Title

Harnessing Change: Human Health through the Lense of Evolution and Dynamical Systems Theory

Author

Maher, M. Cyrus Riley

Publication Date

2014

Harnessing Change: Human Health through the Lens of Evolution and Dynamical Systems Theory

by

M. Cyrus Maher

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Epidemiology and Translational Sciences

in the

GRADUATE DIVISION

of the

[redacted]

Dedication

A mi Daisy, el corazón fuera de mi pecho

Acknowledgements

I would like to thank Dr. Ryan D. Hernandez for his guidance and feedback throughout my graduate career, and for his example as a rigorous and dedicated scientist. Similarly, thanks to the entire Hernandez lab, the Pollard lab, and the Wall lab. Thanks to R. D. H., Katie Pollard, Nevan Krogan, and John Witte for their feedback on this dissertation. A special thanks to John for helping me into the PhD program in the first place! I would also like to thank my former advisors Dr. Travis Porco and Dr. Tom Lietman. Their generous advice, expert tutorship, and unfailing support completely changed my trajectory as a scientist. I would like to thank all my wonderful collaborators in the Krogan and McCune labs. You guys made graduate school fun! I would of course like to thank my parents for loving me, giving me life, and always nourishing my curiousity. You guys listened to me even when I had nothing else to talk about except statistics. That's true unconditional love! Thanks to my siblings for making sure I never lost my sense of humor. Thanks to my Daisy and our doggies for providing support, love, and a safe harbor in even the roughest seas. Finally, many thanks to Maria Glymour, Bob Hiatt, and the Department of Epidemiology for providing an academic home. This work was supported for the first two years by the Epidemiology and Translational Science program. Subsequent years were funded by a National Institutes of Health F31 Predoctoral Fellowship (grant number 1 F31 CA180609-01) and a University of California, San Francisco Lloyd M. Kozloff Fellowship.

Abstract

Over 2000 years ago, Heraclitus noted, "Everything changes and nothing stands still[1]." While this truth has long been evident to the wise, we have only recently developed the tools necessary to scientifically characterize sweeping patterns of change in large dynamical systems. Despite rapid progress, new methods and data sources are still sorely needed to further illuminate the intricate and dynamic nature of reality. In this dissertation, we will focus our investigations on understanding patterns of change with direct relevance to human health. In the first two chapters, we develop novel methodologies that lend insight into the evolutionary history of the human race and the genetic basis of human-specific traits and disease. Chapter 1 presents MOSAIC, a new python package for improved detection of genetically related genes between species. This inference is a foundational step towards understanding the function of proteins and the evolutionary pressures they have faced. This tool, along with a combination of other methods, facilitates our analysis in Chapter 2. In this section, we use the patterns of mutations along the human lineage to discover genes and even specific mutations that may play important roles in intelligence, obesity, mental health, as well as a variety of basic biological functions. These findings provide insight into the genetic architecture of health and disease. At the same time, they leave open questions about how genetic factors interact with the broad array of environmental and ecological variables that fundamentally shape downstream phenotypes. In Chapter 3, we introduce CauseMap, a tool I built to understand causal relationships within complex dynamical systems using time series data. It is our hope that this method will help us to interpret human health and disease as states of the bodily dynamical system embedded inextricably within an evolving social, economic, and environmental network. This perspective, we hope, will allow us to understand the characteristics of human health that emerge from an time-hewn dynamic equilibrium with the world within and around us.

[1] As quoted in Plato, *Cratylus*, 402a

[2] I do not mean to imply, as many do, that the Universe began at the Big Bang or is limited in size to the distance we can see out in space (determined by the age of the universe, its expansion rate, and the speed of light). Such assertions are akin to concluding, in the absence of data, that the world ends at the horizon. We can only draw lower bounds on the extend of spacetime. Further, we have no idea what spacetime(s) actually is (are) or how it (they) came into existence. This is merely to point out that, *especially* on the biggest questions, we should not draw conclusions where we have no data.

Chapter 1 Introduction 1

References 5

Chapter 2 Rock, Paper, Scissors: Harnessing complementarity in ortholog detection methods improves comparative genomic inference. 12

Introduction 12

Materials and methods 14

Results and Discussion 22

Conclusions 32

References 34

Chapter 3 The Yin and Yang of Evolution: Insights into the shared genetic basis of human traits and diseases 41

Introduction 41

Materials and Methods 43

Results 47

Discussion 54

Conclusions 56

References 57

Chapter 4 CauseMap: Fast inference of causality from complex time series 64

Introduction 64

Implementation 66

Results and Discussion 67

Conclusions 71

Availability and Requirements 72

References 73

Appendix A Supplemental Material to Chapter Two 80

Appendix B Supplemental Material to Chapter Three 88

Appendix C Supplemental Materials to Chapter Four 102

Table of Figures

Figure 2-1. Comparison of sequence identity levels between methods........................ 23

Figure 2-2 Illustration of integration process for carbonic anhydrase 12.................... 25

Figure 2-3 OD power and the effect of pooling methods 26

Figure 2-4. MOSAIC improves alignment quality 28

Figure 2-5. A comparison of evolutionary estimates 30

Figure 2-6. Example: a MOSAIC-specific PSS in Tryptase Alpha/Beta 1 32

Figure 3-1. A comparison of simulated selections coefficients versus those inferred by SnIPRE... 48

Figure 3-2. A comparison of SnIPRE-inferred selection coefficient distributions 50

Figure 3-3. Enrichment of disease-trait associations among positively selected genes 53

Figure 4-1. An example visualization from CauseMap using abundances of *Paramecium aurelia* and *Didinium nasutum* 68

Figure 4-2. The effect of time series length on ρ_{ccm} convergence 69

Supp. Figure A-1. Distributions of percent identity relative to the highest scoring ortholog, stratified by species. 80

Supp. Figure A-2. The effect of method integration on sequence identity. 81

Supp. Figure A-3. The cumulative proportion of transcripts for which an ortholog is identified. 82

Supp. Figure A-4. The rate of concordance between functional annotations for proposal orthologs and human transcripts. 82

Supp. Figure A-5. The cumulative proportion of human transcripts as a function of the maximum allowable Robinson-Foulds distance between the gene tree and the species tree........ 83

Supp. Figure A-6. A comparison between MOSAIC and metaPhOrs. 85

Supp. Figure A-7. The distribution of gene-level conservation (measured by dN/dS) for each component method versus MOSAIC$_{\text{matched}}$........ 85

Supp. Figure A-8. A representation of the alignments returned by each method for TPSAB1. 86

Supp. Figure A-9. The MOSAIC alignment of TPSAB1. 86

Supp. Figure A-10. The *Gorilla gorilla* sequence that is orthologous to TPSAB1. 87

Supp. Figure A-11. Manually derived alignments of TPSAB1, reproduced from Trivedi et al. 2007. 87

Supp. Figure B-1. A comparison of selection coefficents calculated using the same data using SnIPRE and MKPRF, respectively. 93

Supp. Figure B-2. A comparison of dN/dS values between previously published (green) and newly calculated (blue) values. 95

Supp. Figure B-3. The distribution of SnIPRE-derived selection coefficients as a function of the allele frequency cutoff used as a criterion for including polymorphic mutations........ 96

Supp. Figure B-4. A heatmap of rank correlations between selection coefficients calculated using varying allele frequency cutoffs........ 97

Supp. Figure B-5. A plot of the joint distributions of SnIPRE-inferred gamma values using allele frequency cutoffs of 15% and 25%. 97

Supp. Figure B-6. Enriched GO processes visualized using MDS-derived semantic similarity distances. 99

Supp. Figure B-7. Enriched GO categories clustered hierarchically........ 100

Supp. Figure B-8. Distributions of postive selection scores within various groups of genes. 100

Supp. Figure B-9. For permuted selection scores, the distribution of positive selection scores as a function of the number of categories associated with a given gene. 101

Supp. Figure C-1. The maximal predictive skill as a function of E, tau p, and the number of included points. 103

Table of Tables

Table 3-1. GO processes enriched among positively selected genes. 51

Table 3-2. Diseases or traits for which Individually significant enrichment was discovered. 54

Supp. Table A-1. SwissProt database BLAST results for each of the putative orthologs of TPSAB1. 87

Chapter 1 Introduction

"Another word for life is change" – Michael Jeffreys

Science is revolutionizing the way we think about ourselves, and the world. Torrents of new data have illuminated the beautifully intricate, and at times, bafflingly complex nature of our universe. Embedded inextricably within this whole is our own complete[2] story—what I call the spacetime history of the human race. I find the staggering magnitude of this universal perspective to be inherently satisfying. In the chapters that follow, I will make the case that this viewpoint also has the potential to spawn breakthroughs in our understanding and treatment of human disease.

Physicist Bryan Swimme famously summarized all of history in one line, "You take a giant ball of hydrogen and helium gas, you leave it alone, and you end up with rosebushes, giraffes, and human beings". Our species emerged out of the creative chaos of a staggeringly immense dynamical system. In this process, we are both observers and participants. We are made from the progression of cause and effect that gave rise to the heavier atoms, then to the planets, then to the first cell, and finally, to us. We were born from change, and in a real sense, we *are* change. We are ourselves dynamical systems, each with the vanishingly rare potential to understand the nature of our world, and to harness that insight for the good of ourselves and others.

[2] I do not mean to imply, as many do, that the Universe began at the Big Bang or is limited in size to the distance we can see out in space (determined by the age of the universe, its expansion rate, and the speed of light). Such assertions are akin to concluding, in the absence of data, that the world ends at the horizon. We can only draw lower bounds on the extend of spacetime. Further, we have no idea what spacetime(s) actually is (are) or how it (they) came into existence. This is merely to point out that, *especially* on the biggest questions, we should not draw conclusions where we have no data.

During my time at UCSF, I have endeavored to develop this potential to the best of my abilities. The following thesis is a fruit of these efforts. In Chapter 1, I will present MOSAIC, a new python package geared towards improving comparative genomic inference by producing more complete and higher quality source data for downstream evolutionary inference. Specifically, MOSAIC identifies evolutionarily related genes in the genomes of distinct species (so-called orthologs). By discovering areas of high similarity along diverged sequences that share common function, researchers are able to infer the most critical components of a given gene. Rapidly acquired and lineage-specific mutations, on the other hand, may provide clues as to the genetic basis of traits unique to a given species.

In Chapter 2, we build on this work to understand the genetic basis of characteristic human traits. We utilize the sequences detected by MOSAIC to reconstruct the genetic code of each human gene just prior to the divergence from chimp. We then analyze the patterns of mutations along the human lineage to discover genes and even specific mutations that may play important roles in intelligence, obesity, mental health, as well as a variety of basic biological functions.

Examining the evolutionary origins of human traits provides useful insight into the genetic links between our strengths and susceptibilities as a species. However, this perspective lacks a dynamic perspective on the interaction of genetic factors with the broad array of environmental and ecological variables that fundamentally shape downstream phenotypes.

In Chapter 3, I will introduce CauseMap, a tool I built to understand causal relationships within complex dynamical systems using time series data. This perspective, I hope, will

allow us to understand the characteristics of human health that emerge from an evolved dynamic equilibrium with the world within and around us. This interdependence is central to our health and physically permeates who we are. For example, recent work has demonstrated the striking importance of the microbiome in protecting against e.g. irritable bowel disease (IBD), obesity, diabetes, asthma, anxiety, and depression (Arrieta, Stiemsma, Amenyogbe, Brown, & Finlay, 2014; Foster & McVey Neufeld, 2013).
Our internal ecosystem is not a static entity, however. Rather, it is constantly evolving as species wax and wane in response to e.g. mutual competition and an ever-changing supply of nutrients (Caporaso et al., 2011; Fisher & Mehta, 2014; Gajer et al., 2012). It is likely that important aspects of human health are shaped by the periodic dynamics of these changes. Yet, our ability to understand these interwoven ecological relationships remains limited by a dearth of appropriate time series methods.

It is for this reason that I developed CauseMap. This tool is the first open-source implementation of Convergent Cross Mapping (CCM), a next-generation algorithm designed to understand causal relationships from ecological time series data. CCM grows out of dynamical systems theory, and while still unproven, holds great promise for understanding how elements of complex systems function *in situ*.

As a proof-of-concept, I apply CauseMap to understand the predator-prey relationship between two species of single celled organisms: *Paramecium aurelia* and *Didinium nasutum*. We use this well-known system to validate our implementation, and to demonstrate the strengths and limitations of CauseMap. We note that, despite its requirement for relatively long time series, CauseMap has the enormous advantage of requiring observations from only a single source system. In dynamical systems with widely varying or context-specific behavior, this would allow researchers to draw conclusions that are tailored to, e.g. a given patient. Rather than acting on population averages, biomedical researchers would be free to fully personalize therapy to the unique biology and ecology of the patient.

We envision many other applications for CauseMap as well. Additional examples include understanding patient-to-patient variability in drug response using time series metabolomics, and examining the basis of e.g. influenza seasonality using global time series. We are in fact in the process of processing data to answer the latter question.

Together, Chapters 1-3 expand the methodological arsenal of those interested in harnessing change to deepen our understanding of human health and disease. In addition, we present several novel findings that themselves lend new insight into the factors that shape us as a species. Like the systems that it examines, however, this work is itself a dynamic entity. What we present is merely a snapshot of a greater work that we hope will continue to evolve, to grow, and eventually, to give back in some small way to the stream of scientific advancement from which it emerged. As always, please enjoy, and let us know if you notice anything that can be improved!

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In F. Czaki & B. N. Petrov (Eds.), *2nd International Symposium on Information Theory* (pp. 267–281). Budapest: Akademiai Kiado.

Akey, J. M. (2009). Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Research*, *19*(5), 711–22. doi:10.1101/gr.086652.108

Alexeyenko, A., Tamas, I., Liu, G., & Sonnhammer, E. L. L. (2006). Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics*, *22*(14), e9–15. doi:10.1093/bioinformatics/btl213

Altenhoff, A. M., & Dessimoz, C. (2009a). Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Computational Biology*, *5*(1), e1000262. doi:10.1371/journal.pcbi.1000262

Altenhoff, A. M., & Dessimoz, C. (2009b). Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Computational Biology*, *5*(1), e1000262. doi:10.1371/journal.pcbi.1000262

Altenhoff, A. M., & Dessimoz, C. (2012). Inferring Orthology. In M. Anisimova (Ed.), *Evolutionary Genomics* (Vol. 855). Totowa, NJ: Humana Press. doi:10.1007/978-1-61779-582-4

Altenhoff, A. M., Schneider, A., Gonnet, G. H., & Dessimoz, C. (2011). OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Research*, *39*(Database issue), D289–94. doi:10.1093/nar/gkq1238

Arrieta, M.-C., Stiemsma, L. T., Amenyogbe, N., Brown, E. M., & Finlay, B. (2014). The Intestinal Microbiome in Early Life: Health and Disease. *Frontiers in Immunology*, *5*, 427. doi:10.3389/fimmu.2014.00427

Babbitt, C. C., Warner, L. R., Fedrigo, O., Wall, C. E., & Wray, G. A. (2011). Genomic signatures of diet-related shifts during human origins. *Proceedings. Biological Sciences / The Royal Society*, *278*(1708), 961–9. doi:10.1098/rspb.2010.2433

Bakewell, M. A., Shi, P., & Zhang, J. (2007). More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(18), 7489–94. doi:10.1073/pnas.0701705104

Berman, H. M. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. doi:10.1093/nar/28.1.235

Bertsekas, D. (1999). *Nonlinear Programming* (p. 780). Athena Scientific; 2nd edition.

Bezanson, J., Karpinski, S., Shah, V. B., & Edelman, A. (2012). Julia: A Fast Dynamic Language for Technical Computing. Programming Languages; Computational Engineering, Finance, and Science.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F. a, Roskin, K. M., … Miller, W. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research*, *14*(4), 708–15. doi:10.1101/gr.1933104

Bowden, R. J., & Turkington, D. A. (1990). *Instrumental Variables*. Cambridge University Press.

Boyko, A. R., Williamson, S. H., Indap, A. R., Degenhardt, J. D., Hernandez, R. D., Lohmueller, K. E., … Bustamante, C. D. (2008). Assessing the evolutionary impact

of amino acid mutations in the human genome. *PLoS Genetics*, *4*(5), e1000083. doi:10.1371/journal.pgen.1000083

Bradley, B. J. (2008). Reconstructing phylogenies and phenotypes: a molecular view of human evolution. *Journal of Anatomy*, *212*(4), 337–53. doi:10.1111/j.1469-7580.2007.00840.x

Burkart, J. M., Allon, O., Amici, F., Fichtel, C., Finkenwirth, C., Heschl, A., … van Schaik, C. P. (2014). The evolutionary origin of human hyper-cooperation. *Nature Communications*, *5*, 4747. doi:10.1038/ncomms5747

Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M. T., Glanowski, S., … Clark, A. G. (2005a). Natural selection on protein-coding genes in the human genome. *Nature*, *437*(7062), 1153–7.

Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M. T., Glanowski, S., … Clark, A. G. (2005b). Natural selection on protein-coding genes in the human genome. *Nature*, *437*(7062), 1153–7. doi:10.1038/nature04240

Caporaso, J. G., Lauber, C. L., Costello, E. K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., … Knight, R. (2011). Moving pictures of the human microbiome. *Genome Biology*, *12*(5), R50. doi:10.1186/gb-2011-12-5-r50

Capra, J. A., Stolzer, M., Durand, D., & Pollard, K. S. (2013). How old is my gene? *Trends in Genetics : TIG*, *29*(11), 659–68. doi:10.1016/j.tig.2013.07.001

Casdagli, M., Eubank, S., Farmer, J. D., & Gibson, J. (1991). State space reconstruction in the presence of noise. *Physica D: Nonlinear Phenomena*, *51*(1-3), 52–98. doi:10.1016/0167-2789(91)90222-U

Chandrasekaran, V., & Jordan, M. I. (2013). Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(13), E1181–90. doi:10.1073/pnas.1302293110

Charlesworth, J., & Eyre-Walker, A. (2008). The McDonald-Kreitman test and slightly deleterious mutations. *Molecular Biology and Evolution*, *25*(6), 1007–15. doi:10.1093/molbev/msn005

Chen, F., Mackey, A. J., Vermunt, J. K., & Roos, D. S. (2007). Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PloS One*, *2*(4), e383. doi:10.1371/journal.pone.0000383

Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B., & Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science (New York, N.Y.)*, *311*(5765), 1283–7. doi:10.1126/science.1123061

Costanzo, M. J., Yabut, S. C., Zhang, H.-C., White, K. B., de Garavilla, L., Wang, Y., … Maryanoff, B. E. (2008). Potent, nonpeptide inhibitors of human mast cell tryptase. Synthesis and biological evaluation of novel spirocyclic piperidine amide derivatives. *Bioorganic & Medicinal Chemistry Letters*, *18*(6), 2114–21. doi:10.1016/j.bmcl.2008.01.093

Dayhoff, M. O., Schwartz, R. M., & Orcutt, B. C. (1978). A model of evolutionary change in proteins. In M. O. Dayhoff (Ed.), *Atlas of protein sequence and structure* (pp. 345–358). Nature Biomedical Research.

Deyle, E. R., Fogarty, M., Hsieh, C., Kaufman, L., MacCall, A. D., Munch, S. B., … Sugihara, G. (2013). Predicting climate effects on Pacific sardine. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(16), 6430–5. doi:10.1073/pnas.1215506110

Dixon, P. A., Milicich, M. J., & Sugihara, G. (1999). Episodic Fluctuations in Larval Supply. *Science*, *283*(5407), 1528–1530. doi:10.1126/science.283.5407.1528

Ebersberger, I., Strauss, S., & von Haeseler, A. (2009). HaMStR: profile hidden markov model based search for orthologs in ESTs. *BMC Evolutionary Biology*, *9*(1), 157. doi:10.1186/1471-2148-9-157

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, *7*(10), e1002195. doi:10.1371/journal.pcbi.1002195

Eden, E., Navon, R., Steinfeld, I., Lipson, D., & Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, *10*(1), 48. doi:10.1186/1471-2105-10-48

Eilertson, K. E., Booth, J. G., & Bustamante, C. D. (2012). SnIPRE: selection inference using a Poisson random effects model. *PLoS Computational Biology*, *8*(12), e1002806. doi:10.1371/journal.pcbi.1002806

Finch, C. E. (2010). Evolution in health and medicine Sackler colloquium: Evolution of the human lifespan and diseases of aging: roles of infection, inflammation, and nutrition. *Proceedings of the National Academy of Sciences of the United States of America*, *107 Suppl* (suppl_1), 1718–24. doi:10.1073/pnas.0909606106

Fisher, C. K., & Mehta, P. (2014). Identifying Keystone Species in the Human Gut Microbiome from Metagenomic Timeseries using Sparse Linear Regression.

Foster, J. A., & McVey Neufeld, K.-A. (2013). Gut-brain axis: how the microbiome influences anxiety and depression. *Trends in Neurosciences*, *36*(5), 305–12. doi:10.1016/j.tins.2013.01.005

Gajer, P., Brotman, R. M., Bai, G., Sakamoto, J., Schütte, U. M. E., Zhong, X., … Ravel, J. (2012). Temporal dynamics of the human vaginal microbiota. *Science Translational Medicine*, *4*(132), 132ra52. doi:10.1126/scitranslmed.3003605

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods Title. *Econometrica*, *37*(3), 424–438.

Haygood, R., Babbitt, C. C., Fedrigo, O., & Wray, G. A. (2010). Contrasts between adaptive coding and noncoding changes during human evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(17), 7853–7. doi:10.1073/pnas.0911249107

Henikoff, S. (1992). Amino Acid Substitution Matrices from Protein Blocks. *Proceedings of the National Academy of Sciences*, *89*(22), 10915–10919.

