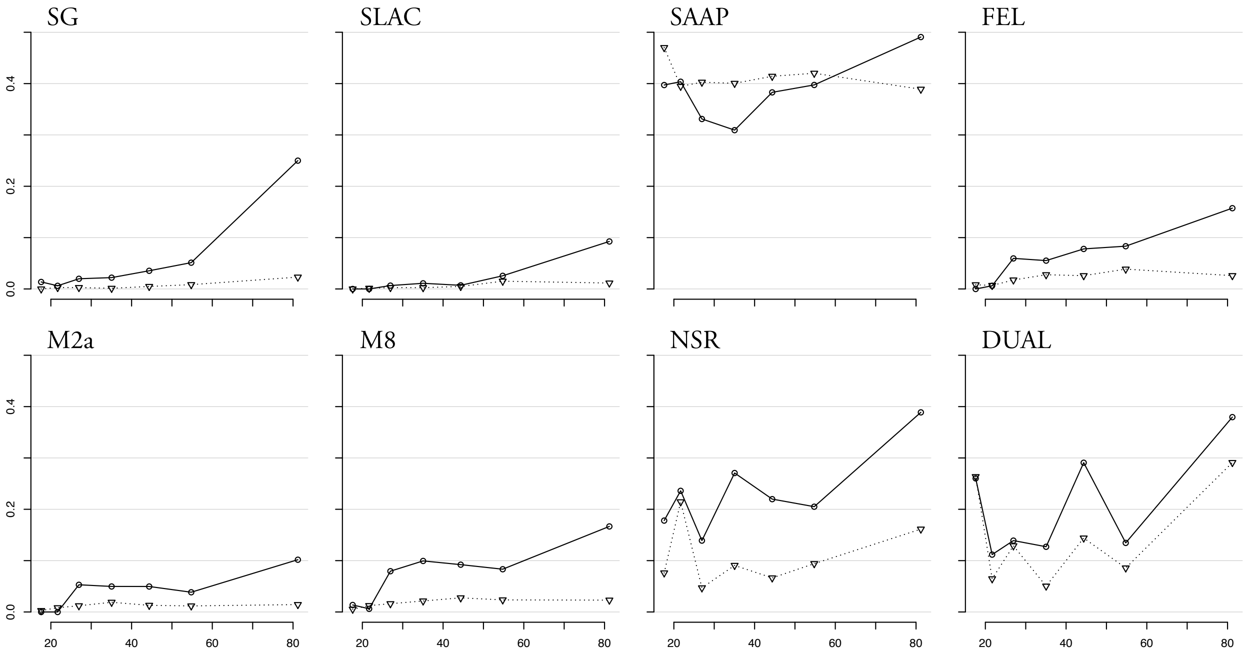


Figure 3: **Proportion of sites identified at variable numbers of sequences.** The y-axis represents the proportion of sites identified; the x-axis represents the number of sequences in the dataset. The proportion of DRMs identified (power) is shown with using circles and solid lines; ONMs have triangles and dotted lines. Datasets were placed into seven bins according the number of sequences used; the bins are 15-19, 20-24, 25-29, 30-39, 40-49, 50-59, and greater than 60. Each bin contained roughly the same number of datasets and total DRMs. The proportion of sites identified was calculated by dividing the number of sites identified at the default CV by the total number of DRMs or ONMs for that bin.

the number of sequences did not appear to correlate with power; however, all methods performed better with over 60 sequences (fig. 3).

We considered several different tree lengths to represent sequence divergence, including the initial Bayesian, synonymous distance, and maximum likelihood estimates. Only tree lengths from the M8 analysis appeared to correlate well with the DRM data. In general, the percentage of DRMs identified increased with increasing sequence divergence. SG, SLAC and FEL demonstrated a consistent increase in power with increasing divergence, although not as dramatic as the increase observed with increasing numbers of sequences. For random-sites approaches, power increased rapidly at a low threshold (tree length ≈ 0.5), then increased gradually with increasing divergence. Power for NSR, which did not have a clear relationship to the number of sequences, corresponds well with sequence divergence. Interestingly,

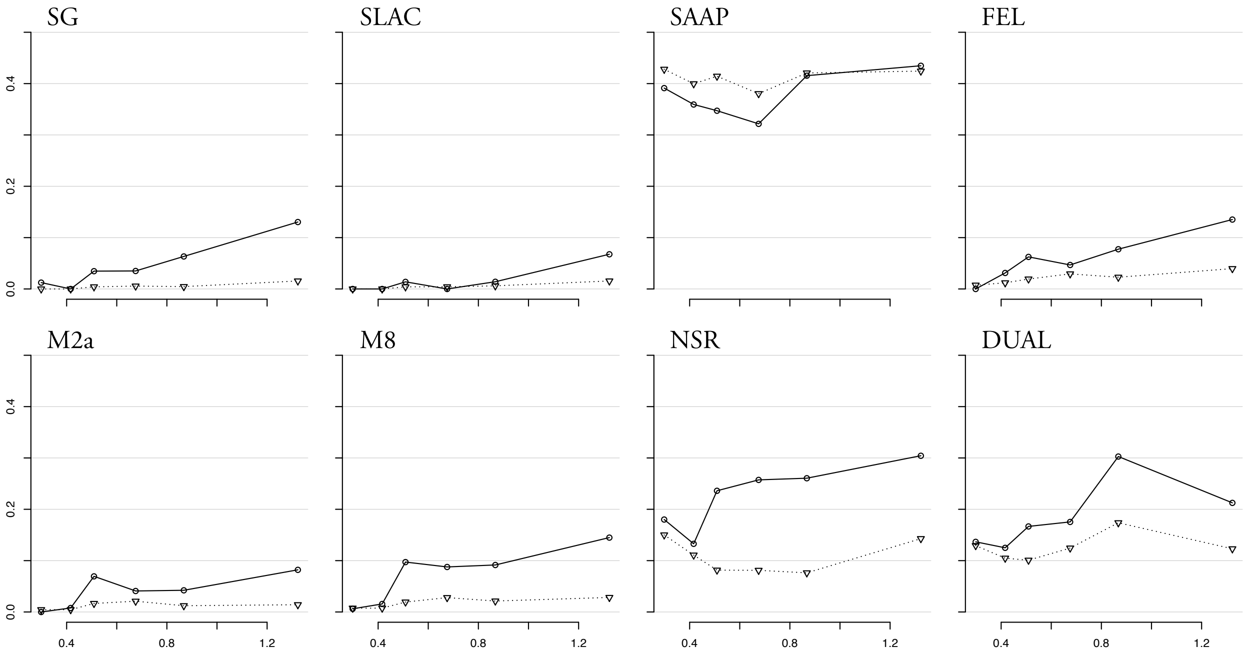


Figure 4: **Proportion of sites identified predicted by sequence divergence.** The y-axis represents the proportion of sites identified; the x-axis represents the sequence divergence as measured by the tree-length estimated by codon model M8. (The tree length is the sum of the branch lengths.) The proportion of DRMs identified (power) is shown with using circles and solid lines; ONMs have triangles and dotted lines. Datasets were placed into six bins according the overall tree length. The six bins were determined by diving all tree lengths into six equal percentiles. The five cutoffs the bins are 0.380, 0.455, 0.588, 0.748, and 0.952. Each bin contained roughly the same number of datasets and total DRMs. The proportion of sites identified was calculated by dividing the number of sites identified at the default CV by the total number of DRMs or ONMs for that bin.

the proportion of DRMs identified by DUAL appears to have a positive correlation with sequence divergence; however, power declines rapidly at high levels of divergence. SAAP did not appear to correlate with sequence divergence; DRM and ONM rates are constant around 40% (fig. 4).

3.3 Conclusions

Site-prediction methods exhibit low power for identifying selected sites in real data. This result is not completely unexpected; other empirical and experimental studies have shown that site prediction methods fail to identify adaptive sites, while giving strong support to sites which have no apparent adaptive function. If this is truly the case, one may question whether such methods should even be used. While the answer to this question ultimately lies with the investigator, we would suggest that some degree of skepticism should be maintained when interpreting results. More importantly, the assumptions of the method should be well understood. Site-prediction methods may be especially useful in instances when the investigator has prior knowledge about selective pressures acting on the protein, but failure to detect supported sites should not necessarily be taken as a rejection of the hypothesis.

Positive identification of unknown sites

All methods considered in this study identified positive selection on sites that presumably do not influence drug susceptibility. However, we do not consider these mutations to be false positives. In general, it is impossible to prove that a site is not under selection; we cannot know with certainty all the functional constraints at a particular site. Yokoyama et al.'s (2008) criticisms of random-sites methods are based on the assumption that mutations that do not influence light absorption are not positively selected. However, the possibility of adaptive evolution with respect to other crucial functions of these residues was not considered and have been suggested in other studies (Crandall and Hillis 1997). On the other hand, much is known about the function and constraints of each amino acid in the HIV-1 *pol* gene; over 100 three-dimensional structures have been determined and complete mutagenesis of the protease has been performed. Despite this published research, we are still hesitant to conclude that any particular mutation is *known* to have no adaptive purpose.