Hernandez, R. D. (2008). A flexible forward simulator for populations subject to selection and demography. *Bioinformatics (Oxford, England)*, *24*(23), 2786–7. doi:10.1093/bioinformatics/btn522

Heskamp, L., Meel-van den Abeelen, A., Katsogridakis, E., Panerai, R., Simpson, D., Lagro, J., & Claassen, J. (2013). Convergent cross mapping: a promising technique for future cerebral autoregulation estimation. *CEREBROVASCULAR DISEASES*, *35*, 15–16.

Hubisz, M. J., & Pollard, K. S. (2014). Exploring the genesis and functions of Human Accelerated Regions sheds light on their role in human evolution. *Current Opinion in Genetics & Development*, *29C*, 15–21. doi:10.1016/j.gde.2014.07.005

Hulsen, T., Huynen, M. A., de Vlieg, J., & Groenen, P. M. A. (2006). Benchmarking ortholog identification methods using functional genomics data. *Genome Biology*, *7*(4), R31. doi:10.1186/gb-2006-7-4-r31

Jordan, G., & Goldman, N. (2012). The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Molecular Biology and Evolution*, *29*(4), 1125–39. doi:10.1093/molbev/msr272

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Research*, *12*(4), 656–64. doi:10.1101/gr.229202. Article published online before March 2002

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, a. D. (2002). The Human Genome Browser at UCSC. *Genome Research*, *12*(6), 996–1006. doi:10.1101/gr.229102

Khoruts, A., & Weingarden, A. R. (2014). Emergence of fecal microbiota transplantation as an approach to repair disrupted microbial gut ecology. *Immunology Letters*. doi:10.1016/j.imlet.2014.07.016

Kosiol, C., Vinar, T., da Fonseca, R. R., Hubisz, M. J., Bustamante, C. D., Nielsen, R., & Siepel, A. (2008). Patterns of positive selection in six Mammalian genomes. *PLoS Genetics*, *4*(8), e1000144. doi:10.1371/journal.pgen.1000144

Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Machine Learning*, *51*(2), 181–207.

Kuzniar, A., van Ham, R. C. H. J., Pongor, S., & Leunissen, J. a M. (2008). The quest for orthologs: finding the corresponding gene across genomes. *Trends in Genetics : TIG*, *24*(11), 539–51. doi:10.1016/j.tig.2008.08.009

Liu, Y., Schmidt, B., & Maskell, D. L. (2010). MSAProbs: multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities. *Bioinformatics (Oxford, England)*, *26*(16), 1958–64. doi:10.1093/bioinformatics/btq338

Maddison, W. P., & Knowles, L. L. (2006). Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology*, *55*(1), 21–30. doi:10.1080/10635150500354928

Maher, M. C., & Hernandez, R. D. (2013). A MOSAIC of methods: Improving ortholog detection through integration of algorithmic diversity. Populations and Evolution; Quantitative Methods.

Mashiyama, S. T., Malabanan, M. M., Akiva, E., Bhosle, R., Branch, M. C., Hillerich, B., … Babbitt, P. C. (2014). Large-scale determination of sequence, structure, and function relationships in cytosolic glutathione transferases across the biosphere. *PLoS Biology*, *12*(4), e1001843. doi:10.1371/journal.pbio.1001843

Massingham, T., & Goldman, N. (2005). Detecting amino acid sites under positive selection and purifying selection. *Genetics*, *169*(3), 1753–62. doi:10.1534/genetics.104.032144

McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, *351*(6328), 652–4.

McEntyre, .J, & Ostell, J. (Eds.). (2002). *The NCBI Handbook*. Bethesda, MD: National Center for Biotechnology Information.

Messer, P. W., & Petrov, D. A. (2013). Frequent adaptation and the McDonald-Kreitman test. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(21), 8615–20. doi:10.1073/pnas.1220835110

Nielsen, R., Bustamante, C., Clark, A. G., Glanowski, S., Sackton, T. B., Hubisz, M. J., … Cargill, M. (2005a). A scan for positively selected genes in the genomes of

humans and chimpanzees. *PLoS Biology*, *3*(6), e170. doi:10.1371/journal.pbio.0030170

Nielsen, R., Bustamante, C., Clark, A. G., Glanowski, S., Sackton, T. B., Hubisz, M. J., … Cargill, M. (2005b). A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biology*, *3*(6), e170. doi:10.1371/journal.pbio.0030170

Nielsen, R., Hubisz, M. J., Hellmann, I., Torgerson, D., Andrés, A. M., Albrechtsen, A., … Clark, A. G. (2009). Darwinian and demographic forces affecting human protein coding genes. *Genome Research*, *19*(5), 838–49. doi:10.1101/gr.088336.108

Preuss, T. M. (2011). The human brain: rewired and running hot. *Annals of the New York Academy of Sciences*, *1225 Suppl*, E182–91. doi:10.1111/j.1749-6632.2011.06001.x

Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., … Ostell, J. M. (2014). RefSeq: an update on mammalian reference sequences. *Nucleic Acids Research*, *42*(Database issue), D756–63. doi:10.1093/nar/gkt1114

Pruitt, K. D., Harrow, J., Harte, R. A., Wallin, C., Diekhans, M., Maglott, D. R., … Lipman, D. (2009). The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Research*, *19*(7), 1316–23. doi:10.1101/gr.080531.108

Pryszcz, L. P., Huerta-Cepas, J., & Gabaldón, T. (2011). MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. *Nucleic Acids Research*, *39*(5), e32. doi:10.1093/nar/gkq953

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., … Finn, R. D. (2012). The Pfam protein families database. *Nucleic Acids Research*, *40*(Database issue), D290–301. doi:10.1093/nar/gkr1065

Remm, M., Storm, C. E., & Sonnhammer, E. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology*, *314*(5), 1041–52. doi:10.1006/jmbi.2000.5197

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., … Sabeti, P. C. (2011). Detecting novel associations in large data sets. *Science (New York, N.Y.)*, *334*(6062), 1518–24. doi:10.1126/science.1205438

Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, *53*(1-2), 131–147. doi:10.1016/0025-5564(81)90043-2

Rokach, L. (2009). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*(1-2), 1–39. doi:10.1007/s10462-009-9124-7

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*, 72–101.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)*, *30*(9), 1312–3. doi:10.1093/bioinformatics/btu033

Stamatakis, A., & Alachiotis, N. (2010). Time and memory efficient likelihood-based tree searches on phylogenomic alignments with missing data. *Bioinformatics (Oxford, England)*, *26*(12), i132–9. doi:10.1093/bioinformatics/btq205

Stolzer, M., Lai, H., Xu, M., Sathaye, D., Vernot, B., & Durand, D. (2012). Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics (Oxford, England)*, *28*(18), i409–i415. doi:10.1093/bioinformatics/bts386

Stone, M. (1977). An Asymptotic Equivalence of Choice of Model by Cross-validation and Akaike's Criterion. *Journal of the Royal Statistical Society*. Retrieved March 30, 2014, from http://www.jstor.org/discover/10.2307/2984877?uid=3739560&uid=2134&uid=2&uid=70&uid=4&uid=3739256&sid=21103766217637

Sugihara, G. (1994). Nonlinear Forecasting for the Classification of Natural Time Series. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *348*(1688), 477–495. doi:10.1098/rsta.1994.0106

Sugihara, G., May, R., Ye, H., Hsieh, C., Deyle, E., Fogarty, M., & Munch, S. (2012). Detecting causality in complex ecosystems. *Science (New York, N.Y.)*, *338*(6106), 496–500. doi:10.1126/science.1227079

Sukumaran, J., & Holder, M. T. (2010). DendroPy: a Python library for phylogenetic computing. *Bioinformatics (Oxford, England)*, *26*(12), 1569–71. doi:10.1093/bioinformatics/btq228

Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS One*, *6*(7), e21800. doi:10.1371/journal.pone.0021800

Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, *34*(Web Server issue), W609–12. doi:10.1093/nar/gkl315

Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and Testing Dependence by Correlation of Distances. *The Annals of Statistics*, *35*(6), 2769–2794.

The Chimpanzee Sequencing and Analysis Consortium. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, *437*(7055), 69–87. doi:10.1038/nature04072

Torgerson, D. G., Boyko, A. R., Hernandez, R. D., Indap, A., Hu, X., White, T. J., … Clark, A. G. (2009). Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. *PLoS Genetics*, *5*(8), e1000592. doi:10.1371/journal.pgen.1000592

Trivedi, N. N., Tong, Q., Raman, K., Bhagwandin, V. J., & Caughey, G. H. (2007). Mast cell alpha and beta tryptases changed rapidly during primate speciation and evolved from gamma-like transmembrane peptidases in ancestral vertebrates. *Journal of Immunology (Baltimore, Md. : 1950)*, *179*(9), 6072–9.

Van der Laan, M. J., & Gruber, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *The International Journal of Biostatistics*, *6*(1), Article 17. doi:10.2202/1557-4679.1181

Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, *6*, Article25. doi:10.2202/1544-6115.1309

Vanderweele, T. J., & Arah, O. A. (2011). Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology (Cambridge, Mass.)*, *22*(1), 42–52. doi:10.1097/EDE.0b013e3181f74493

Varki, A. (2012). Nothing in medicine makes sense, except in the light of evolution. *Journal of Molecular Medicine (Berlin, Germany)*, *90*(5), 481–94. doi:10.1007/s00109-012-0900-5

Vujkovic-Cvijin, I., Dunham, R. M., Iwai, S., Maher, M. C., Albright, R. G., Broadhurst, M. J., … McCune, J. M. (2013). Dysbiosis of the gut microbiota is associated with HIV disease progression and tryptophan catabolism. *Science Translational Medicine*, *5*(193), 193ra91. doi:10.1126/scitranslmed.3006438

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164. doi:10.1093/nar/gkq603

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., … Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, *42*(Database issue), D1001–6. doi:10.1093/nar/gkt1229

Williamson, S., Fledel-Alon, A., & Bustamante, C. D. (2004). Population genetics of polymorphism and divergence for diploid selection models with arbitrary dominance. *Genetics*, *168*(1), 463–75. doi:10.1534/genetics.103.024745

Wolpert, D. H., & Macready, W. G. (1997). No Free Lunch Theorems For Optimization. *IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION*, *1*(1), 67–82.

Wu, J., Sinfield, J. L., Buchanan-Wollaston, V., & Feng, J. (2009). Impact of environmental inputs on reverse-engineering approach to network structures. *BMC Systems Biology*, *3*, 113. doi:10.1186/1752-0509-3-113

Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews. Genetics*, *13*(5), 329–42. doi:10.1038/nrg3174

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–91. doi:10.1093/molbev/msm088

Yu, C., Zavaljevski, N., Desai, V., & Reifman, J. (2011). QuartetS: a fast and accurate algorithm for large-scale orthology detection. *Nucleic Acids Research*, *39*(13), e88. doi:10.1093/nar/gkr308

Zhao, S., Sakai, A., Zhang, X., Vetting, M. W., Kumar, R., Hillerich, B., … Jacobson, M. P. (2014). Prediction and characterization of enzymatic activities guided by sequence similarity and genome neighborhood networks. *eLife*, *3*. doi:10.7554/eLife.03275

Chapter 2 Rock, Paper, Scissors: Harnessing complementarity in ortholog detection methods improves comparative genomic inference.

Introduction

Orthologs are genes that derive from a common ancestral gene, but which have diverged from one another through speciation. This is in contrast to paralogs, which arise through gene duplication within a given genome. It is common in comparative genomics and phylogenetics to extract evolutionary information about a particular gene from its alignment with orthologous sequences. To enable this analysis, orthologs must first be inferred, making ortholog detection (OD) an indispensible first step in a variety of phylogenetic inference tasks (Ciccarelli et al., 2006; Yandell & Ence, 2012).

In general, existing OD methods can be classified as tree-based, graph-based, or a hybrid of the two (Altenhoff & Dessimoz, 2012). Tree-based methods may use reconciliation techniques between gene and species trees or may rely on the gene tree alone. Graph-based methods can employ a variety of metrics to quantify similarity between sequences, including BLAST scores or sequence identity. Information about the conserved gene neighborhood may also be included in this context. Techniques such as Markov clustering may then be applied to create orthologous groups, or one may simply define clusters based on a graph's existing connections (Kuzniar, van Ham, Pongor, & Leunissen, 2008).

Unfortunately, the few benchmarking studies that have sampled broadly from this methodological diversity have provided equivocal results. Although there are general patterns in relative effectiveness between methods, performance is highly context-dependent and does not always favor more sophisticated approaches (Altenhoff & Dessimoz, 2009a; Chen, Mackey, Vermunt, & Roos, 2007; Hulsen, Huynen, de Vlieg, & Groenen, 2006). This is discouraging from the point of view of identifying a single best OD method, but it also suggests a new and relatively facile avenue for methodological improvement. By harnessing differences between OD methods, a wide variety of algorithms may play complementary roles within a cooperative inference framework.

We begin our analysis of orthologs of human genes with a comprehensive comparison of four popular and methodologically distinct OD methods: 1.) MultiParanoid, a reciprocal-BLAST plus Markov clustering method (Alexeyenko, Tamas, Liu, & Sonnhammer, 2006); 2.) TBA, a synteny-based aligner used to produce UCSC's MultiZ alignments (Blanchette et al., 2004); 3.) six-frame translated BLAT, a fast, approximately-scored protein query approach that does not rely on predicted proteomes (W James Kent, 2002); and 4.) OMA, a well-established tree-based method (Altenhoff, Schneider, Gonnet, & Dessimoz, 2011). Applying these methods to OD in a range of primates and closely related mammals, we demonstrate that methodological performance varies widely by species and appears to depend critically on genome quality.

Next, we characterize the striking performance gains yielded by combining these methods. This is demonstrated using sequence identity, phylogenetic tree concordance, and hidden markov model-based functional agreement. This improvement in alignment quality translates to higher estimated levels of overall conservation, while at the same time, detecting up to 180% more positively selected sites. We close by highlighting a novel PSS near the active site of TPSAB1, an enzyme linked to asthma and irritable bowel disease.

The implementation of this novel approach for the integration of diverse ortholog detection methods is presented as the software tool, MOSAIC, or **M**ultiple **O**rthologous **S**equence **A**nalysis and **I**ntegration by **C**luster optimization. MOSAIC is implemented as a well-documented python package that can be installed using easy_install bio-mosaic from the command-line. MOSAIC alignments, source code, and full documentation are available at http://pythonhosted.org/bio-MOSAIC.

Materials and methods

Retrieval of orthologs

For each human consensus coding sequence (version GRCh37.p9), we sought to retrieve orthologs for chimp, gorilla, orangutan, rhesus macaque, marmoset, bushbaby, cat, cow, and horse using four methodologically diverse methods: 1) MultiParanoid (Alexeyenko et al., 2006); 2.) TBA (Blanchette et al., 2004); 3.) six-frame translated BLAT (W James Kent, 2002); and 4.) OMA (Altenhoff et al., 2011). Genomic data was retrieved for the following genome builds:

Genome	Version	Release
Chimp	panTro4	Feb-11
Gorilla	gorGor3.1	May-10
Orangutan	ponAbe2	Jul-11
Rhesus macaque	rheMac3	Oct-10
Marmoset	calJac3	Mar-09
Bushbaby	otoGar3	May-11
Cat	felCat5	Sep-11
Cow	bosTau7	Oct-11
Horse	equCab2	Sep-07

For MultiParanoid (Alexeyenko et al., 2006), an all-versus-all blast search was run using the following command structure:

```
blastp -db $blastdatabase -query [query file] -out [output file] -evalue .01 -num_threads [number of threads] -outfmt 6 -db_soft_mask 21 -word_size 3 -use_sw_tback
```

From this output, ortholog predictions were produced using the standard MultiParanoid protocol.

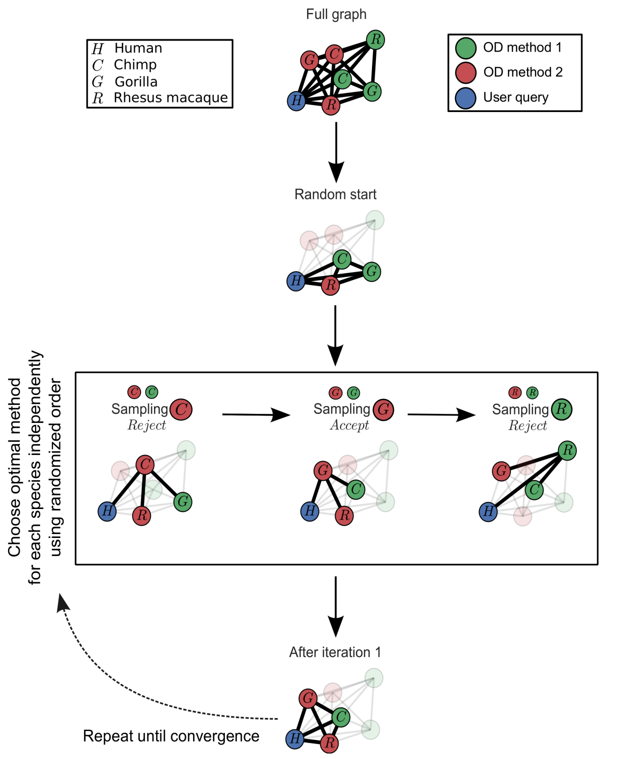
For BLAT (W James Kent, 2002), genomes for each species of interest were downloaded from the NCBI Entrez Genome database (McEntyre & Ostell, 2002). Queries were conducted using the following command structure:

```
blat -q=prot -t=dnax -minIdentity=70 –extendThroughN [genome file] [query file] [output file]
```

In the case of MultiZ (Blanchette et al., 2004), CCDS orthologs were downloaded directly from the UCSC genome browser (W. J. Kent et al., 2002). For OMA (Altenhoff et al., 2011), ortholog predictions were downloaded from omabrowser.org

(December 2012 release). For genes with more than one CCDS, orthologs were mapped to each analyzed transcript. Finally, ortholog predictions from metaPhOrs (Pryszcz, Huerta-Cepas, & Gabaldón, 2011) were retrieved from release v201009 (June 2012).

To remove possibly spurious orthologs, proposals from each method were then filtered according to a species-specific sequence identity cutoff, as described below.



Box 1. A schematic of the sequence selection algorithm. Steps: 1.) Construct graph. 2.) Choose the sequence from a random OD method for each species. 3.) Iterate through species. For each species, pick the orthologs with highest similarity to the current best choices for all other species. 4.) Return current best choices if no changes are made after iterating through all species. 5.) To find global optimum, repeat steps 1-4 with random sampling paths.

MOSAIC: OD integration as cluster optimization

MOSAIC provides a highly flexible, graph-based framework for integrating diverse OD methods. For a given reference sequence, proposal orthologs are conceptualized as nodes in a graph, connected with edges weighted according to the pairwise similarity between sequences (Box 1). The task of OD integration is then to choose proposal orthologs from each species such that a chosen measure of intra-cluster similarity is optimized.

MOSAIC optimizes (weighted) pairwise similarities

To begin, MOSAIC calculates pairwise similarities between all orthologs from different species. Percent identity- and blast-based similarity metrics are provided by default, but user-defined similarity metrics are also accepted. These similarity scores define edge weights, which are used to construct a graph such as the one presented at the top of Box 1. Once this full graph is constructed, it is highly recommended that it be quality filtered using user-specified similarity cutoffs. This step is necessary to minimize the effect of gene loss, duplication, etc. Once the graph is cleaned, MOSAIC then chooses at most one proposal ortholog from each species so that the overall pairwise similarity between accepted sequences is optimized.

To accommodate user priorities, pairwise similarities can be weighted such that sequences from different species contribute unequally to the total similarity score. For uniform weights, this is equivalent to maximizing the average pairwise similarity. In the case where only similarity to a reference sequence is of interest, this reduces to simply accepting the ortholog for each species that is most similar to the reference.

Optimization is carried out using cyclic coordinate descent

For m OD methods and s species, there are up to m^s possible integrated alignments. In the case analyzed in this paper, $m=4$ and $s=10$. This translates to over a million possible integrated alignments for each of the ~25,000 reference sequences considered. It is clear to see from this example that an exhaustive optimization becomes quickly infeasible. Therefore, MOSAIC choses optimal clusters using cyclic coordinate descent (CCD), an efficient non-derivative optimization algorithm (Bertsekas, 1999).

In Box 1, we illustrate the way CCD functions in the context of MOSAIC. After building the full graph that includes all orthologous sequences, random orthologs from each species are chosen as the current best. MOSAIC then loops through the species of interest in a random order. For each species, MOSAIC choses the sequence that optimizes cluster tightness given the current best sequences for all other species. This process is repeated until no further improvements can be made to cluster tightness. Finally, since CCD is prone to finding local rather than global optima, this entire process is repeated multiple times with random starting points and sampling paths.

Scoring and optimization procedures for this study

For the alignments presented here, we consider a protein set with relatively low levels of evolutionary divergence. As described earlier, we chose percent identity as our metric for sequence divergence. Note that several other popular scoring functions are implemented in MOSAIC. For more distantly related species, the application of scoring matrices (Dayhoff, Schwartz, & Orcutt, 1978; Henikoff, 1992) or Hidden Markov Models (Ebersberger, Strauss, & von Haeseler, 2009) may be preferable. To reduce computational costs related to pairwise alignment, we considered only similarities between orthologs and the human target sequence. The optimization procedure was then equivalent to choosing, for each species, the ortholog among all methods that is most similar to the human sequence. This approach corresponds to the arguments edgefunc='perID', optrule='pairwise' when calling the Mosaic constructor in mosaic.py (see: http://pythonhosted.org/bio-MOSAIC/Module.html).