Although selection is unknown, the number of ONMs identified can give valuable clues into the relative performance of these various approaches. ONMs identified provide an upper bound for the false positive rate, and unusually high numbers of ONMs identified can suggest a lack of specificity. For instance, SAAP identified approximately 40% of all nonsynonymous sites, regardless of whether the site was known to be positively selected. We also found that this 40% trend was consistent despite differences in dataset size and tree length (fig. 3 and 4). The same performance could be achieved by randomly choosing 40% of the nonsynonymous sites. One explanation for this apparent lack of discrimination could be the null hypothesis of completely random amino acid replacement. This assumption clearly does not reflect current biological understanding. When given real biological scenarios, an unrealistic null hypothesis will be rejected much more frequently than a plausible one. Another explanation could be the choice and number of physicochemical properties tested. Any particular property may change radically, but have little or no effect on the fitness of the protein. Furthermore, by testing 31 properties, there are essentially 31 hypotheses being tested, which would require multiple test correction to achieve the desired significance. McClellan and Ellison (2010) performed an in-depth study addressing these issues.

NSR and DUAL also identify a relatively high proportion of ONMs. Unlike SAAP, the proportions of ONMs detected was noticeably different from the proportion of DRMs found; however, the number of ONMs identified was much higher than any other method. The naive empirical Bayes approach, which ignores uncertainty in the MLEs during Bayes estimation, may be responsible for this performance. Anisimova et al. (2002) discuss several difficulties with this approach, and the Bayes empirical Bayes approach has been proposed to address these shortcomings (Yang et al. 2005). Using the LRT and AIC to test for variable synonymous and nonsynonymous rates (C+N+D) reduces the overall number of sites detected, but does not provide a true test for the presence of sites under selection. Testing a neutral model against a model that allows adaptive sites may improve specificity.

Missing sites of known impact

It is somewhat disconcerting that 44.0% of the sites *known* to be under positive selection were not identified by *any* of the eight methods. This could be an indication that current methods are insufficient for describing the biological complexities of positive selection; the simplifying assumptions are *too* simple. Ideally, more realistic models which incorporate additional biological reality could be formulated; however, it is not entirely clear which aspects of molecular evolution are most critical to include.

In this study we examined the combined performance of all eight methods for each drug-resistant position found in the data (table 3). We found considerable variation in performance among locations on the protein, ranging from 10.5% to 100% of DRMs identified as adaptively evolving. Since each analysis was performed using 109 different datasets, we can conclude that such overwhelmingly successful (and unsuccessful) results are a reflection of site-specific properties shared in common among homologous proteins. Although we were unable to identify the characteristics that make it easy or difficult to detect selection for a given site, this result presents an interesting topic for future research.

Our results suggest that these methods are sensitive to differences in selective pressure among major and minor DRMs. Detecting minor DRMs at minor-only positions appears to be easier than identifying selection at positions with major DRMs, with nearly three-quarters (197 out of 263) of the DRMs detected. However, when major and minor DRMs are present at the same site, it seems that minor DRMs become very difficult to detect, while major DRMs are easier to detect. This outcome could be due to the effects of some sort of unbalanced, diversifying selection that selects against all drug-susceptible phenotypes and favors phenotypes which confer major drug-resistance. For many positions, this mode of selection would be consistent with the resistance tables compiled by the HIVdb.

We also found that the proportion of correctly identified DRMs may depend on the number of drug-resistant genotypes allowed at a position. According to table 3, selection detection is generally less successful at positions which have several drug-resistant genotypes available. For example, minor-only positions P-24, P-35, R-108, and R-225 each have only one possible drug-resistant genotype; every other genotype is drug-susceptible. Across these four sites, 105 out of 107 DRMs are correctly identified (98.1%). On the other hand, position P-82 has 7 major drug-resistant genotypes available with 11 additional residues classified as minor DRMs. Less than half (23 out of 51) of the major DRMs at this position were detected.

A similar pattern is observed at other positions with multiple drug-resistant residues, i.e., P-54, R-70, and R-219. It is interesting to note that the success rate for positions evolving under a directional mode of selection (those that have a single drug-resistant “target” genotype) is higher than for positions under diversifying selection, despite the fact that all site-prediction methods tested in this study assume diversifying selection.

Dataset properties for optimal performance

For many biologists, the most important question is whether site-prediction methods are suitable for analyzing their particular dataset. We find that the number of sequences and the amount of sequence divergence (tree length) are the best predictors of power for all site-prediction methods. Given only a few sequences or low divergence, we would not expect site-prediction to perform well, since the number of changes observed for each codon is low. However, as divergence increases, multiple changes can be observed at the same site, giving the test greater power. Sequence divergence is particularly important for independent-sites methods since inferences are made using only the data observed at that site (fig. 4). Random-sites models also demonstrate increasing power with higher levels of divergence, but relatively high power can be obtained even at moderate or low levels of divergence. Since random-sites approaches use the entire alignment to estimate the distribution of ω, these methods can gain power by combining information at different sites.

We found that DUAL looses power at high levels of divergence. One explanation could be due to relaxing the assumption of constant synonymous rates. Over short evolutionary distances, the nonsynonymous rate may be higher than the synonymous rate due to substitutions between neutral or similar amino acids. However, over longer evolutionary distances, synonymous substitutions tend to accumulate due to purifying selection. Thus, when sequence divergence is high, we expect to see many more synonymous substitutions. Since DUAL allows the synonymous rate to vary among sites, it may identify a class of sites with high synonymous rates that is invisible to methods that assume a constant synonymous rate. Further research would be necessary to investigate the sensitivity of DUAL to synonymous rate differences when divergence is high.

Another important property that will impact the performance of site-prediction methods is the presence and strength of selection. We do not consider this in the current study since all datasets used have one or more sites under strong selective pressure due to antiviral drug therapy. However, it is highly possible that many proteins operate exclusively under purifying selection or exhibit only weak positive selection.

Independent-sites or random-sites?

It does not appear that the debate regarding independent- and random-sites methods will be resolved in the near future, and we do not attempt to resolve it here. However, this study provides additional evidence that researchers can use to compare these approaches. Our observations indicate that the default CVs for both methods are conservative in most cases, and alternate CVs should be considered to increase power. However, statistical power is not free; there is always a trade-off between sensitivity and the frequency of type-I error.

The observation that both approaches tend to identify the same set of sites (fig. 2) lends additional credibility to all methods. The two approaches agree on 70 DRMs; this is 82% of DRMs identified by independent-sites methods and 25% of those detected by random-sites models. Thus, we feel confident that either method could detect the phylogenetic signal of adaptive evolution given an adequate sampling of the evolutionary history.

The relatively high sensitivity attained by random-sites models in small datasets make this approach appealing to many researchers, especially since sampling additional taxa can be costly or difficult. Despite this apparent statistical power, results from such analyses should be viewed somewhat skeptically in light of the number of ONMs identified in this study. On the other hand, the use of independent-sites in similar situations will lack power and may fail to identify any sites at all. Depending on the nature of the hypothesis and the desired outcomes, researchers should choose which test is most suitable. For example, Nielsen et al. (2005) and Bakewell et al. (2007) analyze thousands of human-chimpanzee orthologous genes using statistical methods based on the ω ratio. Although such a study is likely to produce some false-positive results, hundreds of new hypotheses were identified for future study. Other studies which attempt to make strong conclusions about specific sites may prefer a highly specific method. Yokoyama et al. (2008) and Nozawa et al. (2009a) study the adaptive evolution of vertebrate rhodopsins, which occurs only at particular sites along certain branches of the tree. Clearly, these authors prefer a test that achieves 100% specificity, even if it means that no sites are identified. (A test which never rejects the null hypothesis will always achieve 100% specificity.) Ultimately, it is up to the investigator to decide how much statistical power is needed and what level of false-detection can be tolerated. The best approach is to use a combination of approaches while considering the assumptions made by each approach.

Although the current d_N/d_S based approaches clearly have their merits, perhaps it is time to formulate new approaches to identify positive Darwinian selection. In many cases, there is good evidence regarding which sites might be under selection. Approaches based on Bayesian inference would allow this information to be incorporated into an analysis. MCMC algorithms could make inference under highly-complex models feasible. The statistical unification of phylogenetics and population genetics may lead to interesting statistical models that include mutation, selection and genetic drift (e.g. McVean and Vieira (2001), Yang and Nielsen (2008)). Finally, incorporating information about amino acid properties into models of adaptive evolution would establish the link between non-neutral evolution observed at the codon level and changes in phenotype and fitness (see Bruno (1996) and Wong et al. (2006) for examples). Further research is needed to develop better computational algorithms to implement such models. We hope our study provides some insights into where we might accomplish greater advances in detecting natural selection in nucleotide sequence data.