Example: measuring similarity

Percent identity was then calculated as the percent of sites in the human sequence that were identical in the orthologous sequence. For example, the hypothetical sequence below would be scored as 71% identical (5/7), since there are 2 mismatches between the seven sites present in the human sequence and the character to which those sites are aligned in the chimp sequence (sites where the human sequence has been deleted or the outgroup has an insertion are ignored):

Human A W V A - T F D

Chimp - W V R Y T F D

Filtering putatively non-orthologous sequences

All ortholog detection methods produce false positives. For example, this can result when a gene deletion on one lineage means that no true ortholog exists in a given species. Typically, these issues are dealt with through rigorous filtering of input alignments. The intuition is that by applying a stringent sequence similarity filter, we can remove the vast majority of evolutionarily unrelated genes. MOSAIC also employs this filtering approach prior to integration, guaranteeing that only credibly putatively orthologous sequences are included in the analysis. Cutoffs were chosen considering the known level of genome-wide divergence between human and the species of interest, as well as the overall distributions of percent identity between putative orthologs in the two species. These cutoffs were as follows: chimp: 82%, gorilla: 77%, orangutan: 75%, rhesus macaque: 73%. A cutoff of 70% was employed for marmoset, bushbaby, cat, cow, and horse.

A note on paralogs

For any query protein, there is also a risk that a related gene in another genome has not maintained the same function and so provides inapplicable evolutionary information. This divergence in function would be expected to increase sequence divergence, and so in many cases could also be removed by suitably stringent sequence similarity filters. Paralogs inject additional bias if, compared to the query protein, the most functionally similar of the set is not the most similar at the sequence level. While this is possible, it is the exception and not the rule under reasonable models of evolution. Indeed, this expectation has been validated by experimental data from several model systems (Mashiyama et al., 2014; Zhao et al., 2014). For this reason MOSAIC does not exclude putatively orthologous sequences that have paralogs in the source genomes. We will show that this decision allows us to capture more putative orthologs while simultaneously improving ortholog quality by all commonly used metrics.

In summary, MOSAIC is adapted to producing MSAs that are functionally informative at the site-level. For other applications, researchers may wish to infer genomic events such a gene loss, duplication, horizontal gene transfer, and/or incomplete lineage sorting (e.g. Capra et al. 2013). This involves jointly examining functionally diverged paralogous groups alongside their corresponding orthologs. This task generally requires a combination of tools such as MultiParanoid (to infer paralogs; Remm et al. 2001), RaxML (to build gene and species trees; Stamatakis 2014), and Notung (reconcile gene trees with species trees and infer evolutionary events; Stolzer et al. 2012). For applications such as this, MOSAIC alignments can still be leveraged to guarantee the presence of relevant sequences. Likewise, reconstructed evolutionary histories can be used to flag, among tens of thousands of automatically generated MOSAIC alignments, those exceptional cases that could benefit most from manual inspection.

Multiple sequence alignment

Retrieved sequences were jointly aligned to query proteins using MSAprobs (Liu, Schmidt, & Maskell, 2010), a multithreaded aligner with better performance benchmarks than many top aligners, including ClustalW, MAFFT, MUSCLE, ProbCons, and Probalign (Liu et al., 2010). Importantly, MSAprobs has the further advantage of providing, for each column of an alignment, dependable estimates of the confidence of the alignment at the site.

Quality assessment

Sequence identity

MOSAIC optimizes pairwise sequence similarity. In this example, sequence identity is used as the similarity measure, and pairwise similarities are weighted such that only concordance with the human reference sequence is considered. To achieve greater separation between metrics used for optimization and assessment, comparisons of sequence identity were performed in the context of the full multiple sequence alignments (MSAs). We believe this choice is sensible because it is the quality of the MSA that is of primary importance to many downstream phylogenetic inference tasks. In addition, this approach allows us to indirectly incorporate information about intra-cluster similarity. As an MSA incorporates increasingly divergent sequences, performance relative to pairwise alignments is expected to progressively degrade.

Tree concordance

For each MSA, gene trees were built using RAxML (Stamatakis & Alachiotis, 2010). An unweighted Robinson-Foulds (RF) distance (Robinson & Foulds, 1981) was then calculated between each gene tree and the known species tree using the python module dendropy (Sukumaran & Holder, 2010). Briefly, the unweighted RF distance counts the number of operations required to transform one tree into the other. This quantity is equal to the total number of splits that are present in one tree but not the other. To normalize for variations in tree size, we then divided this distance by the sum of the total number of splits in the gene and species trees (Yu, Zavaljevski, Desai, &

Reifman, 2011). To summarize the genome-wide distribution of normalized RF distances, we took the area under the curve of the cumulative distribution function (CDF). This was limited to values less than 0.4, since beyond this value there is little difference between the observed curves (see fig. SA-5). This metric is superior to, e.g. calculating the proportion of genes below a given threshold because it up-weights smaller RF distances as opposed to, in effect, using non-zero uniform weights below the cutoff value.

Functional concordance

Profile HMMs were downloaded from the PfamA protein families database (Punta et al., 2012). Each sequence was then annotated using the top scoring function retrieved by querying that sequence against the database of all PfamA protein family HMMs. This search was conducted using HMMER3 (Eddy, 2011). Functional concordance was then measured as a binary quantity, corresponding to whether or not a putative orthologous sequence had the same inferred function as its cognate human sequence. It is important to note that not all PfamA HMMs are functionally validated. In cases where experimental validation is unavailable, these HMMs provide a family-specific scoring function that nevertheless yields information not contained in naïve sequence identity measures.

Evolutionary analysis

Gene-level conservation

Alignments were analyzed using Phylogenetic Analysis by Maximum Likelihood (PAML) (Yang, 2007). For each alignment three models were fit: 1.) a neutral model where dN/dS is fixed at one, 2.) a conservation model where dN/dS is less than or equal to one, and 3.) a positive selection model where some fraction of the sequence is fit under the conservation model, while another dN/dS parameter is estimated freely for the remainder of the sequence. Since evolutionary models are not in general nested, we performed model selection via the popular Akaike Information Criterion (AIC), a method that penalizes a model's fit by its number of included variables (Akaike, 1973) and is asymptotically equivalent to maximizing the model's predictive performance on unseen data (Stone, 1977).

Despite rigorous model selection procedures, in rare cases PAML may estimate very high levels of selection over a tiny proportion of a given sequence (even a single site), leading to greatly inflated average levels of dN/dS. To reduce the influence of outlying estimates of selection, all dN/dS values greater than 3 were excluded for the analysis. For all methods, this corresponded to less than .05% of all sequences.

Site-level positive selection

The program Sitewise Likelihood Ratio (SLR) (Massingham & Goldman, 2005) was used to estimate the number of positively selected sites in each sequence. To eliminate false positives due to poorly aligned sites, we filtered out all sites estimated by MSAprobs to be aligned to less than 95% confidence. All included positively selected sites estimated at 95% confidence or greater by SLR were included in the subsequence comparison.

Concordance between positively selected sites

To assess agreement in positively selected sites (PSS), we calculated the degree of overlap between PSS from all pairs of methods. This was calculated as the size of the genome-wide intersection between sites divided by the union of said sites.

Mapping positively selected sites onto three-dimensional structures

We leveraged UniProt mapping files (http://www.uniprot.org/docs/pdbtosp; accessed 9/30/14) to determine which proteins had a relevant structure in the Protein Data Bank (PDB; Berman 2000). We then aligned sequences between PDB structures and candidate genes to determine the degree of coverage and to obtain a mapping between residues. We found 2003 genes for which there was a structure with greater than 70% coverage. Of these, 787 had results from all five ortholog detection methods. Reasons for missing data comprise absence from source data and lack of convergence in the PSS calculation. Within this set of 787 genes, 76 proteins had PSS from MOSAIC that were not found with any of the component methods. From this point, the example of TPSAB1 was quickly identified by manual inspection. We then downloaded PDB structure 2ZEC to visualize the location of positively selected sites. To validate sequences used in the analysis, we blasted each ortholog against the human SwissProt database. This confirmed that TPSAB1 was the most similar human protein in each case.

Results and Discussion

Ortholog detection methods frequently outperform one another

To motivate OD integration, we will begin with a comprehensive comparison of four popular, methodologically diverse OD methods. In figure 2-1, we show the head-to-head performances of these different methods for a range of primates and closely related mammals. Performance is assessed using alignments between all human consensus coding sequences (CCDS) (Pruitt et al., 2009) and their corresponding orthologs from each method. For each possible ortholog (defined by human target sequence and species of origin), we examine whether sequence identity to human is at least five percentage points higher for one method versus another. We otherwise consider the two methods to be tied. By this metric, one method significantly outperforms another 38 to 45% of the time. Importantly, no method uniformly outperforms all others, underlining the complementarity of the chosen algorithms.

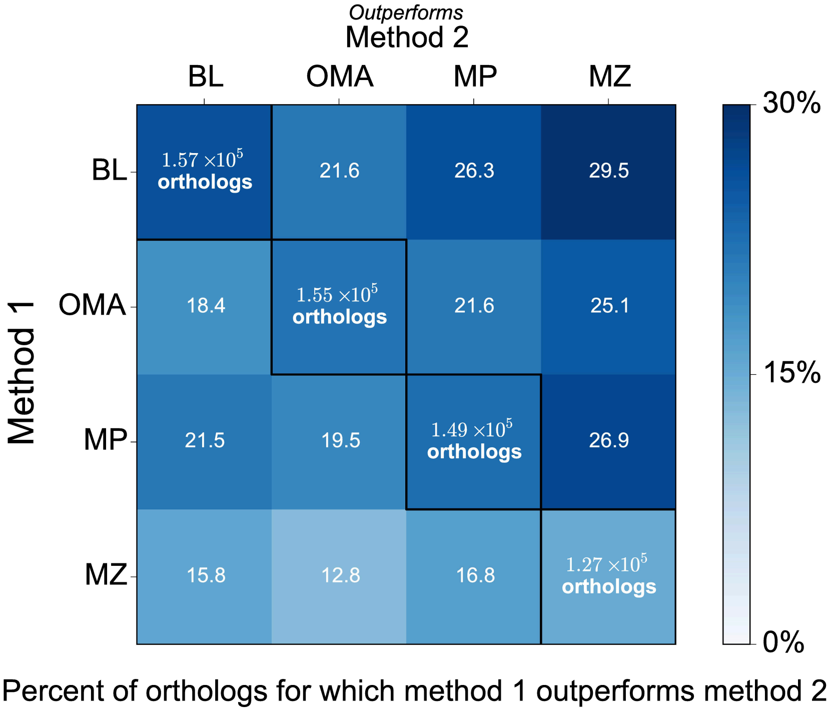


Figure 2-1. Comparison of sequence identity levels between methods. *Heat map of the percent of orthologs for which MultiParanoid (MP), OMA (OMA), BLAT (BL) and MultiZ (MZ) outperform one another. Performance is based on percent identity of each method's orthologs to the human sequence. One method is considered to outperform another method if it improves percent identity by at least five percentage points. Text in diagonal cells shows the number of orthologs identified by each method, colored by the percent of orthologs for which a given method outperforms all the others.*

Combining multiple sequence alignments with MOSAIC

It is well-known in theory (Wolpert & Macready, 1997) and in practice (van der Laan & Gruber, 2010) that the comparative performance of competing statistical inference algorithms often varies by context. Rather than search for a single best algorithm, researchers have sought to integrate a variety of methods in order to reap the benefits of methodological complementarity (Chandrasekaran & Jordan, 2013; Rokach, 2009; van der Laan, Polley, & Hubbard, 2007). As might be expected, the gains yielded by this approach generally scale with the quality of the individual methods integrated, the number of methods included, and, importantly, the diversity of the comprised algorithms (Kuncheva & Whitaker, 2003).

Having observed the complementarity between OD methods, we sought to develop a structure for the automatic integration of methodologically distinct OD methods such as those described above. We term this framework MOSAIC, or

Multiple **O**rthologous **S**equence **A**nalysis and **I**ntegration by **C**luster optimization. MOSAIC allows for the flexible integration of diverse OD methods through the application of standard or user-defined metrics of sequence similarity and ortholog cluster quality. Using the specified similarity metrics, clusters of proposed orthologs are built. These orthologs are then adopted or rejected in order to optimize cluster completeness and quality (e.g., similarity to a reference sequence or average pairwise similarity).

Having presented a schematic of the algorithm itself in box 1, we provide in figure 2-2 a view of example inputs and output MSAs. These are illustrations of real alignments for carbonic anhydrase 12 (CA12), an enzyme critical to a number of biological functions including the formation of bone, saliva, and gastric acid (Pruitt et al., 2014). MSA columns that are aligned to below 95% confidence are displayed in red. Orthologs that were not returned for a given species are denoted with a horizontal black bar. Those that were filtered using pre-integration sequence identity cutoffs (see Materials & Methods) are indicated with grey bars. Note that, just as when employing a single ortholog detection method, this filtering step is critical to guaranteeing alignment quality.

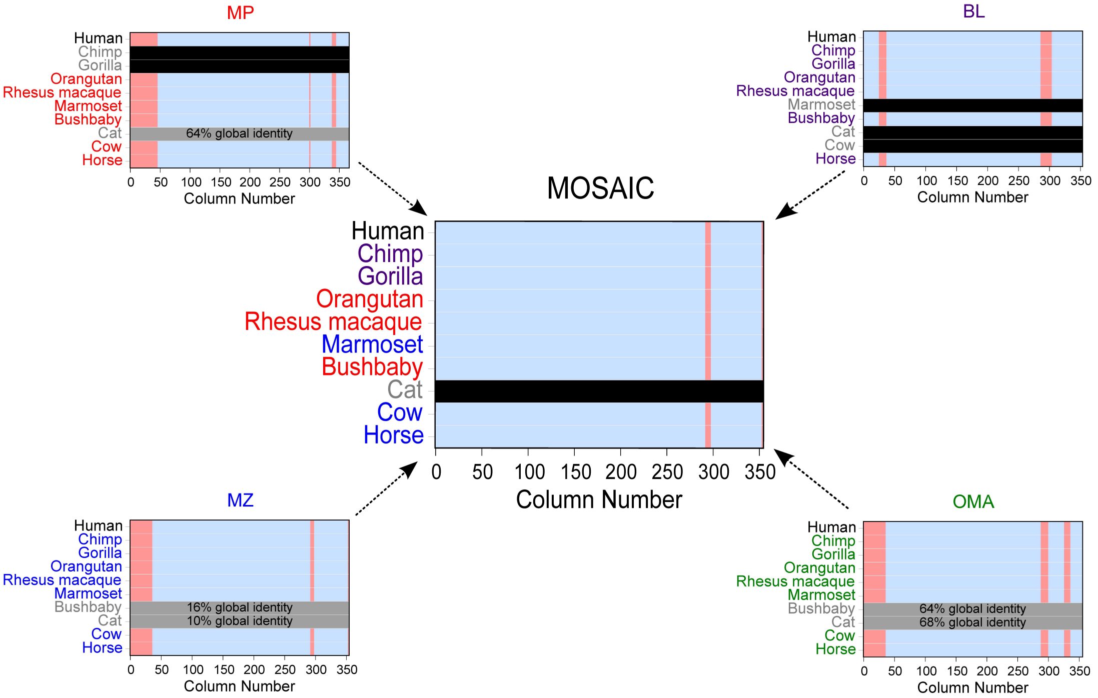


Figure 2-2 Illustration of integration process for carbonic anhydrase 12. *MSA columns that are aligned to below 95% confidence are displayed in red. Orthologs that were not returned for a given species are denoted with a horizontal black bar. Those that were filtered using pre-integration sequence identity cutoffs are indicated with grey bars with the global percent identity included. Species name label colors denote the species of origin for orthologs in the MOSAIC alignment.*

Combining methods increases the number of included sequences

The gains afforded by MOSAIC vary by species and increase with the number of methods that are included (fig. 2-3A). When all four component methods are included, MOSAIC more than quintuples the number of alignments for which all species are present (fig. 2-3B). We observe in fig. 2-3A that the largest improvements are seen for gorilla, bushbaby, and cat. Importantly, orthologs for each of these three species are rescued by different methods (OMA for gorilla, MultiParanoid for bushbaby, and Blat for cat. See fig. SA-3 for further details). In the sections that follow, we will demonstrate that MOSAIC captures these additional sequences while simultaneously improving functional-, phylogenetic-, and sequence identity-based measures of ortholog quality.

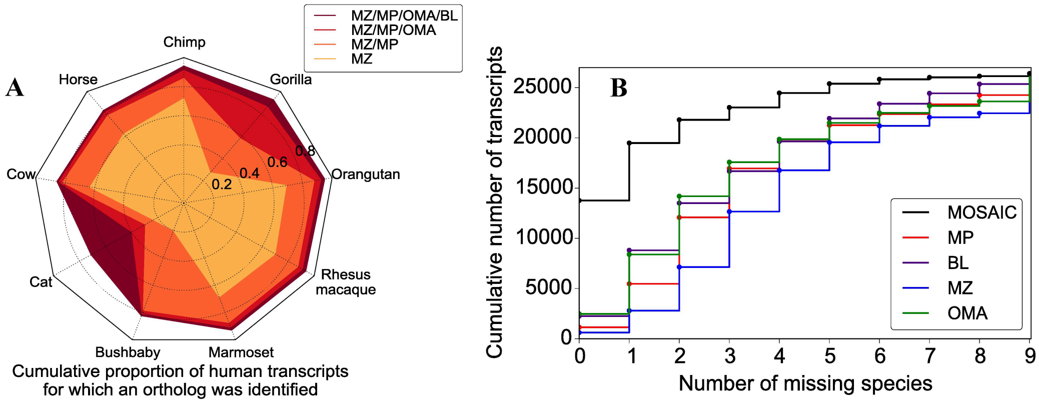


Figure 2-3 OD power and the effect of pooling methods. *A.) The cumulative proportion of human transcripts for which an ortholog was detected, stratified by species. Envelopes illustrate results from pooling an increasing number of methods. B.) The cumulative number of human transcripts as a function of the maximum number of missing species allowed.*

MOSAIC improves functional-, phylogenetic-, and sequence identity-based measures of ortholog quality

MOSAIC improves sequence identity

MOSAIC achieves massive gains in the number of retrieved orthologs while slightly improving average levels of sequence identity. Though MOSAIC directly optimizes sequence identity, this result is non-circular for two reasons. First, average levels of sequence identity could be reduced by preferentially adding sequences from the lower end of the sequence identity distribution. This would be consistent with a scenario in which most methods correctly inferred that a gene was deleted on a particular lineage. Second, MOSAIC optimizes sequence identity measured from pairwise global alignments. In the validation phase, we calculated this metric in the context of the full MSA. We believe this choice is sensible because it is the quality of the MSA that is of primary importance to many downstream phylogenetic inference tasks. In addition, this approach allows us to indirectly incorporate information about intra-cluster similarity. As an MSA incorporates increasingly divergent sequences, performance relative to pairwise alignments is expected to progressively degrade.

MOSAIC improves functional concordance

We employed profile HMMs from the Protein Families Database A (PfamA) (Punta et al., 2012) and HMMER3 (Eddy, 2011) to ascertain putative functional concordance between proposed orthologs and the human CCDS of interest. PfamA builds HMMs via curated alignments of small numbers of representative members from each protein family. It is important to note that not all PfamA HMMs are functionally validated. In cases where experimental validation is unavailable, these HMMs provide a family-specific scoring function that yields information not contained in naïve sequence identity measures.

With HMMER3, we queried protein sequences against all PfamA protein family profiles, annotating each protein according to its top protein family hit. This allowed for an ascertainment of functional concordance that is more comprehensive than relying on gene-by-gene annotation across species, while retaining many of the advantages of manual curation where it exists. This assessment reveals that, for the set of orthologous sequences proposed by all methods, MOSAIC provides levels of functional concordance that are slightly better than the best performing component method (fig. 2-4). Gains are particularly large for gorilla, bushbaby, and cat orthologs (fig. SA-4).

MOSAIC improves phylogenetic concordance

Phylogenetic concordance was ascertained by calculating the normalized, unweighted Robinson-Foulds (RF) distance (Robinson & Foulds, 1981) between gene trees and the established species tree (Altenhoff & Dessimoz, 2009b). This metric is equal to the sum of the number of splits in one tree that are not present in the other, scaled by the total number of splits present across the two trees. Accordingly, larger RF distances correspond to worse agreement between gene and species trees. On a gene-by-gene basis, this metric should be interpreted with caution, since post-speciation admixture and incomplete lineage sorting can lead to true discordance between the species tree and the phylogenetic history of a particular gene (Maddison & Knowles, 2006). However, at the level of the genome, higher concordance between gene trees and the known speciation process strongly suggests a relative improvement in OD.

Figure 2-4 presents a comparison of genome-wide phylogenetic concordance (see Materials & Methods for details on this metric). MultiZ performs the best of any individual method, likely due to its utilization of syntenic information. Surprisingly, the tree-based OD method, OMA, exhibits the worst performance according to this tree-based metric. MOSAIC, on the other hand, provides significant performance gain over all component methods, including a 59% increase in phylogenetic concordance compared to OMA.

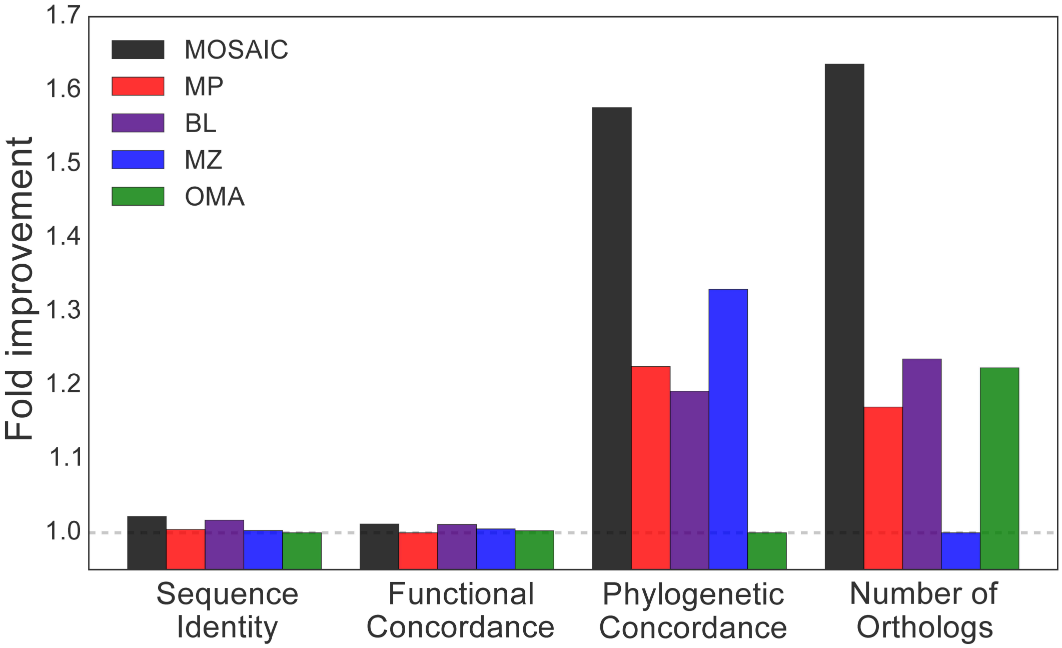


Figure 2-4. MOSAIC improves alignment quality. *We show the fold improvement of each method over the worst performing method in four categories: 1.) sequence identity, 2.) functional concordance, 3.) phylogenetic concordance, and 4.) number of orthologs detected.*

Increased ortholog quality leads to more conservation and more positively selected sites

Having demonstrated an increase in ortholog quality using tree-, function-, and similarity-based measures of quality, we next sought to assess the influence of increased alignment quality on estimated levels of selection. To assess gene-level conservation, we applied Phylogenetic Analysis by Maximum Likelihood (PAML) (Yang, 2007) with automated likelihood-based model selection (see methods below). To ascertain site-level positive selection, we employed Sitewise Likelihood Ratio (SLR), a method shown to have a higher power and a lower false positive rate than PAML's popular Bayes Empirical Bayes (BEB) method (Massingham & Goldman, 2005).

Since varying numbers of sequences can sway evolutionary estimates in unpredictable ways due to, e.g. inhomogeneous levels of selection across organisms, we assessed the performance of MOSAIC relative to each method by matching the species present in each alignment. We refer to this approach as $\text{MOSAIC}_{\text{matched}}$. In the case of both PAML and SLR, synonymous substitution rates in coding DNA are used as a background against which to test for changes in rates of non-synonymous substitution. Since the metaPhOrs database provides only protein sequences for its alignments, no comparison with this method was possible given the available data.

In figure 2-5A, we see that MOSAIC leads to higher gene-level conservation (lower dN/dS) compared to every method except Blat, for which the difference was not statistically significant. Despite higher levels of conservation, MOSAIC was able to detect ~30-180% more positively selected sites than any of its component methods. This was not due to an increase in the inferred rate of positive selection. Rather, most of this increase in power was due to the fact that more sites were aligned to high confidence and therefore included in the analysis. This step of filtering for alignment quality is important because site-wise estimates of positive selection are highly sensitive to short poorly aligned regions (Jordan & Goldman, 2012).

To investigate the quality of the positively selected sites detected by MOSAIC, we assessed concordance with and between component methods. For a pair of method, we measure overlap by dividing the total size of the intersection between positively selected sites by the total size of the union. These results are shown in figure 2-5B. We observe that the minimum overlap between MOSAIC and a component method

(MOSAIC/Blat) is still better than the best overlap between component methods (Multiz/OMA). Averaging over comparisons, we find the improvement in concordance with versus between component methods is statistically significant beyond computational precision (p < 1e-16).

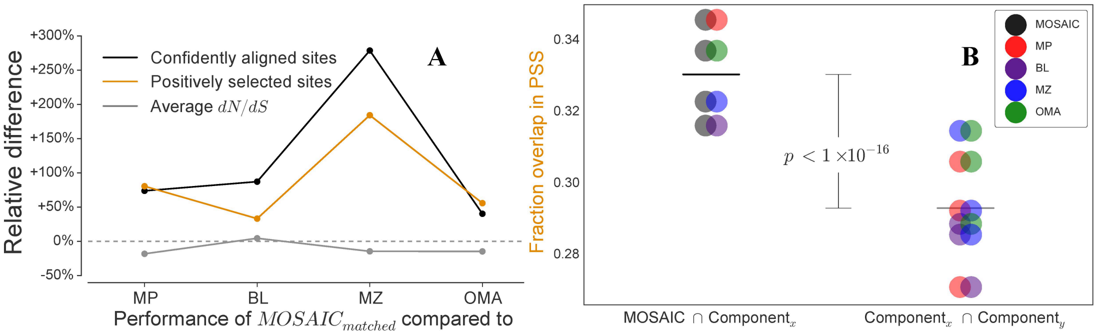


Figure 2-5. A comparison of evolutionary estimates. *A.) The relative difference of MOSAIC$_{matched}$ versus each component method for: 1.) the number of positively selected sites, 2.) the number of confidently aligned sites, and for reference, 3.) the average level of conservation across all alignments. B.) The agreement between positively selected sites 1.) between MOSAIC and component methods, and 2.) among component methods. Fractional overlap values are plotted as Venn diagrams to illustrate the two methods being compared.*

Better alignments may yield new insights into human evolution

We next sought to examine the biological significance of some of the positively selected sites identified uniquely by MOSAIC. This led us to Tryptase Alpha/Beta 1 (TPSAB1), a tetrameric serine protease that has been implicated in the pathogenesis of asthma and other allergic and inflammatory disorders (Pruitt et al., 2014). Shown in 2-ure 6 is the three-dimensional structure of a TPSAB1 tetramer with inhibitor (white) bound (Costanzo et al., 2008). In orange, distal to the active site, is the positively selected residue detected by component methods and by MOSAIC. Note that positive selection at this location is active only outside of the great apes, with a fixed lysine observed in human, chimp, gorilla, and orangutan (fig. SA 9-10).

In red, directly within the proteolytic pore, is the site identified by MOSAIC as positively selected. This residue is a positively charged arginine in humans. This would be expected to modify the electrostatics of ligand binding. In chimp, we instead observe a kink-inducing proline. We might anticipate this change to have a large steric effect, possibly allowing the inward-facing unstructured loop to act as a more rigid lid closing over top of the substrate, or as a modifier of subunit contacts. Importantly, these changes occurred repeatedly in mammals. Proline is observed at this position in rhesus macaque and marmoset. Arginine, on the other hand, is present in gorilla and horse (fig. SA 9-10). In orangutan, we observe a histidine: another positively charged amino acid.

Throughout this examination, we must be cognizant that tryptases evolved rapidly during primate evolution (Trivedi, Tong, Raman, Bhagwandin, & Caughey, 2007). The expansion of this gene family can itself be viewed as an example of positive selection. However, the presence of several paralogs creates the risk of inappropriately aligning pseudo-orthologous sequences that have evolved to serve divergent functions. Given the challenges, this case study provides an excellent opportunity to compare the high-throughput performance of MOSAIC to that of manually curated alignments.

As a first step, we showed that each proposal ortholog was a blast-based best hit to TPSAB1 (Table SA-1). Next, we compared our sequences to those retrieved manually by Trivedi *et al.* in 2007. While we notice a few minor discrepancies between the two sets of alignments (see fig. SA-9 vs. SA-11, reproduced from Trivedi et al. 2007), these differences do not alter our conclusion of human-relevant positive selection at the highlighted site in the proteolytic core of TPSAB1.

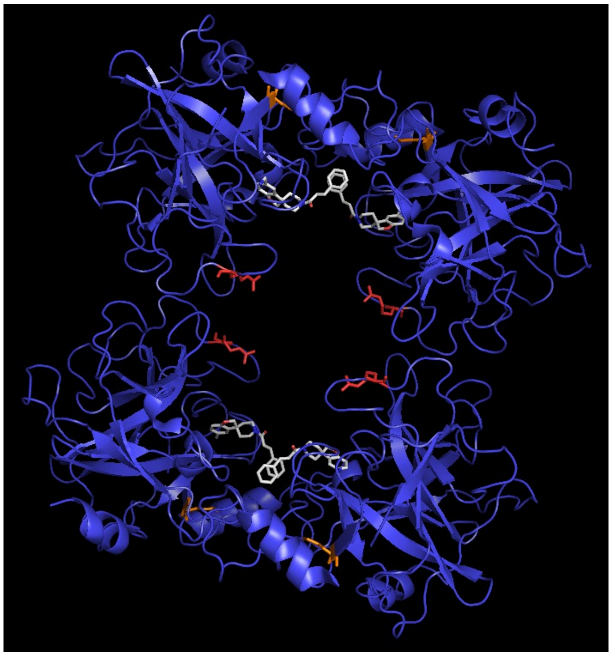


Figure 2-6. **Example: a MOSAIC-specific PSS in Tryptase Alpha/Beta 1.** *The tetrameric TPSAB1 structure is shown with positively selected sites highlighted. The site detected by component methods and by MOSAIC is colored orange, while the MOSAIC-specific PSS is featured in red. A bound inhibitor (white) pinpoints the active site of the enzyme.*

Conclusions

In this paper we have introduced a novel algorithm, MOSAIC, which is capable of integrating an arbitrary number of methodologically diverse ortholog detection methods. We have demonstrated that MOSAIC provides large increases in power relative to its component methods, while simultaneously maintaining or improving functional-, phylogenetic-, and sequence identity-based measures of ortholog quality. Further, given the same number of species, MOSAIC alignments include more columns aligned with high confidence. This translates to higher levels of estimated conservation, and simultaneously, a greatly increased number of positively selected sites detected. Moreover, MOSAIC's positively selected sites agree better with those from component methods than component results do with each other. This suggests that not only does MOSAIC detect more positively selected sites—these sites are more reproducible and are detected due to an increase in alignment quality. Finally, we illustrated the significance of this increase in power by highlighting a positively selected site near the active site of the tryptase TPSAB1. Given the role of this enzyme in asthma and other allergic and inflammatory disorders, we feel that this case study is worthy of experimental follow-up.

In summary, MOSAIC provides the unique flexibility to incorporate any OD method that may be available now or in the future. It can therefore capture the entire swath of methodological diversity, thereby improving OD performance, and allowing researchers to take advantage of methodological gains in a variety of areas of OD research. In addition, it provides the flexibility to adapt scoring and optimization procedures to the set of species under study. In future work, it will be interesting to ascertain how optimal procedures vary between species sets that have differing mean levels of divergence and markedly different patterns of evolution. For example, mammals and prokaryotes will likely have distinct optimal parameter values within MOSAIC. This tool is available a python package that can be installed using easy_install bio-mosaic from the command-line. MOSAIC alignments, source code, and full documentation are available at http://pythonhosted.org/bio-MOSAIC.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In F. Czaki & B. N. Petrov (Eds.), *2nd International Symposium on Information Theory* (pp. 267–281). Budapest: Akademiai Kiado.

Akey, J. M. (2009). Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Research*, *19*(5), 711–22. doi:10.1101/gr.086652.108

Alexeyenko, A., Tamas, I., Liu, G., & Sonnhammer, E. L. L. (2006). Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics*, *22*(14), e9–15. doi:10.1093/bioinformatics/btl213

Altenhoff, A. M., & Dessimoz, C. (2009a). Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Computational Biology*, *5*(1), e1000262. doi:10.1371/journal.pcbi.1000262

Altenhoff, A. M., & Dessimoz, C. (2009b). Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Computational Biology*, *5*(1), e1000262. doi:10.1371/journal.pcbi.1000262

Altenhoff, A. M., & Dessimoz, C. (2012). Inferring Orthology. In M. Anisimova (Ed.), *Evolutionary Genomics* (Vol. 855). Totowa, NJ: Humana Press. doi:10.1007/978-1-61779-582-4

Altenhoff, A. M., Schneider, A., Gonnet, G. H., & Dessimoz, C. (2011). OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Research*, *39*(Database issue), D289–94. doi:10.1093/nar/gkq1238

Arrieta, M.-C., Stiemsma, L. T., Amenyogbe, N., Brown, E. M., & Finlay, B. (2014). The Intestinal Microbiome in Early Life: Health and Disease. *Frontiers in Immunology*, *5*, 427. doi:10.3389/fimmu.2014.00427

Babbitt, C. C., Warner, L. R., Fedrigo, O., Wall, C. E., & Wray, G. A. (2011). Genomic signatures of diet-related shifts during human origins. *Proceedings. Biological Sciences / The Royal Society*, *278*(1708), 961–9. doi:10.1098/rspb.2010.2433

Bakewell, M. A., Shi, P., & Zhang, J. (2007). More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(18), 7489–94. doi:10.1073/pnas.0701705104

Berman, H. M. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. doi:10.1093/nar/28.1.235

Bertsekas, D. (1999). *Nonlinear Programming* (p. 780). Athena Scientific; 2nd edition.

Bezanson, J., Karpinski, S., Shah, V. B., & Edelman, A. (2012). Julia: A Fast Dynamic Language for Technical Computing. Programming Languages; Computational Engineering, Finance, and Science.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F. a, Roskin, K. M., … Miller, W. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research*, *14*(4), 708–15. doi:10.1101/gr.1933104

Bowden, R. J., & Turkington, D. A. (1990). *Instrumental Variables*. Cambridge University Press.

Boyko, A. R., Williamson, S. H., Indap, A. R., Degenhardt, J. D., Hernandez, R. D., Lohmueller, K. E., … Bustamante, C. D. (2008). Assessing the evolutionary impact

of amino acid mutations in the human genome. *PLoS Genetics*, *4*(5), e1000083. doi:10.1371/journal.pgen.1000083

Bradley, B. J. (2008). Reconstructing phylogenies and phenotypes: a molecular view of human evolution. *Journal of Anatomy*, *212*(4), 337–53. doi:10.1111/j.1469-7580.2007.00840.x

Burkart, J. M., Allon, O., Amici, F., Fichtel, C., Finkenwirth, C., Heschl, A., … van Schaik, C. P. (2014). The evolutionary origin of human hyper-cooperation. *Nature Communications*, *5*, 4747. doi:10.1038/ncomms5747

Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M. T., Glanowski, S., … Clark, A. G. (2005a). Natural selection on protein-coding genes in the human genome. *Nature*, *437*(7062), 1153–7.

Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M. T., Glanowski, S., … Clark, A. G. (2005b). Natural selection on protein-coding genes in the human genome. *Nature*, *437*(7062), 1153–7. doi:10.1038/nature04240

Caporaso, J. G., Lauber, C. L., Costello, E. K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., … Knight, R. (2011). Moving pictures of the human microbiome. *Genome Biology*, *12*(5), R50. doi:10.1186/gb-2011-12-5-r50

Capra, J. A., Stolzer, M., Durand, D., & Pollard, K. S. (2013). How old is my gene? *Trends in Genetics : TIG*, *29*(11), 659–68. doi:10.1016/j.tig.2013.07.001

Casdagli, M., Eubank, S., Farmer, J. D., & Gibson, J. (1991). State space reconstruction in the presence of noise. *Physica D: Nonlinear Phenomena*, *51*(1-3), 52–98. doi:10.1016/0167-2789(91)90222-U

Chandrasekaran, V., & Jordan, M. I. (2013). Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(13), E1181–90. doi:10.1073/pnas.1302293110

Charlesworth, J., & Eyre-Walker, A. (2008). The McDonald-Kreitman test and slightly deleterious mutations. *Molecular Biology and Evolution*, *25*(6), 1007–15. doi:10.1093/molbev/msn005

Chen, F., Mackey, A. J., Vermunt, J. K., & Roos, D. S. (2007). Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PloS One*, *2*(4), e383. doi:10.1371/journal.pone.0000383

Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B., & Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science (New York, N.Y.)*, *311*(5765), 1283–7. doi:10.1126/science.1123061

Costanzo, M. J., Yabut, S. C., Zhang, H.-C., White, K. B., de Garavilla, L., Wang, Y., … Maryanoff, B. E. (2008). Potent, nonpeptide inhibitors of human mast cell tryptase. Synthesis and biological evaluation of novel spirocyclic piperidine amide derivatives. *Bioorganic & Medicinal Chemistry Letters*, *18*(6), 2114–21. doi:10.1016/j.bmcl.2008.01.093

Dayhoff, M. O., Schwartz, R. M., & Orcutt, B. C. (1978). A model of evolutionary change in proteins. In M. O. Dayhoff (Ed.), *Atlas of protein sequence and structure* (pp. 345–358). Nature Biomedical Research.

Deyle, E. R., Fogarty, M., Hsieh, C., Kaufman, L., MacCall, A. D., Munch, S. B., … Sugihara, G. (2013). Predicting climate effects on Pacific sardine. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(16), 6430–5. doi:10.1073/pnas.1215506110

Dixon, P. A., Milicich, M. J., & Sugihara, G. (1999). Episodic Fluctuations in Larval Supply. *Science*, *283*(5407), 1528–1530. doi:10.1126/science.283.5407.1528

Ebersberger, I., Strauss, S., & von Haeseler, A. (2009). HaMStR: profile hidden markov model based search for orthologs in ESTs. *BMC Evolutionary Biology*, *9*(1), 157. doi:10.1186/1471-2148-9-157

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, *7*(10), e1002195. doi:10.1371/journal.pcbi.1002195

Eden, E., Navon, R., Steinfeld, I., Lipson, D., & Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, *10*(1), 48. doi:10.1186/1471-2105-10-48

Eilertson, K. E., Booth, J. G., & Bustamante, C. D. (2012). SnIPRE: selection inference using a Poisson random effects model. *PLoS Computational Biology*, *8*(12), e1002806. doi:10.1371/journal.pcbi.1002806

Finch, C. E. (2010). Evolution in health and medicine Sackler colloquium: Evolution of the human lifespan and diseases of aging: roles of infection, inflammation, and nutrition. *Proceedings of the National Academy of Sciences of the United States of America*, *107 Suppl* (suppl_1), 1718–24. doi:10.1073/pnas.0909606106

Fisher, C. K., & Mehta, P. (2014). Identifying Keystone Species in the Human Gut Microbiome from Metagenomic Timeseries using Sparse Linear Regression.

Foster, J. A., & McVey Neufeld, K.-A. (2013). Gut-brain axis: how the microbiome influences anxiety and depression. *Trends in Neurosciences*, *36*(5), 305–12. doi:10.1016/j.tins.2013.01.005

Gajer, P., Brotman, R. M., Bai, G., Sakamoto, J., Schütte, U. M. E., Zhong, X., … Ravel, J. (2012). Temporal dynamics of the human vaginal microbiota. *Science Translational Medicine*, *4*(132), 132ra52. doi:10.1126/scitranslmed.3003605

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods Title. *Econometrica*, *37*(3), 424–438.

Haygood, R., Babbitt, C. C., Fedrigo, O., & Wray, G. A. (2010). Contrasts between adaptive coding and noncoding changes during human evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(17), 7853–7. doi:10.1073/pnas.0911249107

Henikoff, S. (1992). Amino Acid Substitution Matrices from Protein Blocks. *Proceedings of the National Academy of Sciences*, *89*(22), 10915–10919.

Hernandez, R. D. (2008). A flexible forward simulator for populations subject to selection and demography. *Bioinformatics (Oxford, England)*, *24*(23), 2786–7. doi:10.1093/bioinformatics/btn522

Heskamp, L., Meel-van den Abeelen, A., Katsogridakis, E., Panerai, R., Simpson, D., Lagro, J., & Claassen, J. (2013). Convergent cross mapping: a promising technique for future cerebral autoregulation estimation. *CEREBROVASCULAR DISEASES*, *35*, 15–16.

Hubisz, M. J., & Pollard, K. S. (2014). Exploring the genesis and functions of Human Accelerated Regions sheds light on their role in human evolution. *Current Opinion in Genetics & Development*, *29C*, 15–21. doi:10.1016/j.gde.2014.07.005

Hulsen, T., Huynen, M. A., de Vlieg, J., & Groenen, P. M. A. (2006). Benchmarking ortholog identification methods using functional genomics data. *Genome Biology*, *7*(4), R31. doi:10.1186/gb-2006-7-4-r31

Jordan, G., & Goldman, N. (2012). The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Molecular Biology and Evolution*, *29*(4), 1125–39. doi:10.1093/molbev/msr272

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Research*, *12*(4), 656–64. doi:10.1101/gr.229202. Article published online before March 2002

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, a. D. (2002). The Human Genome Browser at UCSC. *Genome Research*, *12*(6), 996–1006. doi:10.1101/gr.229102

Khoruts, A., & Weingarden, A. R. (2014). Emergence of fecal microbiota transplantation as an approach to repair disrupted microbial gut ecology. *Immunology Letters*. doi:10.1016/j.imlet.2014.07.016

Kosiol, C., Vinar, T., da Fonseca, R. R., Hubisz, M. J., Bustamante, C. D., Nielsen, R., & Siepel, A. (2008). Patterns of positive selection in six Mammalian genomes. *PLoS Genetics*, *4*(8), e1000144. doi:10.1371/journal.pgen.1000144

Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Machine Learning*, *51*(2), 181–207.