3.4 Methods

Site-prediction methods

We use eight site-prediction methods to identify adaptive sites. All methods identify specific sites under selection by allowing ω to vary among sites, yet they differ in their assumptions regarding the nature of this variation. Two general approaches are used for accommodating variation in the ω ratio. One approach is to estimate ω at every site, assuming a shared genealogy among all sites. We refer to this as the independent-sites approach, since the distribution of ω is assumed to be independent at each site. This is the approach taken by heuristic counting and site-wise likelihood ratio methods. A second approach is to estimate the distribution of ω among sites and use this distribution to estimate ω at each site a posteriori. Such methods are commonly referred to as random-sites models.

Heuristic counting methods Heuristic counting methods estimate d_N and d_S by inferring the number of synonymous (c_S) and nonsynonymous (c_N) substitutions which occurred across a phylogeny. Since the actual changes that occurred are unknown, the ancestral sequences are reconstructed and assumed to be known. For each codon, the number of nonsynonymous (s_N) and synonymous (s_S) sites is calculated by averaging sites over all branches. The probability of observing a substitution is considered to be proportional to s_N and s_S. Next, the observed number of substitutions c_S and c_N is inferred for each codon, treating ancestral states as known. Finally, a statistical test is applied to determine whether the inferred changes are significantly different than the expected number of changes under neutrality, $c_N/s_N = c_S/s_S$.

The counting framework has been extended to include information about the nature of inferred amino acid substitutions. For example, Hughes et al. (1990) differentiate between radical and conservative amino acid substitutions, as well as sites that undergo both types of changes along the tree. Xia and Li (1998) and McClellan and McCracken (2001) consider the magnitude of change with respect to specific physicochemical properties. Inferred substitutions between similar amino acids is considered to be purifying selection, while large changes may indicate positive selection. Such approaches incorporate information about changes in phenotype, thus are thought to provide insights into whether specific substitutions are adaptive or not.

We use three heuristic counting methods to identify selective sites (table 4). Suzuki and Gojobori's (1999) method (SG), implemented in ADAPTSITE v1.5 (Suzuki et al. 2001), reconstructs ancestral states using maximum parsimony along synonymous distance trees. The significance of the observed c_N is determined using a two-tailed binomial test. Single-likelihood ancestor counting (Kosakovsky Pond and Frost 2005b) (SLAC), implemented in HyPhy v2.0 (Kosakovsky Pond et al. 2005), uses maximum likelihood under a codon substitution model (Muse and Gaut 1994) to infer ancestral codons. Ambiguous reconstructions are resolved using the most frequent codon. The p-value is obtained using the extended binomial distribution. Selection on amino acid properties (McClellan and McCracken 2001) (SAAP) identifies selected sites based on radical changes in physicochemical properties. Ancestral states are inferred using a nucleotide substitution model and assumed to be known. The number of inferred changes c_N and expected changes p_N are calculated across the alignment, where p_N is the number of possible evolutionary pathways (Xia and Li 1998) assumed under completely random amino acid replacement. Inferred and expected changes are categorized according to their magnitude of change with respect to 31 amino acid properties. For each magnitude class m the proportion of inferred to expected changes (c_{N_m}/p_{N_m}) is compared to the total proportion of changes c_N/p_N. Positive selection is suggested when m predicts radical changes in property, and $c_{N_m}/p_{N_m} > c_N/p_N$. Significance is tested using a z-test of two-proportions. The selection on amino acid properties approach is implemented in the software package TreeSAAP (Woolley et al. 2003)

Site-wise likelihood ratio Site-wise likelihood ratio methods (Massingham and Goldman 2005) estimate d_S and d_N (denoted here as α_s and β_s, respectively) for each site using a

Table 4: **Independent-sites methods.** Different assumptions of the independent-sites methods are compared. *ASR* is the method used for ancestral state reconstruction. *Model* is the nucleotide or codon substitution model assumed. *Positive* is the inequality used to determine whether a site is under selection. *Statistical test* is the test used to determine significance. *Software* is the software package where the method is implemented which we used in this study.

	ASR	Model	Positive	Statistical test	Software
SG	Parsimony	—	$\frac{c_N}{s_N} > \frac{c_S}{s_S}$	two-tail binomial	ADAPTSITE [1]
SLAC	Likelihood	MG94 × REV	$\frac{c_N}{s_N} > \frac{c_S}{s_S}$	extended binomial	HyPhy [2]
SAAP	Likelihood	REV	$\frac{c_{N_m}}{p_{N_m}} > \frac{c_N}{p_N}$	two-proportion z-test	TreeSAAP [3]
FEL	—	MG94 × REV	$\alpha_s > \beta_s$	χ^2 (d.f. = 1)	HyPhy

[1] version 1.5 (Suzuki et al. 2001) [2] version 2.0 (Kosakovsky Pond et al. 2005) [3] version 3.2 Woolley et al. (2003)

codon-based substitution model within the maximum likelihood framework. By using a codon model, this approach can incorporate realistic assumptions regarding the substitution rate matrix and equilibrium codon frequencies. In most cases, fitting a full codon model to each site in the alignment is impractical due to the excessive parameters that must be estimated. Instead, the phylogenetic tree, branch lengths, and nucleotide model parameters are estimated for the entire alignment and shared among sites. Estimates of α_s and β_s are optimized at each site while keeping other parameters fixed. Selection is tested using a likelihood ratio test of the null hypothesis $\alpha_s = \beta_s$ against the composite hypothesis $\alpha_s \neq \beta_s$ (Muse and Gaut 1994). We use the fixed-effects likelihood (FEL) method of Kosakovsky Pond and Frost (2005b) implemented in HyPhy v2.0 (Kosakovsky Pond et al. 2005) (table 4). Tree topology, branch lengths, and substitution parameters are fixed under the MG94 X REV codon model. For each site, the likelihood of the neutral hypothesis $\alpha_s = \beta_s$ is tested against $\alpha_s \neq \beta_s$, with $2\Delta\ell$ compared against the asymptotic χ^2 distribution.

Random-sites models Random-sites models identify positively selected sites by estimating the distribution of ω among sites using a codon-based substitution model. The first codon models were described by Muse and Gaut (1994) and Goldman and Yang (1994) (see Anisimova and Kosiol (2009) for review). Muse and Gaut's (1994) model (MG) introduced two separate parameters, α_s and β_s for synonymous and nonsynonymous rates, while Goldman and Yang's (1994) (GY) model estimates the composite parameter ω. Initially, both models assumed constant nonsynonymous and synonymous rates among sites. Random-sites models relax this assumption by considering ω at each site to be a random variable drawn from a statistical distribution. Branch lengths, substitution model, and ω rate distribution parameters are estimated for the entire alignment. After fitting the model, an empirical Bayes approach is used to estimate the posterior distribution of rates for each site. In the naive empirical Bayes (NEB) approach, parameters are fixed at their MLEs while estimating the posterior distribution. The Bayes empirical Bayes approach (BEB) incorporates uncertainty in estimates of the ω distribution, while other parameters are fixed.

In this study, we compare four random-sites methods (table 5). M2a and M8 are the tests described by Nielsen and Yang (1998) and Yang et al. (2000). Each of these tests fits a pair of nested models to the data: a neutral model where $\omega <= 1$ for all sites and an adaptive model with a proportion of sites with $\omega > 1$. The first pair of models, M1a and M2a, assume a proportion of sites with $0 < \omega_0 < 1$ and $\omega_1 = 1$. For M2a, a third class of sites with $\omega > 1$ is assumed. The second pair of models, M7 and M8, assume that ω is beta-distributed among sites. M8 includes an additional category of sites with $\omega > 1$. The likelihood ratio is used to compare the neutral and adaptive models using the asymptotic χ^2 distribution(Anisimova et al. 2001), and BEB is used to estimate the posterior probability of $\omega > 1$. Strong evidence that a site is under positive selection is indicated by a significant likelihood ratio test and a high posterior probability for $\omega > 1$ (Anisimova et al. 2002; Yang et al. 2005). The variable nonsynonymous rates (NSR) and dual variable rates (DUAL) models were proposed by Kosakovsky Pond and Muse (2005). NSR is an MG-based analog of Yang et al.'s (2000) random-sites models. The synonymous rate α_s is assumed to be 1 for all sites, and a general discrete distribution (GDD) is used to describe variation in the nonsynonymous rate. DUAL relaxes the assumption of a constant synonymous rate, and both α_s and β_s are drawn from independent GDDs. Synonymous rate variation can be tested by comparing NSR and DUAL using a likelihood ratio test or Akaike Information Criterion (AIC). NEB is used to estimate the posterior distribution of α_s and β_s. A large Bayes factor for $\beta_s > \alpha_s$ at a particular site is strong evidence of positive selection. Models M2a and M8 are implemented in the software package PAML v4.4 (Yang 2007); NSR and DUAL are implemented in HyPhy v2.0 (Kosakovsky Pond et al. 2005).

Table 5: **Random-sites models.** Different assumptions of the random-sites methods are compared. *Model* is the codon substitution model assumed. *Rate classes* lists the different classes of sites assumed under the ω distribution. *Neutral* rate classes are those for which $\omega \leq 1$; *Adaptive* rate classes have $\omega > 1$. Note that for NSR and DUAL, each site belongs to one rate class for α_s and one for β_s. For NSR, only one rate class exists for $\alpha_s = 1$. Since we are using discrete distributions with three categories, there are three possible rate classes for NSR and nine possible rate classes for DUAL. *Bayes* indicates the empirical Bayes approach used to assign posterior distributions to sites. *Software* is the software package where the method is implemented which we used in this study.

		Rate Classes			
	Model	Neutral	Adaptive	Bayes	Software
M2a	GY94 × HKY+Γ	$\omega_0 < 1$, $\omega_1 = 1$	$\omega_2 > 1$	BEB	PAML [1]
M8	GY94 × HKY+Γ	$\omega \sim \text{beta}(p, q)$	$\omega_s > 1$	BEB	PAML
NSR	MG94 × REV	$\alpha_s = 1$	$\beta_s \sim$ GDD3	NEB	HyPhy [2]
DUAL	MG94 × REV	$\alpha_s \sim$ GDD3	$\beta_s \sim$ GDD3	NEB	HyPhy

[1] version 4.4 (Yang 2007) [2] version 2.0 (Kosakovsky Pond et al. 2005)

Empirical data

We analyzed 109 HIV-1 *pol* nucleotide sequence datasets collected from patients receiving antiretroviral drug therapy during phase I and II clinical studies (Condra et al. 1996; Zhang et al. 1997; Vaillancourt et al. 1999; Bacheler et al. 2000). Each dataset consists of at least 15 sequences isolated from an individual patient at multiple timepoints. Nucleotide sequences, sampling timepoints and drug treatment data were retrieved from the HIV drug resistance database (HIVdb) (Rhee et al. 2003; Shafer 2006). Detailed information about each dataset are presented in table 1.

Classification of drug-resistant sites

Drug-resistant mutations (DRMs) were identified using the Sierra webservice provided by the HIV Drug Resistance Database (HIVdb: http://hivdb.stanford.edu/), algorithm version 6.0.11F (Liu and Shafer 2006). Drug-resistant genotypes were further classified into "major" (primary) and "minor" (secondary) DRMs. In general, a major (or primary) mutation can reduce susceptibility to one or more drugs by itself; minor (or secondary) mutations typically operate by further reducing drug susceptibility or increasing replication fitness in a virus already containing one or more major mutations. Minor mutations may occur naturally in untreated patients, while major mutations are commonly observed in patients during virological failure, and are not observed in untreated patients. Mutations which occur in structurally important and highly conserved regions of the protein are likely to be classified as major.A complete description of this classification scheme can be found on the HIVdb website.

Phylogenetic analysis and population parameters

Phylogenetic relationships and population parameters were estimated for each dataset using the serial-sample coalescent model (Rodrigo et al. 1999). This model is appropriate for intrapatient HIV studies since it considers evolution that may occur during sample intervals. Demographic models, which describe the change in effective population size over time, are inferred under the Bayesian skyline (Drummond et al. 2005) model. HIV-1 is known to exhibit complex patterns of nucleotide substitution (Posada and Crandall 2001); thus we favored the use of general substitution models versus more restrictive models. The pattern of nucleotide substitution is described using the REV model with gamma-distributed rate heterogeneity among sites (Yang 1994). Rates among branches are uncorrelated and drawn independently from a lognormal distribution (Drummond et al. 2006). The phylogenetic tree topology, branch lengths, and coalescent parameters were co-estimated using Bayesian Markov chain

Monte Carlo (MCMC) methods implemented in the software package BEAST (Drummond and Rambaut 2007). Two independent analyses were performed for each dataset with 5×10^8 states sampled. Trace plots were examined to confirm convergence to the stationary distribution; states sampled before reaching convergence are discarded as burn-in. For each analysis, the maximum *a posteriori* tree topology was chosen from the posterior distribution. This tree topology was used in all subsequent analyses unless otherwise indicated.

The posterior distributions of the effective population size N_e and mutation rate μ were obtained from the BEAST analysis. The Bayesian point estimate of genetic diversity $\Theta_B = 2N_e\mu$ was compared to estimates of genetic diversity (Θ_W and Π) calculated using summary statistic methods implemented in VariScan (Vilella et al. 2005). Additional parameters commonly used to identify selection (D, D^* and F^*) were also calculated using VariScan; results for each dataset are shown in Appendix A.

Acknowledgements

We thank Brigham Young University for financial support for this study through their Mentoring Environment Grants program. This study was also partially funded by USDA grant CSREES 2007-01737.

References

M. Anisimova and C. Kosiol. Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Molecular Biology and Evolution*, 26(2):255–271, Feb. 2009.

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

M. Anisimova, J. P. Bielawski, and Z. Yang. Accuracy and power of bayes prediction of amino acid sites under positive selection. *Molecular Biology and Evolution*, 19(6):950–958, June 2002.

L. T. Bacheler, E. D. Anton, P. Kudish, D. Baker, J. Bunville, K. Krakowski, L. Bolling, M. Aujay, X. V. Wang, D. Ellis, M. F. Becker, A. L. Lasut, H. J. George, D. R. Spalding, G. Hollis, and K. Abremski. Human immunodeficiency virus type 1 mutations selected in patients failing efavirenz combination therapy. *Antimicrobial Agents and Chemotherapy*, 44(9):2475–2484, Sept. 2000.

M. A. Bakewell, P. Shi, and J. Zhang. More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proceedings of the National Academy of Sciences*, 104 (18):7489 –7494, May 2007.

W. J. Bruno. Modeling residue usage in aligned protein sequences via maximum likelihood. *Molecular Biology and Evolution*, 13(10):1368–1374, Dec. 1996.

J. H. Condra, D. J. Holder, W. A. Schleif, O. M. Blahy, R. M. Danovich, L. J. Gabryelski, D. J. Graham, D. Laird, J. C. Quintero, A. Rhodes, H. L. Robbins, E. Roth, M. Shivaprakash, T. Yang, J. A. Chodakewitz, P. J. Deutsch, R. Y. Leavitt, F. E. Massari, J. W. Mellors, K. E. Squires, R. T. Steigbigel, H. Teppler, and E. A. Emini. Genetic correlates of in vivo viral resistance to indinavir, a human immunodeficiency virus type 1 protease inhibitor. *Journal of Virology*, 70(12):8270–8276, Dec. 1996.

K. A. Crandall and D. M. Hillis. Rhodopsin evolution in the dark. *Nature*, 387(6634): 667–668, June 1997.

K. A. Crandall, C. R. Kelsey, H. Imamichi, H. C. Lane, and N. P. Salzman. Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. *Molecular Biology and Evolution*, 16(3):372–82, Mar. 1999.

A. J. Drummond and A. Rambaut. BEAST: bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7:214, 2007.

A. J. Drummond, A. Rambaut, B. Shapiro, and O. G. Pybus. Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, 22(5):1185–1192, May 2005.

A. J. Drummond, S. Y. W. Ho, M. J. Phillips, and A. Rambaut. Relaxed phylogenetics and dating with confidence. *PLoS Biology*, 4(5):e88, May 2006.

N. Goldman and Z. Yang. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution*, 11(5):725–36, Sept. 1994.

A. L. Hughes, T. Ota, and M. Nei. Positive darwinian selection promotes charge profile diversity in the antigen-binding cleft of class i major-histocompatibility-complex molecules. *Molecular Biology and Evolution*, 7(6):515–524, Nov. 1990.

S. L. Kosakovsky Pond and S. D. W. Frost. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics (Oxford, England)*, 21 (10):2531–2533, May 2005a.

S. L. Kosakovsky Pond and S. D. W. Frost. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Molecular Biology and Evolution*, 22(5):1208–22, May 2005b.

S. L. Kosakovsky Pond and S. V. Muse. Site-to-Site variation of synonymous substitution rates. *Mol Biol Evol*, 22(12):2375–2385, Dec. 2005.

S. L. Kosakovsky Pond, S. D. W. Frost, and S. V. Muse. HyPhy: hypothesis testing using phylogenies. *Bioinformatics*, 21(5):676–679, Mar. 2005.

T. F. Liu and R. W. Shafer. Web resources for HIV type 1 genotypic-resistance test interpretation. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 42(11):1608–1618, June 2006.

T. Massingham and N. Goldman. Detecting amino acid sites under positive selection and purifying selection. *Genetics*, 169(3):1753–1762, Mar. 2005.

D. A. McClellan and D. D. Ellison. Assessing and improving the accuracy of detecting protein adaptation with the TreeSAAP analytical software. *International Journal of Bioinformatics Research and Applications*, 6(2):120–133, 2010.

D. A. McClellan and K. G. McCracken. Estimating the influence of selection on the variable amino acid sites of the cytochrome b protein functional domains. *Molecular Biology and Evolution*, 18(6):917–25, June 2001.

G. A. T. McVean and J. Vieira. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in drosophila. *Genetics*, 157(1):245–257, Jan. 2001.

S. V. Muse and B. S. Gaut. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution*, 11(5):715–724, Sept. 1994.

R. Nielsen and Z. Yang. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, 148(3):929–36, Mar. 1998.

R. Nielsen, C. Bustamante, A. G. Clark, S. Glanowski, T. B. Sackton, M. J. Hubisz, A. Fledel-Alon, D. M. Tanenbaum, D. Civello, T. J. White, J. J. Sninsky, M. D. Adams, and M. Cargill. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biology*, 3(6):e170, June 2005.