Kuzniar, A., van Ham, R. C. H. J., Pongor, S., & Leunissen, J. a M. (2008). The quest for orthologs: finding the corresponding gene across genomes. *Trends in Genetics : TIG*, *24*(11), 539–51. doi:10.1016/j.tig.2008.08.009

Liu, Y., Schmidt, B., & Maskell, D. L. (2010). MSAProbs: multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities. *Bioinformatics (Oxford, England)*, *26*(16), 1958–64. doi:10.1093/bioinformatics/btq338

Maddison, W. P., & Knowles, L. L. (2006). Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology*, *55*(1), 21–30. doi:10.1080/10635150500354928

Maher, M. C., & Hernandez, R. D. (2013). A MOSAIC of methods: Improving ortholog detection through integration of algorithmic diversity. Populations and Evolution; Quantitative Methods.

Mashiyama, S. T., Malabanan, M. M., Akiva, E., Bhosle, R., Branch, M. C., Hillerich, B., … Babbitt, P. C. (2014). Large-scale determination of sequence, structure, and function relationships in cytosolic glutathione transferases across the biosphere. *PLoS Biology*, *12*(4), e1001843. doi:10.1371/journal.pbio.1001843

Massingham, T., & Goldman, N. (2005). Detecting amino acid sites under positive selection and purifying selection. *Genetics*, *169*(3), 1753–62. doi:10.1534/genetics.104.032144

McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, *351*(6328), 652–4.

McEntyre, .J, & Ostell, J. (Eds.). (2002). *The NCBI Handbook*. Bethesda, MD: National Center for Biotechnology Information.

Messer, P. W., & Petrov, D. A. (2013). Frequent adaptation and the McDonald-Kreitman test. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(21), 8615–20. doi:10.1073/pnas.1220835110

Nielsen, R., Bustamante, C., Clark, A. G., Glanowski, S., Sackton, T. B., Hubisz, M. J., … Cargill, M. (2005a). A scan for positively selected genes in the genomes of

humans and chimpanzees. *PLoS Biology*, *3*(6), e170. doi:10.1371/journal.pbio.0030170

Nielsen, R., Bustamante, C., Clark, A. G., Glanowski, S., Sackton, T. B., Hubisz, M. J., … Cargill, M. (2005b). A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biology*, *3*(6), e170. doi:10.1371/journal.pbio.0030170

Nielsen, R., Hubisz, M. J., Hellmann, I., Torgerson, D., Andrés, A. M., Albrechtsen, A., … Clark, A. G. (2009). Darwinian and demographic forces affecting human protein coding genes. *Genome Research*, *19*(5), 838–49. doi:10.1101/gr.088336.108

Preuss, T. M. (2011). The human brain: rewired and running hot. *Annals of the New York Academy of Sciences*, *1225 Suppl*, E182–91. doi:10.1111/j.1749-6632.2011.06001.x

Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., … Ostell, J. M. (2014). RefSeq: an update on mammalian reference sequences. *Nucleic Acids Research*, *42*(Database issue), D756–63. doi:10.1093/nar/gkt1114

Pruitt, K. D., Harrow, J., Harte, R. A., Wallin, C., Diekhans, M., Maglott, D. R., … Lipman, D. (2009). The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Research*, *19*(7), 1316–23. doi:10.1101/gr.080531.108

Pryszcz, L. P., Huerta-Cepas, J., & Gabaldón, T. (2011). MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. *Nucleic Acids Research*, *39*(5), e32. doi:10.1093/nar/gkq953

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., … Finn, R. D. (2012). The Pfam protein families database. *Nucleic Acids Research*, *40*(Database issue), D290–301. doi:10.1093/nar/gkr1065

Remm, M., Storm, C. E., & Sonnhammer, E. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology*, *314*(5), 1041–52. doi:10.1006/jmbi.2000.5197

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., … Sabeti, P. C. (2011). Detecting novel associations in large data sets. *Science (New York, N.Y.)*, *334*(6062), 1518–24. doi:10.1126/science.1205438

Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, *53*(1-2), 131–147. doi:10.1016/0025-5564(81)90043-2

Rokach, L. (2009). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*(1-2), 1–39. doi:10.1007/s10462-009-9124-7

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*, 72–101.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)*, *30*(9), 1312–3. doi:10.1093/bioinformatics/btu033

Stamatakis, A., & Alachiotis, N. (2010). Time and memory efficient likelihood-based tree searches on phylogenomic alignments with missing data. *Bioinformatics (Oxford, England)*, *26*(12), i132–9. doi:10.1093/bioinformatics/btq205

Stolzer, M., Lai, H., Xu, M., Sathaye, D., Vernot, B., & Durand, D. (2012). Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics (Oxford, England)*, *28*(18), i409–i415. doi:10.1093/bioinformatics/bts386

Stone, M. (1977). An Asymptotic Equivalence of Choice of Model by Cross-validation and Akaike's Criterion. *Journal of the Royal Statistical Society*. Retrieved March 30, 2014, from http://www.jstor.org/discover/10.2307/2984877?uid=3739560&uid=2134&uid=2&uid=70&uid=4&uid=3739256&sid=21103766217637

Sugihara, G. (1994). Nonlinear Forecasting for the Classification of Natural Time Series. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *348*(1688), 477–495. doi:10.1098/rsta.1994.0106

Sugihara, G., May, R., Ye, H., Hsieh, C., Deyle, E., Fogarty, M., & Munch, S. (2012). Detecting causality in complex ecosystems. *Science (New York, N.Y.)*, *338*(6106), 496–500. doi:10.1126/science.1227079

Sukumaran, J., & Holder, M. T. (2010). DendroPy: a Python library for phylogenetic computing. *Bioinformatics (Oxford, England)*, *26*(12), 1569–71. doi:10.1093/bioinformatics/btq228

Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS One*, *6*(7), e21800. doi:10.1371/journal.pone.0021800

Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, *34*(Web Server issue), W609–12. doi:10.1093/nar/gkl315

Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and Testing Dependence by Correlation of Distances. *The Annals of Statistics*, *35*(6), 2769–2794.

The Chimpanzee Sequencing and Analysis Consortium. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, *437*(7055), 69–87. doi:10.1038/nature04072

Torgerson, D. G., Boyko, A. R., Hernandez, R. D., Indap, A., Hu, X., White, T. J., … Clark, A. G. (2009). Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. *PLoS Genetics*, *5*(8), e1000592. doi:10.1371/journal.pgen.1000592

Trivedi, N. N., Tong, Q., Raman, K., Bhagwandin, V. J., & Caughey, G. H. (2007). Mast cell alpha and beta tryptases changed rapidly during primate speciation and evolved from gamma-like transmembrane peptidases in ancestral vertebrates. *Journal of Immunology (Baltimore, Md. : 1950)*, *179*(9), 6072–9.

Van der Laan, M. J., & Gruber, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *The International Journal of Biostatistics*, *6*(1), Article 17. doi:10.2202/1557-4679.1181

Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, *6*, Article25. doi:10.2202/1544-6115.1309

Vanderweele, T. J., & Arah, O. A. (2011). Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology (Cambridge, Mass.)*, *22*(1), 42–52. doi:10.1097/EDE.0b013e3181f74493

Varki, A. (2012). Nothing in medicine makes sense, except in the light of evolution. *Journal of Molecular Medicine (Berlin, Germany)*, *90*(5), 481–94. doi:10.1007/s00109-012-0900-5

Vujkovic-Cvijin, I., Dunham, R. M., Iwai, S., Maher, M. C., Albright, R. G., Broadhurst, M. J., … McCune, J. M. (2013). Dysbiosis of the gut microbiota is associated with HIV disease progression and tryptophan catabolism. *Science Translational Medicine*, *5*(193), 193ra91. doi:10.1126/scitranslmed.3006438

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164. doi:10.1093/nar/gkq603

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., … Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, *42*(Database issue), D1001–6. doi:10.1093/nar/gkt1229

Williamson, S., Fledel-Alon, A., & Bustamante, C. D. (2004). Population genetics of polymorphism and divergence for diploid selection models with arbitrary dominance. *Genetics*, *168*(1), 463–75. doi:10.1534/genetics.103.024745

Wolpert, D. H., & Macready, W. G. (1997). No Free Lunch Theorems For Optimization. *IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION*, *1*(1), 67–82.

Wu, J., Sinfield, J. L., Buchanan-Wollaston, V., & Feng, J. (2009). Impact of environmental inputs on reverse-engineering approach to network structures. *BMC Systems Biology*, *3*, 113. doi:10.1186/1752-0509-3-113

Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews. Genetics*, *13*(5), 329–42. doi:10.1038/nrg3174

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–91. doi:10.1093/molbev/msm088

Yu, C., Zavaljevski, N., Desai, V., & Reifman, J. (2011). QuartetS: a fast and accurate algorithm for large-scale orthology detection. *Nucleic Acids Research*, *39*(13), e88. doi:10.1093/nar/gkr308

Zhao, S., Sakai, A., Zhang, X., Vetting, M. W., Kumar, R., Hillerich, B., … Jacobson, M. P. (2014). Prediction and characterization of enzymatic activities guided by sequence similarity and genome neighborhood networks. *eLife*, *3*. doi:10.7554/eLife.03275

Chapter 3 The Yin and Yang of Evolution: Insights into the shared genetic basis of human traits and diseases

Introduction

A multitude of cognitive and physiological changes have occurred in the ~5-7 million years since humans last shared a common ancestor with chimps (Bradley, 2008; Preuss, 2011). Cranial volume in humans has nearly quintupled during this time (Babbitt, Warner, Fedrigo, Wall, & Wray, 2011). With increased brainpower has come complex sociality, rich culture, and a whirlwind of technological advances (Burkart et al., 2014). Beyond cognitive changes, chimps and humans have diverged in lifespan, as well as susceptibility to cancer and infectious disease (Finch, 2010). In short, there are a large number of human-chimp differences that are both biologically interesting and biomedically relevant (Varki, 2012). Importantly, evolutionary analysis may provide a powerful opportunity to uncover the genetic drivers of these phenotypic changes.

Human evolution has been studied at time scales ranging from many millions (Kosiol et al., 2008) to hundreds of thousands (Bakewell, Shi, & Zhang, 2007; Bustamante et al., 2005a), to tens of thousands of years (Nielsen *et al.* 2009; Akey 2009; Haygood *et al.* 2010). Depending on the source data and the evolutionary tools employed, researchers may examine selection within specific loci, or across the entire genome. While much important work has been done to assess selection genome-wide, functional evolution outside of well-annotated regions remains difficult to understand, though recent progress has been made (Hubisz & Pollard, 2014; Torgerson et al., 2009).

In this study, we focus specifically on the evolution of human proteins since our divergence from chimp. We find this timescale compelling because of the aforementioned massive changes in, e.g. intelligence, sociality, life span, and cancer susceptibility that occurred during this time. A handful of studies have looked at human coding evolution during this time period. Since we are interested in drivers of phenotypic change, we will focus specifically on diversifying or positive selection. Estimates of the number of human genes under positive selection have ranged from 1 to 20% (Bakewell et al., 2007; Boyko et al., 2008; Bustamante et al., 2005a; Nielsen et al., 2005a; The Chimpanzee Sequencing and Analysis Consortium, 2005). Previous studies have shown that positive selection was either not significantly correlated with biological function after adjusting for multiple comparisons (Bustamante et al., 2005b) or was concentrated within classes of genes, such as immunity or olfaction, where high diversity or a preponderance of paralogs may confound evolutionary inference (Nielsen et al., 2005b).

Since the publication of these studies, we have gained access to increased computational power, better sequencing data, and improved statistical models. In this paper, we revisit the question of human coding evolution since divergence from chimp. By bringing together better data with improved methods, well-validated tuning parameter values, and a more precise statistical focus, we are able to uncover a massive signal of selection in genes related to intelligence, life span, cancer susceptibility, and basic cellular functions such as alternative splicing.

In the process, we infer the specific amino acid changes that occurred on the human lineage since divergence with the human-chimp ancestor. This provides us with a well-defined set of mutations in a small subset of the genome that we believe may have helped shape humans as a species. While much experimental follow-up work remains to be done, we hope this study will provide a useful resource for understanding how protein-coding changes have helped define the very essence of what it means to be human.

Materials and Methods

Statistical modeling

We calculated selection coefficients for each gene using SnIPRE, a mixed-effects model described by (Eilertson, Booth, & Bustamante, 2012). This model was shown to vastly increase power relative to the original McDonald-Kreitman (MK) test (McDonald & Kreitman, 1991). Due to low numbers of mutations, particularly in polymorphism data, the MK test can be highly variable. This variance can be decreased by recognizing that selective pressures are correlated across genes within the same genome. By pooling information across genes, SnIPRE is able to reduce noise and improve statistical power.

Input data

The input data for this model is effectively a two-by-two table of counts for each gene. Mutations are classified as either synonymous or non-synonymous, and as being polymorphic or divergent. Polymorphic mutations are defined as those that vary between human individuals. Divergent mutations, on the other hand, are different between human and chimp, but not polymorphic within the human population.

Human-chimp divergence data

We calculated changes between human and chimp using Ancestral Sequence Reconstruction (ASR), as implementing in Phylogenetic Analysis by Maximum

Likelihood (PAML) (Yang, 2007). For each gene, we fit neutral, conservation, positive selection, and human branch-specific positive selection models. The ASR output was accepted for the model with the lowest Akaike Information Criterion (AIC) (Akaike, 1973). This statistic derives from information theory and provides a measure of whether model fit has improved after accounting for potential over-fitting from adding additional parameters (Stone, 1977).

These models were applied to multiple sequences alignments (MSAs) of human, chimp, orangutan, gorilla, and rhesus macaque. To determine which genes derived from a common ancestral gene, we integrated so-called ortholog predictions from MultiParanoid (Alexeyenko et al., 2006), Blat (W James Kent, 2002), MultiZ (Blanchette et al., 2004), and OMA (Altenhoff et al., 2011). This integration was performed using MOSAIC, a software tool developed by our group which can integrate an arbitrary number of methodologically diverse ortholog detection methods (Maher & Hernandez, 2013). Once orthologs were identified, they were aligned at the amino acid level using MSAprobs (Liu et al., 2010). These amino acid alignments were then converted to codon alignments using PAL2NAL (Suyama, Torrents, & Bork, 2006). This step was undertaken because higher levels of conservation at the protein level allow for higher confidence alignments (Jordan & Goldman, 2012).

Human polymorphism data

Phase 3 genotypes were downloaded from the 1000 genomes project ftp site (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp). We then intersected variants with CCDS release 17 coding sequences (Pruitt et al., 2009) and determined effects on proteins sequences using

ANNOVAR (Wang, Li, & Hakonarson, 2010). We next calculated allele frequencies for each variant for use in downstream filtering.

For the MK test, it is therefore necessary to filter out low frequency variants. The MK test uses polymorphism data as an estimator of neutral evolution. This assumption of neutrality often does not hold for low frequency variants, however. While many low frequency variants are uncommon because they have emerged recently, others exist at low frequency because they are harmful and selected against evolutionarily.

To address this issue, we used a sample-wide frequency cutoff of fifteen percent, which is value commonly employed for the MK test (Charlesworth & Eyre-Walker, 2008). We also tested cutoffs of 5, 10, 20, and 25 percent. Using a cutoff of 5% led to poor agreement with other cutoff levels, however, pairwise correlations between selection estimates for all higher frequency cutoffs were in excess of 0.95 (Figure SB-4).

Comparison to previous result

We downloaded raw data from (Bustamante et al., 2005b). These tables were then run through SnIPRE to produce equivalently estimated selection values.

Simulation study

To test SnIPRE's performance, we performed forward simulations of selection on coding sequences with SFS_CODE (Hernandez, 2008). For each gene in the human genome, we computed the total length of its exons using the CCDS (Pruitt et al., 2009). We then performed a single simulation for each gene. Hence, the distribution of gene lengths in our simulations exactly matches the genome-wide distribution in humans.

For each simulated gene, we drew a single value of the selection coefficient, 2Ns, from a normal distribution with mean 0 and standard deviation 100/6. In our simulations, all non-synonymous mutations within each gene have the same selection coefficient. A sample command line for these simulations is provided below. For each simulation, after 10,000 generations we sampled 50 diploid individuals from a population of 500 and calculated the number of non-synonymous and synonymous substitutions that occurred in the previous 5,000 generations. We tallied the number of currently segregating non-synonymous and synonymous polymorphisms.

Sample command line:

```
sfs_code 1 1 -t 0.001 -r 0.001 -N 500 -n 50 -TE 10 -L 1 <L> -W L 0 1 <gamma> 0 1 -s <random>
```

These values were used as input to SnIPRE. Note that we have performed this analysis without simulating an out-group, meaning that we assume perfect knowledge of the substitutions and have not inferred them from the data. Thus, our analyses with simulated data may contain less noise than real data.

Briefly, -t sets theta = 4Nu (mutation rate), -r sets rho=4Nr (recombination rate), -N the population size, -n the sample size, -TE the length of the simulation in units of 2N generations, -L the length of the simulated gene in base pairs, -W the value of the selection coefficient (2Ns), and –s is a random seed.

GO analysis

We examined enriched Gene Ontology (GO) functions using GOrilla (Eden, Navon, Steinfeld, Lipson, & Yakhini, 2009). We performed a rank-based test for enrichment that does not use a cutoff for gamma. Results were robust to this decision, however. Cutoff-based tests yielded qualitatively similar results that were not sensitive to the minimum level of positive selection necessary to consider a gene rapidly evolving.

GWAS enrichment

We next investigated whether positively selected genes were associated with diseases and phenotypes present in the Genome Wide Associate Study (GWAS) Catalog (Welter et al., 2014). Rather than re-implementing the rank-based enrichment test used by GOrilla, we defined positively selected genes as those with a gamma greater than 0.05. Enrichment was then assessed using a fisher's exact test. To control the number of statistical comparisons, we calculated the minimum category size necessary to detect an odds ratio of five, 75% of the time. Only categories with at least this number of genes were analyzed. As an exploratory analysis, we then performed the enrichment test on all categories.

Results

In simulations, SnIPRE has excellent ability to detect positive selection in humans

Previous studies have suffered from a severe lack of power at the timescale we are interested in. Fortunately, some of this data can be reanalyzed with more advanced tools such as SnIPRE, a linear-mixed effects model implementation of the McDonald-Kreitman test. To assess the ability of SnIPRE to detect positive selection at the timescales that we are interested in, we simulated evolution on human coding regions over 10,000 generations. We then assessed the correlation between true values of gamma, the population-scaled selection coefficient, and the gamma estimates from SnIPRE.

For positively selected genes, that is, genes with a true gamma values greater than 0, we observed a pearson correlation between estimated and observed values of 0.8 (see Figure 3-1). This result holds for simulations including complex demography and heterogeneous selection within genes (see Figure SB-2). As expected, this correlation sharply degrades for conserved genes. This is because conservation tends to remove non-synonymous mutations, leading to zero counts in the underlying data. This then impedes differentiating between varying strengths of conservation.

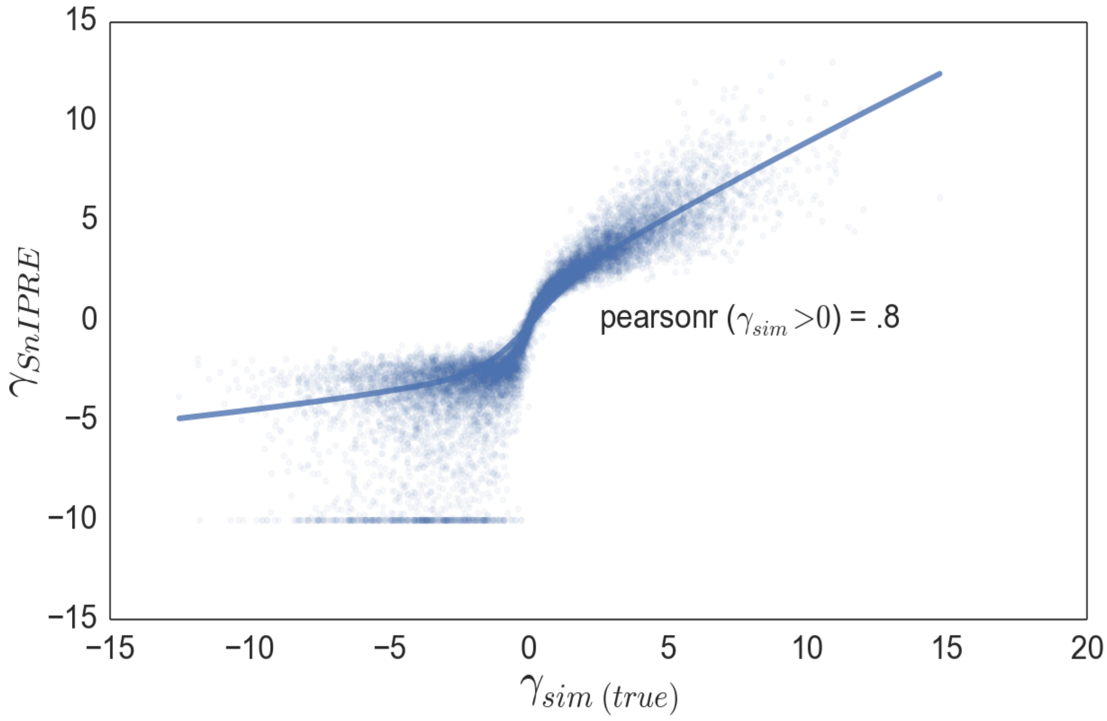


Figure 3-1. A comparison of simulated selections coefficients versus those inferred by SnIPRE.

Fitting SnIPRE on previous data yields little signal of positive selection

Despite the power of SnIPRE to detect positive selection, we found that applying this model to previously published data yielded only 93 positively selected genes (Figure 3-2). This is in sharp contrast to the MKPRF model fit in the original study, which inferred nearly a third of the genome to be under positive selection (Figure SB-1).

Updating input data leads to the detection of many more positively selected genes

Given the lack of signal under the otherwise extremely powerful SnIPRE model, we sought to improve the input data in hopes of uncovering novel evolutionary signatures. We did this in three ways. First, we leveraged high-quality multiple sequence alignments produced by MOSAIC, a software tool developed by our group. Second, we better filtered human polymorphism data based on population genetic theory as well as a thorough examination of the data-at-hand. Third, we leveraged ancestral sequence reconstruction (ASR) to remove mutations that occurred on the chimp lineage.

In doing so, we filtered out chimp-lineage mutations from incoming cross-species divergence data, and we improved human polymorphism data by removing deleterious mutations that would otherwise contribute to an overly conservative null model (see materials & methods for a more thorough explanation). We plot the result of these changes in Figure 3-2. For comparison, we also show the gamma values calculated similarly from previous published data. In all, the methodological improves describe above lead to an increase from 93 to 1773 genes inferred to be under positive selection.

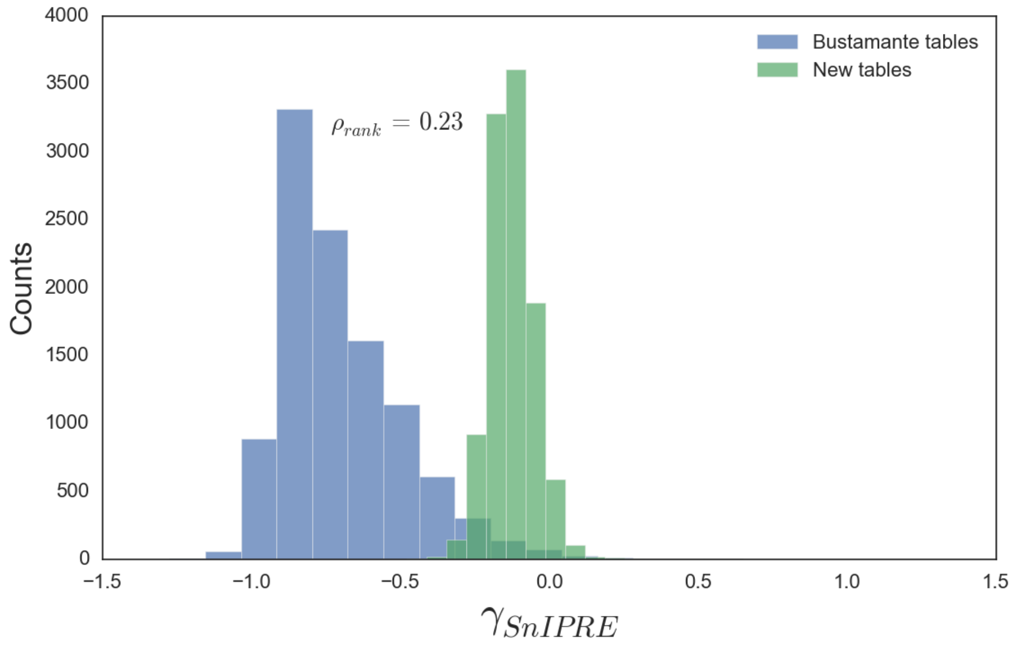


Figure 3-2. A comparison of SnIPRE-inferred selection coefficient distributions. *The blue distribution are the gamma values inferred using tables published by Bustamante et al. 2005. The green distribution is calculated using data generated by the present study.*

Incorporation of ancestral sequence reconstruction greatly increases GO enrichment

Examining species divergence without ancestral sequence reconstruction (ASR) pools both human- and chimp-lineage mutations. Under this approach, we find only GO terms related to olfaction (q-value = 7e-5) and gonadal development (q-value = 1.5e-3) to be enriched among positively selected genes. It is important to note, however, that olfaction-related genes are enriched among those that are most *conserved* on the human lineage (q-value=3.1e-54). This suggests that previous results of positive selection on olfaction in humans (Nielsen et al., 2005b) are in fact driven by adaptation on the chimp rather than human lineage. Indeed, this is much more consistent with expectation given the comparatively poor sense of smell in humans relative to chimp.

Through this analysis, we demonstrate that ASR is useful in eliminating spurious and even misleading evolutionary signal. Moreover, it gives us the newfound ability to detect enriched positive selection in a variety of pathways related to brain development, alternative splicing, tissue organization, fat storage, and locomotion. These are many of the pathways where we would expect to find the genetic basis for the phenotypic divergence between human and chimp. A larger, representative sampling of the pathways with more than two-fold enrichment are shown in Table 1.

	FDR q-value	Fold Enrichment
Positive regulation of neuron projection development	1.2E-04	5.4
Regulation of extent of cell growth	3.6E-04	3.9
mRNA splicing, via spliceosome	1.4E-07	3.4
Memory	9.7E-05	3.4
Regulation of phospholipase activity	5.8E-04	2.9
Neurotrophin signaling pathway	1.7E-05	2.6
Cell junction organization	8.5E-05	2.6
Axon guidance	2.2E-10	2.2
Regulation of locomotion	2.3E-06	2.1

Table 3-1. GO processes enriched among positively selected genes.

Extending our focus to pathways enriched less that two-fold, we find that a number of other processes are also represented, such as transport, behavior, signaling and regulation of signaling, and response to stimulation and its regulation. Finally, we found that positive selection rises markedly as genes belong to increasingly more enriched categories (Figure SB-8). This dose-dependence is consistent with increased diversifying pressure as a gene is involved in a larger number of positively selected traits.

We also find enrichment for association with disease

To further examine whether signals of positive selection were associated with known biological results, we tested whether positively selected genes were over-represented among genes associated by GWAS to particular diseases and traits. Given the small-ish number of proteins known to be involved with most traits, tests for enrichment are expected to be underpowered to make statements about particular diseases and traits. We can, however, easily test whether positively selected genes are, on the whole, more associated with diseases and traits than we would expect by chance.

In Figure 3-3A, we see the distribution of enrichment p-values for enrichment of positive selection for each disease or phenotype. We determined that just above 8% of tests were below the nominal significance threshold. Note that we do not know *a priori* how many tests should exceed this threshold due to the correlation between categories induced by overlapping genes.

We examined this with a permutation test. For each of 1000 replicates, we randomly shuffled whether genes where positively selected or not. We then calculated enrichment across all disease/phenotype categories and summarized the proportion of enrichment p-values that were below the nominal significance threshold of 0.05. For this summary statistic, we found that the expected value under the null was 1.6%, more than four-fold below the observed value of 8%. Further, Figure 3-3B shows that our observed proportion is nearly double the most extreme value observed across 1000 null replicates. We can therefore conclude the positively selected genes are highly enriched for associations with diseases and traits (p << .001).

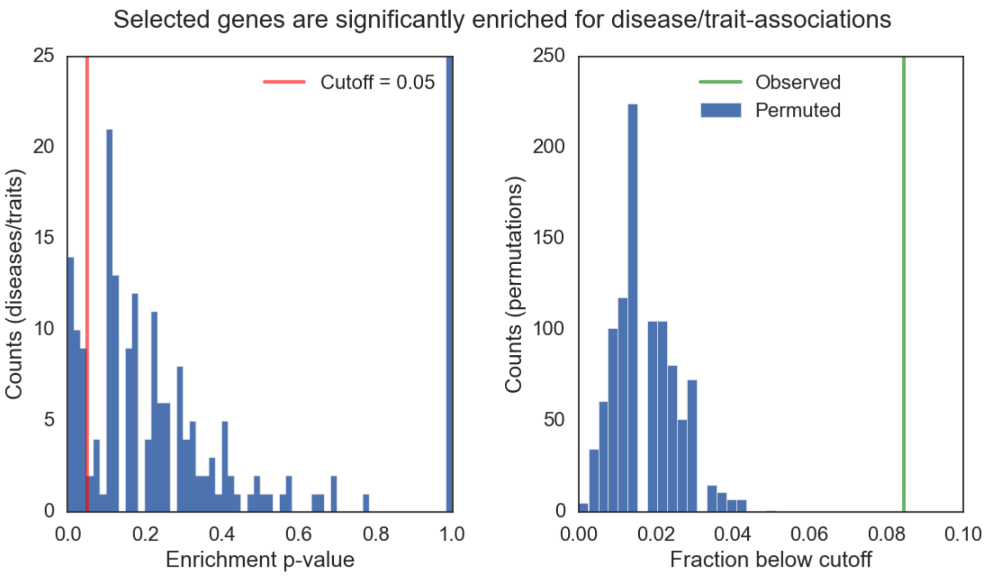


Figure 3-3. Enrichment of disease-trait associations among positively selected genes. *A.) The observed distribution for p-values testing whether individual diseases or traits are enriched among positively selected genes. The red line shows the nominal significance cutoff of 0.05. The fraction of tests below this cutoff is 0.082. B.) To assess the null distribution for fraction of nominally significant tests, we conducted 1000 permutations. This generated the blue distribution, which has a mean of .016." For comparison, the observed value is plotted in green.*

We next sought to examine enrichment for particular phenotypes. To control our number of comparisons, we restricted our analysis to phenotypes with a large enough number of genes to provide 75% power at five-fold enrichment. This yielded significant associations with IgG glycosylation (FDR q-value=5e-5), multiple sclerosis (FDR q-value=.041), and cognitive performance (FDR q-value=.041). The relative enrichments for these categories were 4.5x, 4x, and 4.7x, respectively.

As an exploratory analysis, we then widened our test to the entire list of phenotypes, adjusting for multiple comparisons accordingly. In doing so, we found significant enrichment for association to interstitial lung disease (q-value=.045, enrichment=16x) and eating disorders(q-value=.040, enrichment=53x). These results are summarized in Table 2. As with the GO analysis, levels of enrichment were also dosedependent for disease/phenotype associations (Figure SB-8). We do not observed this trend in permuted data (Figure SB-9).

	Enrich-ment	Q-val (all)	Q-val (restricted)	Tot. Genes
IgG glycosylation	4.5	1.38E-03	7.5E-05	131
Bulimia (purging)	52.5	0.040	N/A	5
Interstitial lung disease	15.6	0.045	N/A	13
Multiple sclerosis	4.0	0.33	0.041	59
Cognitive performance	4.7	0.33	0.041	42

Table 3-2. **Diseases or traits for which Individually significant enrichment was discovered.**

Discussion

In this study we have provided new insight into the genetic basis of several important phenotypic changes that have occurred since the divergence of human and chimp. This advance was facilitated by several improvements on previous efforts. First, we applied SnIPRE, a recently developed, state-of-the-art statistical model for estimating MK statistics (Eilertson et al., 2012). Second, we leverage high-quality multiple sequence alignments produced by MOSAIC, a software tool developed by our group which can integrate an arbitrary number of methodologically diverse ortholog detection methods (Maher & Hernandez, 2013). Third, we better filtered human polymorphism data based on population genetic theory as well as a thorough examination of the data-at-hand. The effect of this improvement is to remove an overly conservative bias from the estimates of rates of neutral evolution. Finally, we leveraged ancestral sequence reconstruction (ASR) to whittle down human-chimp divergence data to only those mutations that took place on the human lineage. Since almost all of the adaptation in, e.g. cranial volume since divergence from chimp is known to have occurred on the human lineage, we would expect this modification to remove the signal-diluting effect of mutations from the chimp lineage.

We find that positive selection has disproportionately affected genes related to brain development, alternative splicing, tissue organization, fat storage, and locomotion. Compared to previous results, these enriched categories cover a much broader swath of the known phenotypic changes since human's divergence from chimp ~5-7 MYA.

Furthermore, we uncover a 4.5 fold enrichment for associations to diseases and traits represented in the NHGRI GWAS catalog. Top associations are found to immune-related function (IgG glycosylation), lung performance (interstitial lung disease), caloric regulation (bulimia), and brain function (cognitive performance, multiple sclerosis).

We observe striking concordance between enrichment in biological function and enrichment for disease association. In particular, we see signal for categories related to brain function, mobility, and diet. Indeed, many of these phenotypic changes could have emerged together. For example, man is thought to have utilized newfound intellectual resources and more efficient locomotion to undertake daylong persistence hunts of large game. Persistence hunters such as the modern-day Kalahari bushmen secured brain-fueling meat by walking, running, and tracking prey to the point of fatal exhaustion. This practice is one of the earliest forms of human hunting.

Taken together, these results support the hypothesis that some of the genotypic changes that had positive effects on human survival may have also left us vulnerable to new or different types of disease. In this paper, we identify the genes where this signal of adaptation is strongest. Further, we are able to focus on genes related to biological processes, human diseases, and well-studied traits. In the process, we are able to identify a set of possibly evolutionarily important genes with known links to phenotypes such as intelligence, obesity, and mental illness.

Conclusions

We have provided here what we believe to be the clearest view yet into the important evolutionary changes that have occurred in coding regions since our separation from our common ancestor with chimps ~5-7 million years ago.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In F. Czaki & B. N. Petrov (Eds.), *2nd International Symposium on Information Theory* (pp. 267–281). Budapest: Akademiai Kiado.

Akey, J. M. (2009). Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Research*, *19*(5), 711–22. doi:10.1101/gr.086652.108

Alexeyenko, A., Tamas, I., Liu, G., & Sonnhammer, E. L. L. (2006). Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics*, *22*(14), e9–15. doi:10.1093/bioinformatics/btl213

Altenhoff, A. M., & Dessimoz, C. (2009a). Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Computational Biology*, *5*(1), e1000262. doi:10.1371/journal.pcbi.1000262

Altenhoff, A. M., & Dessimoz, C. (2009b). Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Computational Biology*, *5*(1), e1000262. doi:10.1371/journal.pcbi.1000262

Altenhoff, A. M., & Dessimoz, C. (2012). Inferring Orthology. In M. Anisimova (Ed.), *Evolutionary Genomics* (Vol. 855). Totowa, NJ: Humana Press. doi:10.1007/978-1-61779-582-4

Altenhoff, A. M., Schneider, A., Gonnet, G. H., & Dessimoz, C. (2011). OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Research*, *39*(Database issue), D289–94. doi:10.1093/nar/gkq1238

Arrieta, M.-C., Stiemsma, L. T., Amenyogbe, N., Brown, E. M., & Finlay, B. (2014). The Intestinal Microbiome in Early Life: Health and Disease. *Frontiers in Immunology*, *5*, 427. doi:10.3389/fimmu.2014.00427

Babbitt, C. C., Warner, L. R., Fedrigo, O., Wall, C. E., & Wray, G. A. (2011). Genomic signatures of diet-related shifts during human origins. *Proceedings. Biological Sciences / The Royal Society*, *278*(1708), 961–9. doi:10.1098/rspb.2010.2433

Bakewell, M. A., Shi, P., & Zhang, J. (2007). More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(18), 7489–94. doi:10.1073/pnas.0701705104

Berman, H. M. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. doi:10.1093/nar/28.1.235

Bertsekas, D. (1999). *Nonlinear Programming* (p. 780). Athena Scientific; 2nd edition.

Bezanson, J., Karpinski, S., Shah, V. B., & Edelman, A. (2012). Julia: A Fast Dynamic Language for Technical Computing. Programming Languages; Computational Engineering, Finance, and Science.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F. a, Roskin, K. M., … Miller, W. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research*, *14*(4), 708–15. doi:10.1101/gr.1933104

Bowden, R. J., & Turkington, D. A. (1990). *Instrumental Variables*. Cambridge University Press.

Boyko, A. R., Williamson, S. H., Indap, A. R., Degenhardt, J. D., Hernandez, R. D., Lohmueller, K. E., … Bustamante, C. D. (2008). Assessing the evolutionary impact

of amino acid mutations in the human genome. *PLoS Genetics*, *4*(5), e1000083. doi:10.1371/journal.pgen.1000083

Bradley, B. J. (2008). Reconstructing phylogenies and phenotypes: a molecular view of human evolution. *Journal of Anatomy*, *212*(4), 337–53. doi:10.1111/j.1469-7580.2007.00840.x

Burkart, J. M., Allon, O., Amici, F., Fichtel, C., Finkenwirth, C., Heschl, A., … van Schaik, C. P. (2014). The evolutionary origin of human hyper-cooperation. *Nature Communications*, *5*, 4747. doi:10.1038/ncomms5747

Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M. T., Glanowski, S., … Clark, A. G. (2005a). Natural selection on protein-coding genes in the human genome. *Nature*, *437*(7062), 1153–7.

Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M. T., Glanowski, S., … Clark, A. G. (2005b). Natural selection on protein-coding genes in the human genome. *Nature*, *437*(7062), 1153–7. doi:10.1038/nature04240

Caporaso, J. G., Lauber, C. L., Costello, E. K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., … Knight, R. (2011). Moving pictures of the human microbiome. *Genome Biology*, *12*(5), R50. doi:10.1186/gb-2011-12-5-r50

Capra, J. A., Stolzer, M., Durand, D., & Pollard, K. S. (2013). How old is my gene? *Trends in Genetics : TIG*, *29*(11), 659–68. doi:10.1016/j.tig.2013.07.001

Casdagli, M., Eubank, S., Farmer, J. D., & Gibson, J. (1991). State space reconstruction in the presence of noise. *Physica D: Nonlinear Phenomena*, *51*(1-3), 52–98. doi:10.1016/0167-2789(91)90222-U

Chandrasekaran, V., & Jordan, M. I. (2013). Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(13), E1181–90. doi:10.1073/pnas.1302293110

Charlesworth, J., & Eyre-Walker, A. (2008). The McDonald-Kreitman test and slightly deleterious mutations. *Molecular Biology and Evolution*, *25*(6), 1007–15. doi:10.1093/molbev/msn005

Chen, F., Mackey, A. J., Vermunt, J. K., & Roos, D. S. (2007). Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PloS One*, *2*(4), e383. doi:10.1371/journal.pone.0000383

Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B., & Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science (New York, N.Y.)*, *311*(5765), 1283–7. doi:10.1126/science.1123061

Costanzo, M. J., Yabut, S. C., Zhang, H.-C., White, K. B., de Garavilla, L., Wang, Y., … Maryanoff, B. E. (2008). Potent, nonpeptide inhibitors of human mast cell tryptase. Synthesis and biological evaluation of novel spirocyclic piperidine amide derivatives. *Bioorganic & Medicinal Chemistry Letters*, *18*(6), 2114–21. doi:10.1016/j.bmcl.2008.01.093

Dayhoff, M. O., Schwartz, R. M., & Orcutt, B. C. (1978). A model of evolutionary change in proteins. In M. O. Dayhoff (Ed.), *Atlas of protein sequence and structure* (pp. 345–358). Nature Biomedical Research.

Deyle, E. R., Fogarty, M., Hsieh, C., Kaufman, L., MacCall, A. D., Munch, S. B., … Sugihara, G. (2013). Predicting climate effects on Pacific sardine. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(16), 6430–5. doi:10.1073/pnas.1215506110

Dixon, P. A., Milicich, M. J., & Sugihara, G. (1999). Episodic Fluctuations in Larval Supply. *Science*, *283*(5407), 1528–1530. doi:10.1126/science.283.5407.1528

Ebersberger, I., Strauss, S., & von Haeseler, A. (2009). HaMStR: profile hidden markov model based search for orthologs in ESTs. *BMC Evolutionary Biology*, *9*(1), 157. doi:10.1186/1471-2148-9-157

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, *7*(10), e1002195. doi:10.1371/journal.pcbi.1002195

Eden, E., Navon, R., Steinfeld, I., Lipson, D., & Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, *10*(1), 48. doi:10.1186/1471-2105-10-48

Eilertson, K. E., Booth, J. G., & Bustamante, C. D. (2012). SnIPRE: selection inference using a Poisson random effects model. *PLoS Computational Biology*, *8*(12), e1002806. doi:10.1371/journal.pcbi.1002806

Finch, C. E. (2010). Evolution in health and medicine Sackler colloquium: Evolution of the human lifespan and diseases of aging: roles of infection, inflammation, and nutrition. *Proceedings of the National Academy of Sciences of the United States of America*, *107 Suppl* (suppl_1), 1718–24. doi:10.1073/pnas.0909606106

Fisher, C. K., & Mehta, P. (2014). Identifying Keystone Species in the Human Gut Microbiome from Metagenomic Timeseries using Sparse Linear Regression.

Foster, J. A., & McVey Neufeld, K.-A. (2013). Gut-brain axis: how the microbiome influences anxiety and depression. *Trends in Neurosciences*, *36*(5), 305–12. doi:10.1016/j.tins.2013.01.005

Gajer, P., Brotman, R. M., Bai, G., Sakamoto, J., Schütte, U. M. E., Zhong, X., … Ravel, J. (2012). Temporal dynamics of the human vaginal microbiota. *Science Translational Medicine*, *4*(132), 132ra52. doi:10.1126/scitranslmed.3003605

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods Title. *Econometrica*, *37*(3), 424–438.