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

M. Nozawa, Y. Suzuki, and M. Nei. Response to yang et al.: Problems with bayesian methods of detecting positive selection at the DNA sequence level. *Proceedings of the National Academy of Sciences*, 106(36):E96, 2009b.

C. J. Petropoulos, N. T. Parkin, K. L. Limoli, Y. S. Lie, T. Wrin, W. Huang, H. Tian, D. Smith, G. A. Winslow, D. J. Capon, and J. M. Whitcomb. A novel phenotypic drug susceptibility assay for human immunodeficiency virus type 1. *Antimicrobial Agents and Chemotherapy*, 44(4):920–928, Apr. 2000.

M. L. Porter, T. W. Cronin, D. A. McClellan, and K. A. Crandall. Molecular characterization of crustacean visual pigments and the evolution of pancrustacean opsins. *Molecular Biology and Evolution*, 24(1):253–268, Jan. 2007.

D. Posada and K. A. Crandall. Selecting models of nucleotide substitution: an application to human immunodeficiency virus 1 (HIV-1). *Molecular Biology and Evolution*, 18(6): 897–906, June 2001.

S. Rhee, M. J. Gonzales, R. Kantor, B. J. Betts, J. Ravela, and R. W. Shafer. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Research*, 31(1):298–303, Jan. 2003.

A. G. Rodrigo, E. G. Shpaer, E. L. Delwart, A. K. Iversen, M. V. Gallo, J. Brojatsch, M. S. Hirsch, B. D. Walker, and J. I. Mullins. Coalescent estimates of HIV-1 generation time in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, 96(5):2187–2191, Mar. 1999.

S. L. Sawyer, L. I. Wu, M. Emerman, and H. S. Malik. Positive selection of primate TRIM5alpha identifies a critical species-specific retroviral restriction domain. *Proceedings of the National Academy of Sciences of the United States of America*, 102(8):2832–2837, Feb. 2005.

R. W. Shafer. Rationale and uses of a public HIV drug-resistance database. *The Journal of Infectious Diseases*, 194 Suppl 1:S51–58, Sept. 2006.

R. W. Shafer, P. Hsu, A. K. Patick, C. Craig, and V. Brendel. Identification of biased amino acid substitution patterns in human immunodeficiency virus type 1 isolates from patients treated with protease inhibitors. *Journal of Virology*, 73(7):6197–6202, July 1999.

R. W. Shafer, S. Rhee, D. Pillay, V. Miller, P. Sandstrom, J. M. Schapiro, D. R. Kuritzkes, and D. Bennett. HIV-1 protease and reverse transcriptase mutations for drug resistance surveillance. *AIDS (London, England)*, 21(2):215–23, Jan. 2007.

P. M. Sharp. In search of molecular darwinism. *Nature*, 385(6612):111–112, Jan. 1997.

C. B. Stewart, J. W. Schilling, and A. C. Wilson. Adaptive evolution in the stomach lysozymes of foregut fermenters. *Nature*, 330(6146):401–404, Dec. 1987.

Y. Suzuki and T. Gojobori. A method for detecting positive selection at single amino acid sites. *Molecular Biology and Evolution*, 16(10):1315–1328, Oct. 1999.

Y. Suzuki and M. Nei. 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, Nov. 2002.

Y. Suzuki, T. Gojobori, and M. Nei. ADAPTSITE: detecting natural selection at single amino acid sites. *Bioinformatics (Oxford, England)*, 17(7):660–661, July 2001.

M. Vaillancourt, D. Irlbeck, T. Smith, R. W. Coombs, and R. Swanstrom. The HIV type 1 protease inhibitor saquinavir can select for multiple mutations that confer increasing resistance. *AIDS Research and Human Retroviruses*, 15(4):355–363, Mar. 1999.

A. J. Vilella, A. Blanco-Garcia, S. Hutter, and J. Rozas. VariScan: analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. *Bioinformatics (Oxford, England)*, 21(11):2791–2793, June 2005.

W. S. W. Wong, Z. Yang, N. Goldman, and R. Nielsen. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics*, 168(2):1041–1051, Oct. 2004.

W. S. W. Wong, R. Sainudiin, and R. Nielsen. Identification of physicochemical selective pressure on protein encoding nucleotide sequences. *BMC Bioinformatics*, 7:148, 2006.

S. Woolley, J. Johnson, M. J. Smith, K. A. Crandall, and D. A. McClellan. TreeSAAP: selection on amino acid properties using phylogenetic trees. *Bioinformatics (Oxford, England)*, 19(5):671–2, Mar. 2003.

X. Xia and W. H. Li. What amino acid properties affect protein evolution? *Journal of Molecular Evolution*, 47(5):557–564, Nov. 1998.

Z. Yang. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution*, 39(3):306–314, Sept. 1994.

Z. Yang. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8):1586–91, Aug. 2007.

Z. Yang and M. dos Reis. Statistical properties of the branch-site test of positive selection. *Molecular Biology and Evolution*, 28(3):1217–1228, Mar. 2011.

Z. Yang and R. Nielsen. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Molecular Biology and Evolution*, 25(3): 568–79, Mar. 2008.

Z. Yang, R. Nielsen, N. Goldman, and A. M. Pedersen. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155(1):431–49, May 2000.

Z. Yang, W. S. W. Wong, and R. Nielsen. Bayes empirical bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution*, 22(4):1107–1118, Apr. 2005.

Z. Yang, R. Nielsen, and N. Goldman. In defense of statistical methods for detecting positive selection. *Proceedings of the National Academy of Sciences of the United States of America*, 106(36):E95; author reply E96, Sept. 2009.

S. Yokoyama, T. Tada, H. Zhang, and L. Britt. Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates. *Proceedings of the National Academy of Sciences of the United States of America*, 105(36):13480–13485, Sept. 2008.

J. Zhang, S. Rhee, J. Taylor, and R. W. Shafer. Comparison of the precision and sensitivity of the antivirogram and PhenoSense HIV drug susceptibility assays. *Journal of Acquired Immune Deficiency Syndromes (1999)*, 38(4):439–444, Apr. 2005.

Y. M. Zhang, H. Imamichi, T. Imamichi, H. C. Lane, J. Falloon, M. B. Vasudevachari, and N. P. Salzman. Drug resistance during indinavir therapy is caused by mutations in the protease gene and in its gag substrate cleavage sites. *Journal of Virology*, 71(9):6662–6670, Sept. 1997.

Chapter 4

Conclusion

In this thesis, I have presented two studies that evaluate several approaches for identifying positively selected sites. In the first study, I examined whether methods for detecting sites under positive selection could successfully identify substitutions known to affect protein function. Empirical data from ANGPTL4, with known loss-of-function mutations, was used to test this hypothesis. The second study considered the overall and relative performance of eight different site-prediction methods using empirical data. Drug-resistant mutations that had accumulated during antiretroviral therapy were treated as known adaptive sites.

In both studies, SAAP successfully recognizes many of the known sites. However, this approach also recognizes many additional sites that are not known to affect function or evolve adaptively. Although the role of these sites remains a mystery, I presume that SAAP has a high false-positive rate in these data. PolyPhen, an approach explicitly designed for detecting functional variants, successfully identified all substitutions known to prevent protein secretion, yet failed to detect variants that cause reduced protein function. The reason for this discrepancy is unclear and warrants further research into the physical and chemical differences brought about by these substitutions.

Independent- and random-sites approaches evaluated in the HIV-1 drug-resistance study exhibited low power in predicting positively selected sites. Nearly half (44.0%) of the sites known to be under positive selection were not identified by any of the eight methods. Only SAAP, NSR, and DUAL identified greater than 10% of the known sites. Considerable overlap was found in the sites identified, particularly among similar methods. Thus, although each method takes a different approach, there are some aspects of the evolutionary signal that are detected by multiple approaches. All of the methods identified several substitutions where the selective pressure is unknown (ONMs). Although there is no evidence that these sites are non-adaptive, differences in the rate of ONM detection may indicate that specificity also varies among methods. It is clear that independent-sites approaches, aside from SAAP, were less prone to identifying ONMs than random-sites methods, even when model selection criteria is used. Thus, I conclude that the investigator must consider the goals of the study when choosing among independent- and random-sites methods; studies that require high-specificity may prefer using independent-sites approaches, while random-sites approaches are excellent tools for generating hypotheses. Ultimately, the assumptions and statistical properties each method must be carefully considered when interpreting results.

My goal in this thesis was to provide some insights into the strengths and weakness of existing computational approaches for identifying critical sites in protein-coding sequences. My hope is that this research enhances our understanding of molecular evolution and that these findings can be used to inform future research towards improved computational methods.

Appendices

Appendix A

Software utilities

Listing A.1: Perl script for parsing and manipulating TCS network files in GML format.

```perl
#! /usr/bin/perl

use strict;
use integer;
use Bio::PrimarySeq;

### Parse command line
my $usage   = "USAGE: $0 [INFILE] [ACC_NUM]";
@ARGV == 2 or die "$usage\n";
my $infile   = $ARGV[0];
my $acc_num = $ARGV[1];

### Setup input files
open( FH, "$infile" ) or die "not $infile\n";
my @lines = <FH>;
my $slurp = join('',@lines);
close FH;

### Build arrays for nodes and edges
$slurp =~ s/\s*//g;
my @n_arr   = split (/node\[/,$slurp);
my $last    = pop(@n_arr);
my @e_arr   = split (/edge\[/,$last);
my $n = shift(@e_arr);
push (@n_arr, $n);
```

```
shift @n_arr;

### Setup output files
open (POLYIN,”>polyin.txt”) or die ”not polyin.txt”;
open (OF, ”>$infile.AA”) or die ”not $infile.AA”;
my $i = 0;
while ($lines[$i] !~ /edge\s*\[/) {print OF ”$lines[$i++]”;}

### Create hash mapping node ID to sequences
my $nodes = {};
foreach my $node (@n_arr) {
  $node =~ /Sequence\”(\w+)\”Fre.*id(\d+)lab/;
  my $seqobj = Bio::PrimarySeq->new (-seq       => ”$1”,
                                     -id        => ”$2”,
                                     -alphabet  => ”dna”
                                    );
  $nodes->{”$2”} = $seqobj;
}

### Determine what changes occur on each edge
foreach my $edge (@e_arr) {
  my ($sty, $pos, $c_s, $c_t, $brl, $src, $tar) =
  ($edge =~ /linestyle\”(\w+)
            \”label\”(\d+)
            \”data\[Changes\”(\w)(\w)
            \”BranchLength\”(\d+\.?\d*)
            \”\]source(\d+)target(\d+)\]/x);

  my $seq_src = $nodes->{”$src”};
  my $seq_tar = $nodes->{”$tar”};

  my $aa_src = $seq_src->translate;
  my $aa_tar = $seq_tar->translate;
  my $aa_pos = (($pos-1) / 3) + 1;
  $pos = $aa_pos;
  $c_s = $aa_src->subseq($aa_pos,$aa_pos);
  $c_t = $aa_tar->subseq($aa_pos,$aa_pos);
  if ($c_s ne $c_t) {
    $sty = ”dashed”;  # blue
    print POLYIN ”$acc_num $pos $c_s $c_t\n”;
  }
  my $formatted = edge2str($sty, $pos, $c_s, $c_t, $brl, $src, $tar);
  print OF ”$formatted”;
}
print OF ”]\n”;

### Clean up
close POLYIN;
close OF;
print ”$0 execution complete\n”;

###  Parameters: linestyle, label, change src, change tar,
###              branch length, source, target
###  Returns string with GML formatted edge
sub edge2str {
  my @a = @_;
  my $s = ”    edge [\n”;
  $s = $s.”        linestyle \”$a[0]\”\n”;
  $s = $s.”        label \”$a[1]\”\n”;
  $s = $s.”        data [\n”;
  $s = $s.”            Changes \”$a[2] $a[3]\”\n”;
  $s = $s.”            BranchLength \”$a[4] \”\n”;
  $s = $s.”        ]\n”;
  $s = $s.”        source $a[5]\n”;
  $s = $s.”        target $a[6]\n”;
  $s = $s.”    ]\n”;
  return $s;
}
```

Listing A.2: Perl script for parsing PolyPhen tab-delimited output files. “Possibly damaging” and “Probably damaging” substitutions identified by PolyPhen are used to annotate a TCS network.

```
#!/usr/bin/perl

use strict;

### Parse command line
my $usage = "USAGE: $0 [PP_FILE] [TCS_FILE]";
(@ARGV == 2) or die "$usage\n";
my $ppfile  = $ARGV[0];
my $tcfile  = $ARGV[1];
my $verbose = 1;

### Parse the PolyPhen output file
my $polyphen = {};
open(PFH, "$ppfile") or die "Could not open $ppfile\n";
my @plines = <PFH>;
shift @plines;
foreach my $line (@plines) {
  if ($line =~ /^\s*$/) {next;}
  else {
    my @fields = split(/\t/,$line);
    $polyphen->{$fields[3]} = { _src => $fields[4],
                                _tar => $fields[5],
                                _cat => $fields[6],
                              };
  }
}
```

```
### Print messages to terminal
if ($verbose) {
  my $i = 1;
  print ”PolyPhen Sites:\n”;
  while ( my ($pos, $params) = each(%$polyphen) ) {
    print $i++.”: ”.
    print $params->{_src}.” $pos ”.
    print $params->{_tar}.” ”;
    print $params->{_cat}.”\n”;
  }
}

### Parse the TCS AA file
open(TFH, ”$tcfile” ) or die ”Could not open $tcfile\n”;
my @lines = <TFH>;
my $slurp = join(’’,@lines);
close TFH;

### Build arrays for nodes and edges
$slurp =˜ s/\s*//g;
my @n_arr  = split (/node\[/,$slurp);
my $last   = pop(@n_arr);
my @e_arr  = split (/edge\[/,$last);
my $n = shift(@e_arr);
push (@n_arr, $n);
shift @n_arr;

### Print the node part of the file
open (OF, ”>$tcfile.PP”) or die ”Could not open $tcfile.PP”;
my $i = 0;
while ($lines[$i] !˜ /edge\s*\[/) {
  print OF ”$lines[$i]”;
  $i++;
}

### Determine what changes occur on each edge
foreach my $edge (@e_arr) {
  my ($sty, $pos, $c_s, $c_t, $brl, $src, $tar) =
      ($edge =~ /linestyle\"(\w+)
                \"label\"(\d+)
                \"data\[Changes\"(\w)(\w)
                \"BranchLength\"(\d+\.?\d*)
                \"\]source(\d+)target(\d+)\]/x);
  $sty = "solid";
  if (my $params = $polyphen->{$pos}) {
    if ($params->{_src} eq $c_s && $params->{_tar} eq $c_t) {
      if ($params->{_cat} eq "probably damaging") {$sty = "dashed";}
      if ($params->{_cat} eq "possibly damaging") {$sty = "dotted";}
    }
  }
  my $formatted = edge2str($sty, $pos, $c_s, $c_t, $brl, $src, $tar);
  print OF "$formatted";
}
print OF "]\n";

### Clean up
close OF;
print "$0 execution complete\n";

### Parameters: linestyle, label, change src, change tar,
###             branch length, source, target
### Returns string with GML formatted edge
sub edge2str {
  my @a = @_;
  my $s = "    edge [\n";
  $s = $s."        linestyle \"$a[0]\"\n";
  $s = $s."        label \"$a[1]\"\n";
  $s = $s."        data [\n";
  $s = $s."            Changes \"$a[2] $a[3]\"\n";
  $s = $s."            BranchLength \"$a[4] \"\n";
  $s = $s."        ]\n";
  $s = $s."        source $a[5]\n";
  $s = $s."        target $a[6]\n";
  $s = $s."    ]\n";
  return $s;
}
```

Listing A.3: Perl script for parsing stablizing and destabilizing selection files generated from TreeSAAP output. Sites under stabilizing and destabilizing selection are used to annotate a TCS network.

```
#!/usr/bin/perl

use strict;

### Parse command line
my $usage = "USAGE: $0 [TS_FILE] [TCS_FILE]";
(@ARGV == 2) or die "$usage\n";
my $tsfile  = $ARGV[0];
my $tcfile  = $ARGV[1];

### Parse the TreeSAAP output file
my @suds;
open(SAAP, "$tsfile") or die "Could not open $tsfile\n";
my @plines = <SAAP>;
my $all = $plines[1];
$all =~ /\(resi ([\d|,]*) \)/;
my @sites = split(/,/,$1);
```

```

### Parse the TCS AA file
open(TFH, "$tcfile" ) or die "Could not open $tcfile\n";
my @lines = <TFH>;
my $slurp = join('',@lines);
close TFH;

### Build arrays for nodes and edges
$slurp =~ s/\s*//g;
my @n_arr  = split (/node\[/,$slurp);
my $last   = pop(@n_arr);
my @e_arr  = split (/edge\[/,$last);
my $n = shift(@e_arr);
push (@n_arr, $n);
shift @n_arr;

### Print the node part of the file
open (OF, ">$tcfile.TS") or die "Could not open $tcfile.PPnot";
my $i = 0;
while ($lines[$i] !~ /edge\s*\[/) {
  print OF "$lines[$i]";
  $i++;
}

###  Determine what changes occur on each edge
foreach my $edge (@e_arr) {
  my ($sty, $pos, $c_s, $c_t, $brl, $src, $tar) =
      ($edge =~ /linestyle\"(\w+)
                  \"label\"(\d+)
                  \"data\[Changes
                  \"(\w)(\w)\"BranchLength\"(\d+\.?\d*)
                  \"\]source(\d+)target(\d+)\]/x);
  $sty = "solid";
  if ($c_s ne $c_t) {
    foreach my $site (@sites) {
      if ($site == $pos) {$sty = "dashed";}
    }
  }
  my $formatted = edge2str($sty, $pos, $c_s, $c_t, $brl, $src, $tar);
  print OF "$formatted";
}
print OF "]\n";

### Clean up
close OF;
print "$0 execution complete\n";

###  Parameters: linestyle, label, change src, change tar,
###              branch length, source, target
###  Returns string with GML formatted edge
sub edge2str {
  my @a = @_;
  my $s = "    edge [\n";
  $s = $s."        linestyle \"$a[0]\"\n";
  $s = $s."        label \"$a[1]\"\n";
  $s = $s."        data [\n";
  $s = $s."            Changes \"$a[2] $a[3]\"\n";
  $s = $s."            BranchLength \"$a[4] \"\n";
  $s = $s."        ]\n";
  $s = $s."        source $a[5]\n";
  $s = $s."        target $a[6]\n";
  $s = $s."    ]\n";
  return $s;
}
```