Haygood, R., Babbitt, C. C., Fedrigo, O., & Wray, G. A. (2010). Contrasts between adaptive coding and noncoding changes during human evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(17), 7853–7. doi:10.1073/pnas.0911249107

Henikoff, S. (1992). Amino Acid Substitution Matrices from Protein Blocks. *Proceedings of the National Academy of Sciences*, *89*(22), 10915–10919.

Hernandez, R. D. (2008). A flexible forward simulator for populations subject to selection and demography. *Bioinformatics (Oxford, England)*, *24*(23), 2786–7. doi:10.1093/bioinformatics/btn522

Heskamp, L., Meel-van den Abeelen, A., Katsogridakis, E., Panerai, R., Simpson, D., Lagro, J., & Claassen, J. (2013). Convergent cross mapping: a promising technique for future cerebral autoregulation estimation. *CEREBROVASCULAR DISEASES*, *35*, 15–16.

Hubisz, M. J., & Pollard, K. S. (2014). Exploring the genesis and functions of Human Accelerated Regions sheds light on their role in human evolution. *Current Opinion in Genetics & Development*, *29C*, 15–21. doi:10.1016/j.gde.2014.07.005

Hulsen, T., Huynen, M. A., de Vlieg, J., & Groenen, P. M. A. (2006). Benchmarking ortholog identification methods using functional genomics data. *Genome Biology*, *7*(4), R31. doi:10.1186/gb-2006-7-4-r31

Jordan, G., & Goldman, N. (2012). The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Molecular Biology and Evolution*, *29*(4), 1125–39. doi:10.1093/molbev/msr272

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Research*, *12*(4), 656–64. doi:10.1101/gr.229202. Article published online before March 2002

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, a. D. (2002). The Human Genome Browser at UCSC. *Genome Research*, *12*(6), 996–1006. doi:10.1101/gr.229102

Khoruts, A., & Weingarden, A. R. (2014). Emergence of fecal microbiota transplantation as an approach to repair disrupted microbial gut ecology. *Immunology Letters*. doi:10.1016/j.imlet.2014.07.016

Kosiol, C., Vinar, T., da Fonseca, R. R., Hubisz, M. J., Bustamante, C. D., Nielsen, R., & Siepel, A. (2008). Patterns of positive selection in six Mammalian genomes. *PLoS Genetics*, *4*(8), e1000144. doi:10.1371/journal.pgen.1000144

Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Machine Learning*, *51*(2), 181–207.

Kuzniar, A., van Ham, R. C. H. J., Pongor, S., & Leunissen, J. a M. (2008). The quest for orthologs: finding the corresponding gene across genomes. *Trends in Genetics : TIG*, *24*(11), 539–51. doi:10.1016/j.tig.2008.08.009

Liu, Y., Schmidt, B., & Maskell, D. L. (2010). MSAProbs: multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities. *Bioinformatics (Oxford, England)*, *26*(16), 1958–64. doi:10.1093/bioinformatics/btq338

Maddison, W. P., & Knowles, L. L. (2006). Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology*, *55*(1), 21–30. doi:10.1080/10635150500354928

Maher, M. C., & Hernandez, R. D. (2013). A MOSAIC of methods: Improving ortholog detection through integration of algorithmic diversity. Populations and Evolution; Quantitative Methods.

Mashiyama, S. T., Malabanan, M. M., Akiva, E., Bhosle, R., Branch, M. C., Hillerich, B., … Babbitt, P. C. (2014). Large-scale determination of sequence, structure, and function relationships in cytosolic glutathione transferases across the biosphere. *PLoS Biology*, *12*(4), e1001843. doi:10.1371/journal.pbio.1001843

Massingham, T., & Goldman, N. (2005). Detecting amino acid sites under positive selection and purifying selection. *Genetics*, *169*(3), 1753–62. doi:10.1534/genetics.104.032144

McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, *351*(6328), 652–4.

McEntyre, .J, & Ostell, J. (Eds.). (2002). *The NCBI Handbook*. Bethesda, MD: National Center for Biotechnology Information.

Messer, P. W., & Petrov, D. A. (2013). Frequent adaptation and the McDonald-Kreitman test. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(21), 8615–20. doi:10.1073/pnas.1220835110

Nielsen, R., Bustamante, C., Clark, A. G., Glanowski, S., Sackton, T. B., Hubisz, M. J., … Cargill, M. (2005a). A scan for positively selected genes in the genomes of

humans and chimpanzees. *PLoS Biology*, *3*(6), e170. doi:10.1371/journal.pbio.0030170

Nielsen, R., Bustamante, C., Clark, A. G., Glanowski, S., Sackton, T. B., Hubisz, M. J., … Cargill, M. (2005b). A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biology*, *3*(6), e170. doi:10.1371/journal.pbio.0030170

Nielsen, R., Hubisz, M. J., Hellmann, I., Torgerson, D., Andrés, A. M., Albrechtsen, A., … Clark, A. G. (2009). Darwinian and demographic forces affecting human protein coding genes. *Genome Research*, *19*(5), 838–49. doi:10.1101/gr.088336.108

Preuss, T. M. (2011). The human brain: rewired and running hot. *Annals of the New York Academy of Sciences*, *1225 Suppl*, E182–91. doi:10.1111/j.1749-6632.2011.06001.x

Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., … Ostell, J. M. (2014). RefSeq: an update on mammalian reference sequences. *Nucleic Acids Research*, *42*(Database issue), D756–63. doi:10.1093/nar/gkt1114

Pruitt, K. D., Harrow, J., Harte, R. A., Wallin, C., Diekhans, M., Maglott, D. R., … Lipman, D. (2009). The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Research*, *19*(7), 1316–23. doi:10.1101/gr.080531.108

Pryszcz, L. P., Huerta-Cepas, J., & Gabaldón, T. (2011). MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. *Nucleic Acids Research*, *39*(5), e32. doi:10.1093/nar/gkq953

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., … Finn, R. D. (2012). The Pfam protein families database. *Nucleic Acids Research*, *40*(Database issue), D290–301. doi:10.1093/nar/gkr1065

Remm, M., Storm, C. E., & Sonnhammer, E. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology*, *314*(5), 1041–52. doi:10.1006/jmbi.2000.5197

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., … Sabeti, P. C. (2011). Detecting novel associations in large data sets. *Science (New York, N.Y.)*, *334*(6062), 1518–24. doi:10.1126/science.1205438

Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, *53*(1-2), 131–147. doi:10.1016/0025-5564(81)90043-2

Rokach, L. (2009). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*(1-2), 1–39. doi:10.1007/s10462-009-9124-7

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*, 72–101.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)*, *30*(9), 1312–3. doi:10.1093/bioinformatics/btu033

Stamatakis, A., & Alachiotis, N. (2010). Time and memory efficient likelihood-based tree searches on phylogenomic alignments with missing data. *Bioinformatics (Oxford, England)*, *26*(12), i132–9. doi:10.1093/bioinformatics/btq205

Stolzer, M., Lai, H., Xu, M., Sathaye, D., Vernot, B., & Durand, D. (2012). Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics (Oxford, England)*, *28*(18), i409–i415. doi:10.1093/bioinformatics/bts386

Stone, M. (1977). An Asymptotic Equivalence of Choice of Model by Cross-validation and Akaike's Criterion. *Journal of the Royal Statistical Society*. Retrieved March 30, 2014, from http://www.jstor.org/discover/10.2307/2984877?uid=3739560&uid=2134&uid=2&uid=70&uid=4&uid=3739256&sid=21103766217637

Sugihara, G. (1994). Nonlinear Forecasting for the Classification of Natural Time Series. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *348*(1688), 477–495. doi:10.1098/rsta.1994.0106

Sugihara, G., May, R., Ye, H., Hsieh, C., Deyle, E., Fogarty, M., & Munch, S. (2012). Detecting causality in complex ecosystems. *Science (New York, N.Y.)*, *338*(6106), 496–500. doi:10.1126/science.1227079

Sukumaran, J., & Holder, M. T. (2010). DendroPy: a Python library for phylogenetic computing. *Bioinformatics (Oxford, England)*, *26*(12), 1569–71. doi:10.1093/bioinformatics/btq228

Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS One*, *6*(7), e21800. doi:10.1371/journal.pone.0021800

Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, *34*(Web Server issue), W609–12. doi:10.1093/nar/gkl315

Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and Testing Dependence by Correlation of Distances. *The Annals of Statistics*, *35*(6), 2769–2794.

The Chimpanzee Sequencing and Analysis Consortium. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, *437*(7055), 69–87. doi:10.1038/nature04072

Torgerson, D. G., Boyko, A. R., Hernandez, R. D., Indap, A., Hu, X., White, T. J., … Clark, A. G. (2009). Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. *PLoS Genetics*, *5*(8), e1000592. doi:10.1371/journal.pgen.1000592

Trivedi, N. N., Tong, Q., Raman, K., Bhagwandin, V. J., & Caughey, G. H. (2007). Mast cell alpha and beta tryptases changed rapidly during primate speciation and evolved from gamma-like transmembrane peptidases in ancestral vertebrates. *Journal of Immunology (Baltimore, Md. : 1950)*, *179*(9), 6072–9.

Van der Laan, M. J., & Gruber, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *The International Journal of Biostatistics*, *6*(1), Article 17. doi:10.2202/1557-4679.1181

Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, *6*, Article25. doi:10.2202/1544-6115.1309

Vanderweele, T. J., & Arah, O. A. (2011). Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology (Cambridge, Mass.)*, *22*(1), 42–52. doi:10.1097/EDE.0b013e3181f74493

Varki, A. (2012). Nothing in medicine makes sense, except in the light of evolution. *Journal of Molecular Medicine (Berlin, Germany)*, *90*(5), 481–94. doi:10.1007/s00109-012-0900-5

Vujkovic-Cvijin, I., Dunham, R. M., Iwai, S., Maher, M. C., Albright, R. G., Broadhurst, M. J., … McCune, J. M. (2013). Dysbiosis of the gut microbiota is associated with HIV disease progression and tryptophan catabolism. *Science Translational Medicine*, *5*(193), 193ra91. doi:10.1126/scitranslmed.3006438

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164. doi:10.1093/nar/gkq603

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., … Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, *42*(Database issue), D1001–6. doi:10.1093/nar/gkt1229

Williamson, S., Fledel-Alon, A., & Bustamante, C. D. (2004). Population genetics of polymorphism and divergence for diploid selection models with arbitrary dominance. *Genetics*, *168*(1), 463–75. doi:10.1534/genetics.103.024745

Wolpert, D. H., & Macready, W. G. (1997). No Free Lunch Theorems For Optimization. *IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION*, *1*(1), 67–82.

Wu, J., Sinfield, J. L., Buchanan-Wollaston, V., & Feng, J. (2009). Impact of environmental inputs on reverse-engineering approach to network structures. *BMC Systems Biology*, *3*, 113. doi:10.1186/1752-0509-3-113

Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews. Genetics*, *13*(5), 329–42. doi:10.1038/nrg3174

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–91. doi:10.1093/molbev/msm088

Yu, C., Zavaljevski, N., Desai, V., & Reifman, J. (2011). QuartetS: a fast and accurate algorithm for large-scale orthology detection. *Nucleic Acids Research*, *39*(13), e88. doi:10.1093/nar/gkr308

Zhao, S., Sakai, A., Zhang, X., Vetting, M. W., Kumar, R., Hillerich, B., … Jacobson, M. P. (2014). Prediction and characterization of enzymatic activities guided by sequence similarity and genome neighborhood networks. *eLife*, *3*. doi:10.7554/eLife.03275

Introduction

Establishing health-related causal relationships is a pivotal objective in biomedical research. Yet, the interdependent non-linearity of biological systems often impedes a thorough understanding of causal dynamics. Existing and forthcoming time series data will likely play an important role in taming this complexity. While one-time or repeated cross-sectional sampling may average out non-linear patterns by pooling data across subjects, long time series from a single source allow us to observe dynamic and context-specific patterns of change.

We are just beginning to understand the biomedical relevance of such a dynamical systems perspective. Consider for example the human microbiome. Dysbiosis in the gut has been implicated in, e.g. irritable bowel disease (IBD), obesity, diabetes, asthma, anxiety, and depression (Arrieta et al., 2014; Foster & McVey Neufeld, 2013). Meanwhile, recent studies on microbiome dynamics have found that the ecological makeup of the human microbiome is dynamic and individual-specific (Caporaso et al., 2011; Fisher & Mehta, 2014; Gajer et al., 2012). These dynamics may also interact with pathogens in interesting and therapeutically important ways. For example, there is evidence that ecological time series dynamics within the body may play a role in the progression from HIV to AIDS (Vujkovic-Cvijin *et al.*, 2013).

Complex, dynamically evolving interdependent systems such as the microbiome pose a significant challenge to existing time series methods. Several metrics exist for detecting static non-linear relationships. These include: spearman correlation (Spearman, 1904), distance correlation (Székely, Rizzo, & Bakirov, 2007), and mutual information content (Reshef *et al.*, 2011). Causal relationships, on the other hand, can be examined using methods such as time-lagged regression (Granger, 1969), instrumental variables (Bowden & Turkington, 1990), and dynamical Bayesian networks (Wu, Sinfield, Buchanan-Wollaston, & Feng, 2009).

These causal methods are heavily model-based, however. As a result, they often falter when examining arbitrary non-linear or context-dependent relationships. Furthermore, the approaches mentioned above cannot adequately handle feedback loops, and they frequently generate both false positives and false negatives due to the influence of unmeasured confounders (Vanderweele & Arah, 2011). These are significant liabilities, particularly in biomedicine, where relationships are often embedded within a broad network of only partially observed interactions.

In this paper, we present the first publicly available, open source implementation of convergent cross mapping (CCM), a model-free approach to detecting dependencies and inferring causality in complex non-linear systems (even in the presence of feedback loops and unmeasured confounding; Sugihara *et al.*, 2012). CCM derives this power from explicitly capturing time-dependent dynamics through a technique known as state-space reconstruction (SSR). SSR has demonstrated utility for problems as diverse as wildlife management (Deyle et al., 2013; Dixon, Milicich, & Sugihara, 1999) and cerebral autoregulation (Heskamp *et al.*, 2013). In practice, this analysis typically requires at least 25 data points, measured with sufficient density to capture system dynamics.

CCM builds on SSR, leveraging the fact that time series can be viewed as projections of higher-dimensional system dynamics (Sugihara *et al.*, 2012). As a result of this property, the time series of individual variables must contain information about the full causal system. Causal dynamics (conceptualized as the state space, or manifold) can then be reconstructed using individual time series. These reconstructions can be thought of as shadows of the true causal system. If the shadows reconstructed from distinct variables can be used to predict points from each other's time series, we can infer that these variables provide views of the same causal system and so are causally related. Since these relationships are fundamentally asymmetric, this test can also establish the directionality of causation.

Further details on CCM are available in the supplementary material of this paper, as well as in that of Sugihara *et al.* 2012. Additional explanatory resources can also be accessed through the project website (http://cyrusmaher.github.io/CauseMap.jl).

Implementation

CauseMap implements CCM in Julia, a high-performance programming language designed for facile technical computing (Bezanson, Karpinski, Shah, & Edelman, 2012). Via intelligent JIT (just in time) compilation, Julia offers much of the speed of low-level, low-productivity languages like C, while also providing the ease of use and platform independence of much slower high-level languages like Python, R, or Matlab.

Beyond the speed and comparative simplicity resulting from cutting-edge JIT compilation, CauseMap offers a number of conveniences and performance enhancements. For CCM, it is particularly important to optimize two tuning parameters: E and τ_p. E is related to the assumed dimensionality of the full causal system. This quantity is used to determine the dimensions of the reconstructed manifolds. τ_p, on the other hand, denotes the time delay of the causal effect of interest. By examining the optimal values of these two parameters, we may place bounds on the number of variables involved in the full causal system. In addition, we gain insight into the timeframe of causal effects.

CauseMap precomputes all necessary manifolds and pairwise distances using a state-of-the-art, BLAS-based protocol (for benchmarks, see: https://github.com/JuliaStats/Distance.jl). E and τ_p are then optimized by multiple iterations of cyclic coordinate descent (Bertsekas, 1999). Note that while convergence of the cross map signal as a function of the time series length (L) is taken as a practical criterion for causality, the dependence of this signal on E and τ_p is also useful for qualitatively estimating the credibility of the observed signal. CauseMap therefore also includes a plotting function to visualize the dependence of the predictive skill (ρ_{ccm}) on L, as well as on the joint values of E and τ_p.

Results and Discussion

To demonstrate CauseMap's functionality and performance, we examined the predator-prey relationship between *Paramecium aurelia* and *Didinium nasutum* (George Sugihara et al., 2012). Observations were collected every 12 hours for 30 days, yielding a total of 60 data points. Plotted in Figure 4-1 is the CauseMap visualization of the dependence of predictive skill (ρ_{ccm}) on *L, E,* and τ_p. In Figure 4-1A, we observe convergence in ρ_{ccm} with respect to *L*, the number of data points used for prediction of held-out observations. This convergence is a practical criterion for causality and the source of the name *convergent* cross mapping. Figures 4-1B and 1C show the dependence of the max ρ_{ccm} on E (proportional to the assumed dimensionality of the system), and the supposed time lag of the causal effect (τ_p). While the max ρ_{ccm} is relatively insensitive to the assumed dimensionality, the best-performing τ_p values correspond to either immediate causal effects, or those delayed by five days. Note that τ_p=5 corresponds to the principal frequency of the *Paramecium aurelia* and *Didinium nasutum* time series, as determined by fourier transform analysis (see supplemental materials for further details).

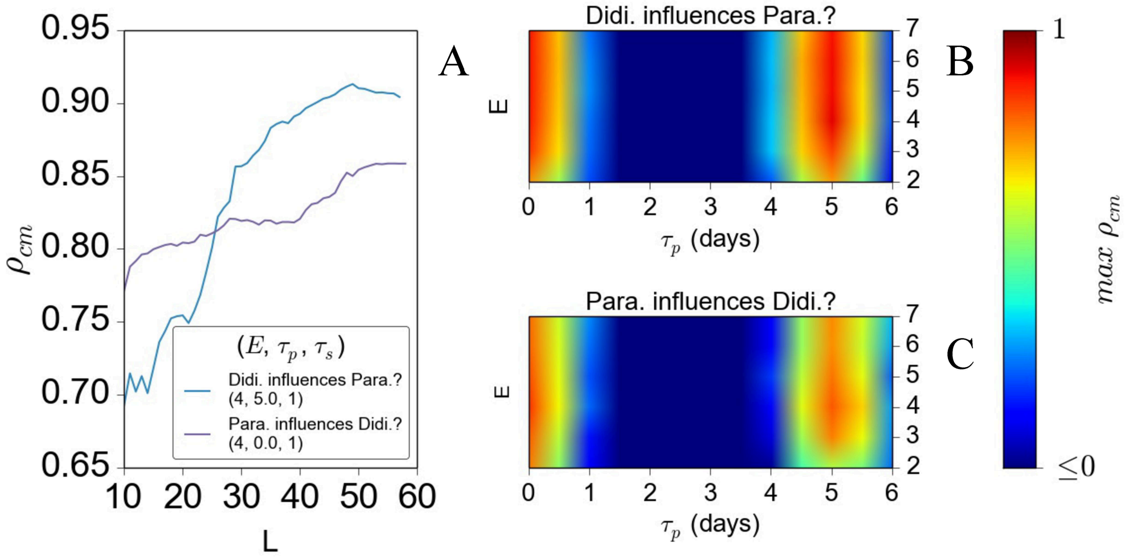


Figure 3-4. An example visualization from CauseMap using abundances of *Paramecium aurelia* and *Didinium nasutum* *(see supplemental materials for more information on this system). A.) For optimal parameter values, the convergence of the cross-map correlation with library size. B-C.) The dependence of the maximum cross-map correlation on assumed dimensionality (measured by E) and the time lag of the causal effect (measured by τ_p). Note that the second maximum at τ_p=5 corresponds to the principal frequency of the P. aurelia and D. nasutum time series, as determined by fourier transform analysis.*

Performance

Approximately 300 CCM evaluations were conducted to produce Figure 4-1. These calculations finished in less than 30 seconds on a single 2.6 GHz processor. Each of these evaluations involved the prediction of over 60,000 points, compiled across all sliding windows of libraries of varying lengths. At an average of 1.7 microseconds per prediction, this is a highly efficient implementation given the computational challenges.

Intended use and Benefits

CauseMap is designed to examine causal relationships in time series with 25 or more observations. In order to illustrate the effects of shorter time series, we thinned the *Paramedium-Didinium* data set by one-half and by one-third, yielding series of 30 and 20 observations, respectively. Figure 4-2 demonstrates the effect of this reduction on the convergence of predictive skill (ρ_{ccm}). We see that the 1/2 thinned data set recapitulates the trends observed in the full series, including the relative magnitudes of ρ_{ccm} between the mappings of *Didinium* to *Paramecium* and vice versa. The 1/3 thinned sample set, on the other hand, no longer demonstrates convergence. In addition, compared to the longer sets, it exhibits the opposite trend in relative predictive skill between the two mappings. Patterns in max ρ_{ccm} versus E and τ_p are approximately conserved, however (fig. SC-1).