Appendix B

HIV-1 Patient Datasets

Table 1:

Dataset statistics and parameter estimates. The first column contains the abbreviation for the dataset. The first two letters indicate the published study for the sequences, followed by the patient identifier. Studies used are: YZ, Zhang et al. (1997); LB, Bacheler et al. (2000); MV, Vaillancourt et al. (1999); JC, Condra et al. (1996). LB sequences are 984 bp (protease A.A. 1-99 and reverse transcriptase A.A. 1-229). YZ, MV and JC sequences are 297 bp (protease A.A. 1-99). n is the number of sequences used in the analysis. TPs is the number of timepoints sampled. Wks is the number of weeks between the first and last sample timepoints. N_e and μ are the median effective population size and mean mutation rate, respectively, estimated from the Bayesian skyline Drummond et al. (2005) analysis. $\Theta_{\mathrm{B}} = 2N_e\mu$ is the estimate of genetic diversity from the Bayesian analysis. Θ_W is Watterson's (1975) estimator of nucleotide diversity per site based on the total number of mutations (η) . Π is the nucleotide diversity, the average pairwise nucleotide differences per site. D is Tajima's (1989) statistical test for neutrality. D^* and F^* are the tests proposed by Fu and Li (1993).

	n	TPs	Wks	N_e	μ	Θ_{B}	Θ_{W}	Π	D	D^*	F^*
YZ-p3	89	9	75	1085.4	5.40	0.117	0.053	0.024	-1.77	-4.25	-3.82
YZ-p4	71	8	75	1280.9	3.01	0.077	0.033	0.015	-1.80	-3.79	-3.52
YZ-p5	65	7	59	1636.0	4.86	0.159	0.045	0.025	-1.48	-3.07	-2.88
YZ-p6	126	13	71	964.5	8.62	0.166	0.062	0.018	-2.31	-6.93	-5.83
YZ-p7	79	9	60	1810.6	6.25	0.226	0.057	0.026	-1.82	-4.66	-4.15
YZ-p8	129	14	72	1045.8	12.74	0.266	0.076	0.034	-1.78	-5.99	-4.90
LB-p1	48	8	69	1422.7	4.79	0.136	0.037	0.024	-1.25	-3.13	-2.87
LB-p3	30	8	58	2591.7	2.31	0.120	0.028	0.014	-1.90	-3.39	-3.35

Table 1 (continued)

	n	TPs	Wks	N_e	μ	Θ_B	Θ_W	Π	D	D^*	F^*
LB-p4	29	3	34	4125.8	2.68	0.222	0.027	0.014	-1.81	-2.95	-2.98
LB-p5	40	5	89	1505.1	4.00	0.120	0.032	0.014	-2.08	-3.86	-3.78
LB-p6	20	3	10	3960.2	2.73	0.216	0.027	0.017	-1.60	-2.19	-2.29
LB-p7	20	3	94	2296.1	2.07	0.095	0.019	0.012	-1.58	-2.83	-2.75
LB-p8	21	3	72	2829.5	1.82	0.103	0.016	0.007	-2.22	-3.31	-3.34
LB-p11	26	5	21	2510.9	2.28	0.115	0.021	0.014	-1.34	-2.89	-2.73
LB-p13	24	4	43	1514.4	2.46	0.074	0.023	0.013	-1.72	-2.65	-2.69
LB-p14	24	6	75	1689.1	2.76	0.093	0.029	0.015	-1.88	-2.63	-2.75
LB-p15	37	8	25	2224.0	2.09	0.093	0.026	0.017	-1.20	-3.33	-3.01
LB-p17	27	5	21	4441.0	2.08	0.185	0.025	0.012	-2.07	-3.67	-3.62
LB-p20	36	5	70	3377.4	3.61	0.244	0.038	0.017	-2.04	-3.86	-3.79
LB-p21	58	8	96	801.1	11.63	0.186	0.048	0.020	-2.08	-4.54	-4.26
LB-p22	34	4	70	2009.7	4.78	0.192	0.039	0.020	-1.87	-3.37	-3.34
LB-p24	26	4	94	3777.0	3.81	0.288	0.027	0.017	-1.35	-2.17	-2.19
LB-p26	58	6	23	5256.2	4.36	0.458	0.042	0.021	-1.79	-4.34	-3.97
LB-p28	50	9	70	4255.0	4.35	0.370	0.046	0.023	-1.82	-3.41	-3.33
LB-p32	36	7	70	1848.4	3.00	0.111	0.025	0.012	-1.97	-3.68	-3.59
LB-p39	30	5	84	3917.4	3.25	0.254	0.033	0.018	-1.78	-3.13	-3.12
LB-p44	28	7	44	3405.2	3.27	0.223	0.038	0.020	-1.83	-3.17	-3.17
LB-p47	46	7	71	1303.5	5.88	0.153	0.034	0.021	-1.40	-2.98	-2.82
LB-p50	51	6	96	5845.2	2.78	0.325	0.038	0.014	-2.23	-5.61	-5.13
LB-p55	59	8	108	1843.1	4.55	0.168	0.041	0.016	-2.17	-5.14	-4.73
LB-p57	38	6	87	1910.7	4.07	0.156	0.032	0.015	-1.94	-3.38	-3.36
LB-p58	25	4	16	3823.1	2.40	0.183	0.031	0.015	-2.09	-3.32	-3.37
LB-p60	47	8	96	1631.1	5.67	0.185	0.041	0.017	-2.15	-4.76	-4.48
LB-p63	27	5	24	4130.7	2.09	0.172	0.022	0.010	-2.18	-3.54	-3.55

Table 1 (continued)

	n	TPs	Wks	N_e	μ	Θ_{B}	Θ_{W}	Π	D	D^*	F^*
LB-p66	23	4	48	1804.7	2.28	0.082	0.022	0.011	-1.89	-2.95	-2.98
LB-p69	46	7	44	3796.4	3.60	0.273	0.044	0.019	-2.10	-4.38	-4.18
LB-p70	53	6	44	2400.1	6.32	0.303	0.041	0.018	-2.03	-4.72	-4.38
LB-p71	36	7	86	2244.9	2.72	0.122	0.032	0.017	-1.75	-3.57	-3.44
LB-p72	35	6	36	2733.4	3.35	0.183	0.035	0.022	-1.39	-2.86	-2.75
LB-p73	26	5	36	4482.7	2.46	0.220	0.032	0.013	-2.31	-4.05	-4.03
LB-p77	47	8	109	983.3	10.83	0.213	0.052	0.021	-2.20	-3.95	-3.91
LB-p80	16	3	100	3893.6	1.81	0.141	0.020	0.009	-2.28	-3.17	-3.22
LB-p81	35	6	72	5865.6	2.24	0.263	0.032	0.014	-2.16	-4.31	-4.17
LB-p83	27	4	72	1612.6	3.33	0.108	0.026	0.014	-1.82	-2.65	-2.74
LB-p84	63	9	96	1300.1	6.13	0.159	0.042	0.014	-2.30	-5.52	-5.04
LB-p86	25	5	73	1915.9	2.98	0.114	0.028	0.018	-1.47	-2.47	-2.47
LB-p87	29	4	72	2503.7	2.63	0.132	0.025	0.014	-1.65	-2.70	-2.72
LB-p89	68	11	96	1314.6	8.23	0.216	0.048	0.020	-2.01	-5.01	-4.53
LB-p91	22	6	98	3293.0	3.03	0.200	0.029	0.021	-1.15	-1.91	-1.91
LB-p93	30	7	63	3011.2	3.29	0.198	0.035	0.018	-1.90	-3.19	-3.21
LB-p94	34	8	97	2229.7	2.44	0.109	0.026	0.009	-2.48	-4.40	-4.35
LB-p98	24	6	45	2516.0	2.17	0.109	0.024	0.014	-1.75	-3.17	-3.11
LB-p99	35	5	20	2756.4	2.36	0.130	0.023	0.012	-1.73	-3.41	-3.28
LB-p100	46	8	56	1837.1	3.49	0.128	0.032	0.016	-1.81	-3.43	-3.33
LB-p101	23	4	72	2671.5	2.55	0.136	0.026	0.015	-1.72	-2.62	-2.67
LB-p102	25	5	42	1632.9	2.73	0.089	0.026	0.017	-1.46	-2.59	-2.56
LB-p106	39	6	61	5872.0	3.19	0.375	0.036	0.017	-1.97	-3.58	-3.54
LB-p107	40	8	60	9659.1	2.31	0.447	0.042	0.014	-2.42	-4.85	-4.69
LB-p109	20	6	56	2284.6	2.16	0.099	0.023	0.013	-1.74	-2.40	-2.48
LB-p110	29	5	59	2773.5	3.03	0.168	0.025	0.016	-1.26	-2.19	-2.17

Table 1 (continued)

	n	TPs	Wks	N_e	μ	Θ_B	Θ_W	Π	D	D^*	F^*
LB-p111	23	4	71	1964.6	2.75	0.108	0.024	0.016	-1.31	-2.40	-2.35
LB-p112	18	4	63	1801.8	2.81	0.101	0.028	0.020	-1.20	-1.78	-1.82
LB-p113	26	5	68	4734.4	3.12	0.296	0.034	0.019	-1.69	-2.93	-2.93
LB-p114	22	5	59	4991.1	1.79	0.179	0.021	0.009	-2.26	-3.45	-3.49
LB-p115	19	5	20	2082.3	2.18	0.091	0.020	0.012	-1.76	-2.38	-2.46
LB-p117	20	5	82	2549.9	2.73	0.139	0.025	0.016	-1.41	-2.00	-2.06
LB-p119	27	5	68	2774.5	1.97	0.109	0.022	0.009	-2.37	-4.07	-4.03
LB-p120	38	5	28	3353.0	2.06	0.138	0.029	0.013	-2.06	-4.49	-4.25
LB-p121	57	10	46	10004.5	2.98	0.597	0.041	0.016	-2.21	-4.43	-4.23
LB-p123	16	5	19	1922.3	3.00	0.115	0.025	0.022	-0.48	-0.95	-0.91
LB-p124	17	4	36	1800.0	3.22	0.116	0.028	0.021	-1.09	-1.59	-1.62
LB-p126	41	7	59	4887.2	3.37	0.329	0.038	0.018	-1.95	-2.97	-3.06
LB-p130	26	4	12	3449.5	2.79	0.192	0.028	0.018	-1.33	-2.61	-2.53
LB-p132	54	9	71	4182.3	4.10	0.343	0.041	0.018	-1.99	-4.33	-4.08
LB-p134	19	5	65	2427.6	2.54	0.123	0.026	0.016	-1.51	-2.27	-2.31
LB-p135	15	3	16	3150.7	2.06	0.130	0.020	0.011	-2.00	-2.45	-2.56
LB-p141	36	6	24	4912.9	2.91	0.286	0.032	0.016	-1.79	-3.32	-3.26
LB-p142	41	9	62	3650.4	2.65	0.194	0.029	0.013	-2.07	-4.08	-3.94
LB-p143	46	8	61	3393.6	3.15	0.213	0.036	0.016	-2.02	-4.00	-3.85
LB-p144	20	3	20	2054.9	3.21	0.132	0.035	0.024	-1.35	-1.87	-1.95
LB-p145	53	8	67	3174.1	4.70	0.298	0.041	0.019	-1.89	-4.32	-4.02
LB-p146	25	4	16	1830.4	4.06	0.149	0.041	0.025	-1.51	-2.44	-2.49
LB-p147	23	6	42	2608.1	2.53	0.132	0.026	0.016	-1.57	-2.25	-2.33
LB-p148	21	6	50	4430.3	2.76	0.245	0.031	0.018	-1.72	-2.42	-2.52
LB-p151	18	4	12	3003.9	2.87	0.173	0.022	0.016	-1.20	-1.94	-1.93
LB-p152	24	4	24	2377.1	1.65	0.079	0.015	0.008	-1.92	-2.77	-2.82

Table 1 (continued)

	n	TPs	Wks	N_e	μ	Θ_B	Θ_W	Π	D	D^*	F^*
LB-p154	22	4	69	1803.5	3.07	0.111	0.030	0.017	-1.75	-2.74	-2.79
LB-p155	20	5	16	7270.1	2.41	0.350	0.029	0.016	-1.91	-2.69	-2.79
LB-p156	28	4	16	6194.2	2.01	0.249	0.024	0.010	-2.20	-3.95	-3.88
LB-p158	22	3	56	2516.8	3.02	0.152	0.024	0.015	-1.56	-2.37	-2.41
LB-p159	20	3	40	2251.4	3.12	0.141	0.030	0.022	-1.09	-1.77	-1.78
LB-p167	16	4	50	2302.2	3.37	0.155	0.034	0.022	-1.51	-1.87	-1.99
LB-p168	21	5	20	2208.8	2.83	0.125	0.031	0.018	-1.73	-2.25	-2.39
LB-p170	18	3	60	3382.2	1.80	0.121	0.017	0.009	-2.01	-2.72	-2.79
LB-p171	18	3	54	4961.2	3.19	0.316	0.037	0.024	-1.53	-2.17	-2.26
MV-p219	39	4	56	2011.8	4.32	0.174	0.043	0.025	-1.53	-3.04	-2.92
MV-p833	29	3	40	2892.2	3.50	0.202	0.042	0.024	-1.62	-2.26	-2.33
MV-p847	34	3	32	1826.8	2.56	0.094	0.025	0.015	-1.44	-2.42	-2.35
JC-pA	62	7	52	1352.6	4.20	0.114	0.036	0.035	-0.07	-1.17	-0.87
JC-pB	60	7	52	1442.5	2.56	0.074	0.037	0.021	-1.44	-4.01	-3.54
JC-pD	21	3	44	1441.9	2.47	0.071	0.012	0.013	0.21	-0.86	-0.56
JC-pE	29	4	32	1261.6	2.29	0.058	0.027	0.029	0.28	-0.81	-0.51
JC-pI	18	3	16	1417.3	2.36	0.067	0.025	0.021	-0.63	-0.81	-0.80
JC-pJ	34	4	36	1085.1	2.52	0.055	0.029	0.023	-0.68	-1.66	-1.50
JC-pK	20	3	60	1482.9	3.43	0.102	0.028	0.027	-0.06	-0.08	-0.08
JC-pO	22	3	60	965.0	2.26	0.044	0.018	0.018	0.10	-0.45	-0.30
JC-pP	19	3	48	1407.0	2.28	0.064	0.023	0.018	-0.91	-1.50	-1.41
JC-pQ	18	3	60	1219.6	2.71	0.066	0.029	0.029	-0.05	-0.67	-0.52
JC-pR	19	3	60	1387.4	2.15	0.060	0.024	0.014	-1.64	-2.45	-2.36

Note — μ estimates are scaled by 10^5

References

L. T. Bacheler, E. D. Anton, P. Kudish, D. Baker, J. Bunville, K. Krakowski, L. Bolling, M. Aujay, X. V. Wang, D. Ellis, M. F. Becker, A. L. Lasut, H. J. George, D. R. Spalding, G. Hollis, and K. Abremski. Human immunodeficiency virus type 1 mutations selected in patients failing efavirenz combination therapy. *Antimicrobial agents and chemotherapy*, 44(9):2475–84, Sept. 2000.

J. H. Condra, D. J. Holder, W. A. Schleif, O. M. Blahy, R. M. Danovich, L. J. Gabryelski, D. J. Graham, D. Laird, J. C. Quintero, A. Rhodes, H. L. Robbins, E. Roth, M. Shivaprakash, T. Yang, J. A. Chodakewitz, P. J. Deutsch, R. Y. Leavitt, F. E. Massari, J. W. Mellors, K. E. Squires, R. T. Steigbigel, H. Teppler, and E. A. Emini. Genetic correlates of in vivo viral resistance to indinavir, a human immunodeficiency virus type 1 protease inhibitor. *Journal of virology*, 70(12):8270–6, Dec. 1996.

A. J. Drummond, A. Rambaut, B. Shapiro, and O. G. Pybus. Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular biology and evolution*, 22(5):1185–92, May 2005.

Y. X. Fu and W. H. Li. Statistical tests of neutrality of mutations. *Genetics*, 133(3):693–709, Mar. 1993.

F. Tajima. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–95, Nov. 1989.

M. Vaillancourt, D. Irlbeck, T. Smith, R. W. Coombs, and R. Swanstrom. The HIV type 1 protease inhibitor saquinavir can select for multiple mutations that confer increasing resistance. *AIDS research and human retroviruses*, 15(4):355–63, Mar. 1999.

G. A. Watterson. On the number of segregating sites in genetical models without recombination. *Theoretical population biology*, 7(2):256–76, Apr. 1975.

Y. M. Zhang, H. Imamichi, T. Imamichi, H. C. Lane, J. Falloon, M. B. Vasudevachari, and N. P. Salzman. Drug resistance during indinavir therapy is caused by mutations in the protease gene and in its Gag substrate cleavage sites. *Journal of virology*, 71(9):6662–70, Sept. 1997.