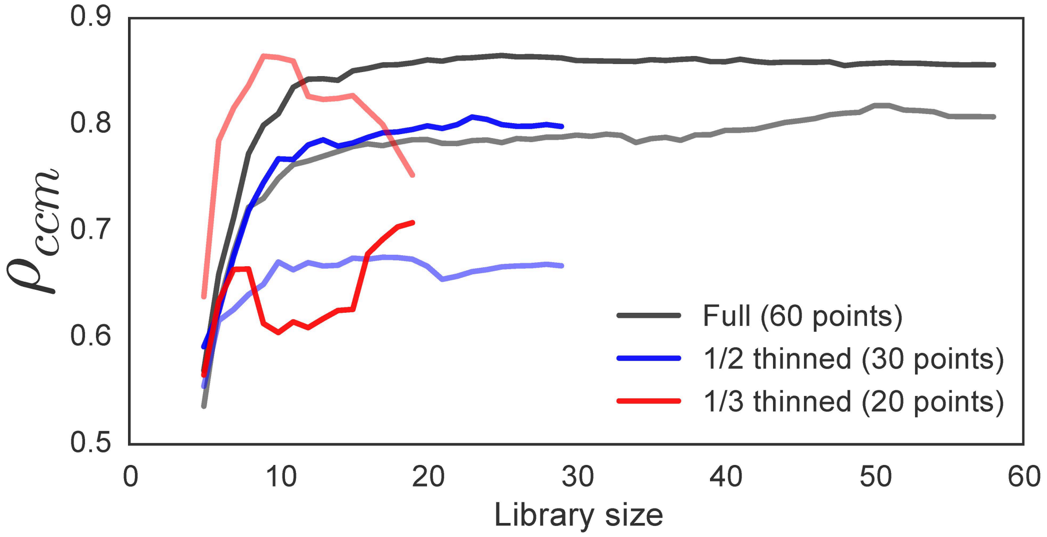


Figure 3-5. The effect of time series length on ρ_{ccm} convergence. *Black, blue, and red lines illustrate ρ_{ccm} for the full, 1/2 thinned, and 1/3 thinned datasets, respectively. For a given color, darker lines show ρ_{ccm} for the test of whether Didinium abundance influences Paramecium abundance. Lighter lines examine the converse.*

This points to the advantage of examining how predictive skill depends on all relevant variables, as opposed to *L* alone. At shorter time series lengths, we are better able to qualitatively differentiate weak signal from random noise if we examine *L,* E, and τ_p together. This may streamline the process of hypothesis generation and validation, allowing for a more intelligent allocation of resources for exploring and verifying causal relationships.

Despite its requirement for relatively long time series (>25 observations), CauseMap has the advantage of requiring only a single time series for each variable. In dynamical systems with widely varying or context-specific behavior, this would allow researchers to draw conclusions that are tailored to, e.g. a given patient. Rather than acting on population averages, biomedical researchers would be free to fully personalize therapy to the unique biology and ecology of the patient. One example of this is in the treatment of microbiome dysbiosis. Imbalances in the microbiome have been implicated in, e.g. irritable bowel disease (IBD), obesity, diabetes, asthma, anxiety, and depression (Arrieta et al., 2014; Foster & McVey Neufeld, 2013). While fecal transplantation therapy is effective in treating specific types of dysbiosis (Khoruts & Weingarden, 2014), next generation therapeutics may offer a blend of purified strains, tailored to the gut ecology of the patient. We believe CauseMap has the potential to be a valuable tool for designing such breakthrough therapies.

Additional examples include understanding patient-to-patient variability in drug response using time series metabolomics, and examining the basis of e.g. influenza seasonality using global time series. We expect that such applications will continue to proliferate as the costs of data collection decrease over the coming years. For this reason, we believe it is vitally important that the biomedical research community have access to an efficient implementation of CCM that is user-friendly and available for immediate field testing.

Planned future development

In future versions, we will include S-map calculations to evaluate the non-linearity of the causal system (G. Sugihara, 1994). We will also add a bootstrap-based procedure for library selection, as opposed to the current approach using sliding windows. This has been shown to reduce the effect of secular trends on the cross map correlation (Hao Ye, George Sugihara, *personal communication*). In addition, we will re-implement the plotting functionality in Julia, removing the requirements of Python and matplotlib for visualization. Finally, we will design Python and R wrappers for CauseMap functions so that our codebase can be easily leveraged from those environments as well. User suggestions will also be considered as we decide how best to develop the tool.

Conclusions

CauseMap provides a fast, user-friendly implementation of CCM, a powerful new method for exploring dependencies and even establishing causality in complex, highly non-linear datasets with many unobserved variables. We believe that CCM holds a great deal of promise for a wide range of applications, including personalized microbiome therapy and metabolic dynamics analysis. As novel time series datasets continue to emerge, it is our hope that CauseMap will allow researchers to uncover interesting and biomedically actionable causal relationships using this next-generation time series method.

Availability and Requirements

Project name: CauseMap

Project home page: http://cyrusmaher.github.io/CauseMap.jl/

Operating system(s): Platform independent

Programming language: Julia

Other requirements: Python and matplotlib (for graphing)

License: MIT

Any restrictions to use by non-academics: No

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In F. Czaki & B. N. Petrov (Eds.), *2nd International Symposium on Information Theory* (pp. 267–281). Budapest: Akademiai Kiado.

Akey, J. M. (2009). Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Research*, *19*(5), 711–22. doi:10.1101/gr.086652.108

Alexeyenko, A., Tamas, I., Liu, G., & Sonnhammer, E. L. L. (2006). Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics*, *22*(14), e9–15. doi:10.1093/bioinformatics/btl213

Altenhoff, A. M., & Dessimoz, C. (2009a). Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Computational Biology*, *5*(1), e1000262. doi:10.1371/journal.pcbi.1000262

Altenhoff, A. M., & Dessimoz, C. (2009b). Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Computational Biology*, *5*(1), e1000262. doi:10.1371/journal.pcbi.1000262

Altenhoff, A. M., & Dessimoz, C. (2012). Inferring Orthology. In M. Anisimova (Ed.), *Evolutionary Genomics* (Vol. 855). Totowa, NJ: Humana Press. doi:10.1007/978-1-61779-582-4

Altenhoff, A. M., Schneider, A., Gonnet, G. H., & Dessimoz, C. (2011). OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Research*, *39*(Database issue), D289–94. doi:10.1093/nar/gkq1238

Arrieta, M.-C., Stiemsma, L. T., Amenyogbe, N., Brown, E. M., & Finlay, B. (2014). The Intestinal Microbiome in Early Life: Health and Disease. *Frontiers in Immunology*, *5*, 427. doi:10.3389/fimmu.2014.00427

Babbitt, C. C., Warner, L. R., Fedrigo, O., Wall, C. E., & Wray, G. A. (2011). Genomic signatures of diet-related shifts during human origins. *Proceedings. Biological Sciences / The Royal Society*, *278*(1708), 961–9. doi:10.1098/rspb.2010.2433

Bakewell, M. A., Shi, P., & Zhang, J. (2007). More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(18), 7489–94. doi:10.1073/pnas.0701705104

Berman, H. M. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. doi:10.1093/nar/28.1.235

Bertsekas, D. (1999). *Nonlinear Programming* (p. 780). Athena Scientific; 2nd edition.

Bezanson, J., Karpinski, S., Shah, V. B., & Edelman, A. (2012). Julia: A Fast Dynamic Language for Technical Computing. Programming Languages; Computational Engineering, Finance, and Science.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F. a, Roskin, K. M., … Miller, W. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research*, *14*(4), 708–15. doi:10.1101/gr.1933104

Bowden, R. J., & Turkington, D. A. (1990). *Instrumental Variables*. Cambridge University Press.

Boyko, A. R., Williamson, S. H., Indap, A. R., Degenhardt, J. D., Hernandez, R. D., Lohmueller, K. E., … Bustamante, C. D. (2008). Assessing the evolutionary impact

of amino acid mutations in the human genome. *PLoS Genetics*, *4*(5), e1000083. doi:10.1371/journal.pgen.1000083

Bradley, B. J. (2008). Reconstructing phylogenies and phenotypes: a molecular view of human evolution. *Journal of Anatomy*, *212*(4), 337–53. doi:10.1111/j.1469-7580.2007.00840.x

Burkart, J. M., Allon, O., Amici, F., Fichtel, C., Finkenwirth, C., Heschl, A., … van Schaik, C. P. (2014). The evolutionary origin of human hyper-cooperation. *Nature Communications*, *5*, 4747. doi:10.1038/ncomms5747

Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M. T., Glanowski, S., … Clark, A. G. (2005a). Natural selection on protein-coding genes in the human genome. *Nature*, *437*(7062), 1153–7.

Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M. T., Glanowski, S., … Clark, A. G. (2005b). Natural selection on protein-coding genes in the human genome. *Nature*, *437*(7062), 1153–7. doi:10.1038/nature04240

Caporaso, J. G., Lauber, C. L., Costello, E. K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., … Knight, R. (2011). Moving pictures of the human microbiome. *Genome Biology*, *12*(5), R50. doi:10.1186/gb-2011-12-5-r50

Capra, J. A., Stolzer, M., Durand, D., & Pollard, K. S. (2013). How old is my gene? *Trends in Genetics : TIG*, *29*(11), 659–68. doi:10.1016/j.tig.2013.07.001

Casdagli, M., Eubank, S., Farmer, J. D., & Gibson, J. (1991). State space reconstruction in the presence of noise. *Physica D: Nonlinear Phenomena*, *51*(1-3), 52–98. doi:10.1016/0167-2789(91)90222-U

Chandrasekaran, V., & Jordan, M. I. (2013). Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(13), E1181–90. doi:10.1073/pnas.1302293110

Charlesworth, J., & Eyre-Walker, A. (2008). The McDonald-Kreitman test and slightly deleterious mutations. *Molecular Biology and Evolution*, *25*(6), 1007–15. doi:10.1093/molbev/msn005

Chen, F., Mackey, A. J., Vermunt, J. K., & Roos, D. S. (2007). Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PloS One*, *2*(4), e383. doi:10.1371/journal.pone.0000383

Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B., & Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science (New York, N.Y.)*, *311*(5765), 1283–7. doi:10.1126/science.1123061

Costanzo, M. J., Yabut, S. C., Zhang, H.-C., White, K. B., de Garavilla, L., Wang, Y., … Maryanoff, B. E. (2008). Potent, nonpeptide inhibitors of human mast cell tryptase. Synthesis and biological evaluation of novel spirocyclic piperidine amide derivatives. *Bioorganic & Medicinal Chemistry Letters*, *18*(6), 2114–21. doi:10.1016/j.bmcl.2008.01.093

Dayhoff, M. O., Schwartz, R. M., & Orcutt, B. C. (1978). A model of evolutionary change in proteins. In M. O. Dayhoff (Ed.), *Atlas of protein sequence and structure* (pp. 345–358). Nature Biomedical Research.

Deyle, E. R., Fogarty, M., Hsieh, C., Kaufman, L., MacCall, A. D., Munch, S. B., … Sugihara, G. (2013). Predicting climate effects on Pacific sardine. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(16), 6430–5. doi:10.1073/pnas.1215506110

Dixon, P. A., Milicich, M. J., & Sugihara, G. (1999). Episodic Fluctuations in Larval Supply. *Science*, *283*(5407), 1528–1530. doi:10.1126/science.283.5407.1528

Ebersberger, I., Strauss, S., & von Haeseler, A. (2009). HaMStR: profile hidden markov model based search for orthologs in ESTs. *BMC Evolutionary Biology*, *9*(1), 157. doi:10.1186/1471-2148-9-157

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, *7*(10), e1002195. doi:10.1371/journal.pcbi.1002195

Eden, E., Navon, R., Steinfeld, I., Lipson, D., & Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, *10*(1), 48. doi:10.1186/1471-2105-10-48

Eilertson, K. E., Booth, J. G., & Bustamante, C. D. (2012). SnIPRE: selection inference using a Poisson random effects model. *PLoS Computational Biology*, *8*(12), e1002806. doi:10.1371/journal.pcbi.1002806

Finch, C. E. (2010). Evolution in health and medicine Sackler colloquium: Evolution of the human lifespan and diseases of aging: roles of infection, inflammation, and nutrition. *Proceedings of the National Academy of Sciences of the United States of America*, *107 Suppl* (suppl_1), 1718–24. doi:10.1073/pnas.0909606106

Fisher, C. K., & Mehta, P. (2014). Identifying Keystone Species in the Human Gut Microbiome from Metagenomic Timeseries using Sparse Linear Regression.

Foster, J. A., & McVey Neufeld, K.-A. (2013). Gut-brain axis: how the microbiome influences anxiety and depression. *Trends in Neurosciences*, *36*(5), 305–12. doi:10.1016/j.tins.2013.01.005

Gajer, P., Brotman, R. M., Bai, G., Sakamoto, J., Schütte, U. M. E., Zhong, X., … Ravel, J. (2012). Temporal dynamics of the human vaginal microbiota. *Science Translational Medicine*, *4*(132), 132ra52. doi:10.1126/scitranslmed.3003605

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods Title. *Econometrica*, *37*(3), 424–438.

Haygood, R., Babbitt, C. C., Fedrigo, O., & Wray, G. A. (2010). Contrasts between adaptive coding and noncoding changes during human evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(17), 7853–7. doi:10.1073/pnas.0911249107

Henikoff, S. (1992). Amino Acid Substitution Matrices from Protein Blocks. *Proceedings of the National Academy of Sciences*, *89*(22), 10915–10919.

Hernandez, R. D. (2008). A flexible forward simulator for populations subject to selection and demography. *Bioinformatics (Oxford, England)*, *24*(23), 2786–7. doi:10.1093/bioinformatics/btn522

Heskamp, L., Meel-van den Abeelen, A., Katsogridakis, E., Panerai, R., Simpson, D., Lagro, J., & Claassen, J. (2013). Convergent cross mapping: a promising technique for future cerebral autoregulation estimation. *CEREBROVASCULAR DISEASES*, *35*, 15–16.

Hubisz, M. J., & Pollard, K. S. (2014). Exploring the genesis and functions of Human Accelerated Regions sheds light on their role in human evolution. *Current Opinion in Genetics & Development*, *29C*, 15–21. doi:10.1016/j.gde.2014.07.005

Hulsen, T., Huynen, M. A., de Vlieg, J., & Groenen, P. M. A. (2006). Benchmarking ortholog identification methods using functional genomics data. *Genome Biology*, *7*(4), R31. doi:10.1186/gb-2006-7-4-r31

Jordan, G., & Goldman, N. (2012). The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Molecular Biology and Evolution*, *29*(4), 1125–39. doi:10.1093/molbev/msr272

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Research*, *12*(4), 656–64. doi:10.1101/gr.229202. Article published online before March 2002

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, a. D. (2002). The Human Genome Browser at UCSC. *Genome Research*, *12*(6), 996–1006. doi:10.1101/gr.229102

Khoruts, A., & Weingarden, A. R. (2014). Emergence of fecal microbiota transplantation as an approach to repair disrupted microbial gut ecology. *Immunology Letters*. doi:10.1016/j.imlet.2014.07.016

Kosiol, C., Vinar, T., da Fonseca, R. R., Hubisz, M. J., Bustamante, C. D., Nielsen, R., & Siepel, A. (2008). Patterns of positive selection in six Mammalian genomes. *PLoS Genetics*, *4*(8), e1000144. doi:10.1371/journal.pgen.1000144

Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Machine Learning*, *51*(2), 181–207.

Kuzniar, A., van Ham, R. C. H. J., Pongor, S., & Leunissen, J. a M. (2008). The quest for orthologs: finding the corresponding gene across genomes. *Trends in Genetics : TIG*, *24*(11), 539–51. doi:10.1016/j.tig.2008.08.009

Liu, Y., Schmidt, B., & Maskell, D. L. (2010). MSAProbs: multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities. *Bioinformatics (Oxford, England)*, *26*(16), 1958–64. doi:10.1093/bioinformatics/btq338

Maddison, W. P., & Knowles, L. L. (2006). Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology*, *55*(1), 21–30. doi:10.1080/10635150500354928

Maher, M. C., & Hernandez, R. D. (2013). A MOSAIC of methods: Improving ortholog detection through integration of algorithmic diversity. Populations and Evolution; Quantitative Methods.

Mashiyama, S. T., Malabanan, M. M., Akiva, E., Bhosle, R., Branch, M. C., Hillerich, B., … Babbitt, P. C. (2014). Large-scale determination of sequence, structure, and function relationships in cytosolic glutathione transferases across the biosphere. *PLoS Biology*, *12*(4), e1001843. doi:10.1371/journal.pbio.1001843

Massingham, T., & Goldman, N. (2005). Detecting amino acid sites under positive selection and purifying selection. *Genetics*, *169*(3), 1753–62. doi:10.1534/genetics.104.032144

McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, *351*(6328), 652–4.

McEntyre, .J, & Ostell, J. (Eds.). (2002). *The NCBI Handbook*. Bethesda, MD: National Center for Biotechnology Information.

Messer, P. W., & Petrov, D. A. (2013). Frequent adaptation and the McDonald-Kreitman test. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(21), 8615–20. doi:10.1073/pnas.1220835110

Nielsen, R., Bustamante, C., Clark, A. G., Glanowski, S., Sackton, T. B., Hubisz, M. J., … Cargill, M. (2005a). A scan for positively selected genes in the genomes of

humans and chimpanzees. *PLoS Biology*, *3*(6), e170. doi:10.1371/journal.pbio.0030170

Nielsen, R., Bustamante, C., Clark, A. G., Glanowski, S., Sackton, T. B., Hubisz, M. J., … Cargill, M. (2005b). A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biology*, *3*(6), e170. doi:10.1371/journal.pbio.0030170

Nielsen, R., Hubisz, M. J., Hellmann, I., Torgerson, D., Andrés, A. M., Albrechtsen, A., … Clark, A. G. (2009). Darwinian and demographic forces affecting human protein coding genes. *Genome Research*, *19*(5), 838–49. doi:10.1101/gr.088336.108

Preuss, T. M. (2011). The human brain: rewired and running hot. *Annals of the New York Academy of Sciences*, *1225 Suppl*, E182–91. doi:10.1111/j.1749-6632.2011.06001.x

Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., … Ostell, J. M. (2014). RefSeq: an update on mammalian reference sequences. *Nucleic Acids Research*, *42*(Database issue), D756–63. doi:10.1093/nar/gkt1114

Pruitt, K. D., Harrow, J., Harte, R. A., Wallin, C., Diekhans, M., Maglott, D. R., … Lipman, D. (2009). The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Research*, *19*(7), 1316–23. doi:10.1101/gr.080531.108

Pryszcz, L. P., Huerta-Cepas, J., & Gabaldón, T. (2011). MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. *Nucleic Acids Research*, *39*(5), e32. doi:10.1093/nar/gkq953

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., … Finn, R. D. (2012). The Pfam protein families database. *Nucleic Acids Research*, *40*(Database issue), D290–301. doi:10.1093/nar/gkr1065

Remm, M., Storm, C. E., & Sonnhammer, E. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology*, *314*(5), 1041–52. doi:10.1006/jmbi.2000.5197

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., … Sabeti, P. C. (2011). Detecting novel associations in large data sets. *Science (New York, N.Y.)*, *334*(6062), 1518–24. doi:10.1126/science.1205438

Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, *53*(1-2), 131–147. doi:10.1016/0025-5564(81)90043-2

Rokach, L. (2009). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*(1-2), 1–39. doi:10.1007/s10462-009-9124-7

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*, 72–101.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)*, *30*(9), 1312–3. doi:10.1093/bioinformatics/btu033

Stamatakis, A., & Alachiotis, N. (2010). Time and memory efficient likelihood-based tree searches on phylogenomic alignments with missing data. *Bioinformatics (Oxford, England)*, *26*(12), i132–9. doi:10.1093/bioinformatics/btq205

Stolzer, M., Lai, H., Xu, M., Sathaye, D., Vernot, B., & Durand, D. (2012). Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics (Oxford, England)*, *28*(18), i409–i415. doi:10.1093/bioinformatics/bts386

Stone, M. (1977). An Asymptotic Equivalence of Choice of Model by Cross-validation and Akaike's Criterion. *Journal of the Royal Statistical Society*. Retrieved March 30, 2014, from http://www.jstor.org/discover/10.2307/2984877?uid=3739560&uid=2134&uid=2&uid=70&uid=4&uid=3739256&sid=21103766217637

Sugihara, G. (1994). Nonlinear Forecasting for the Classification of Natural Time Series. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *348*(1688), 477–495. doi:10.1098/rsta.1994.0106

Sugihara, G., May, R., Ye, H., Hsieh, C., Deyle, E., Fogarty, M., & Munch, S. (2012). Detecting causality in complex ecosystems. *Science (New York, N.Y.)*, *338*(6106), 496–500. doi:10.1126/science.1227079

Sukumaran, J., & Holder, M. T. (2010). DendroPy: a Python library for phylogenetic computing. *Bioinformatics (Oxford, England)*, *26*(12), 1569–71. doi:10.1093/bioinformatics/btq228

Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS One*, *6*(7), e21800. doi:10.1371/journal.pone.0021800

Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, *34*(Web Server issue), W609–12. doi:10.1093/nar/gkl315

Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and Testing Dependence by Correlation of Distances. *The Annals of Statistics*, *35*(6), 2769–2794.

The Chimpanzee Sequencing and Analysis Consortium. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, *437*(7055), 69–87. doi:10.1038/nature04072

Torgerson, D. G., Boyko, A. R., Hernandez, R. D., Indap, A., Hu, X., White, T. J., … Clark, A. G. (2009). Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. *PLoS Genetics*, *5*(8), e1000592. doi:10.1371/journal.pgen.1000592

Trivedi, N. N., Tong, Q., Raman, K., Bhagwandin, V. J., & Caughey, G. H. (2007). Mast cell alpha and beta tryptases changed rapidly during primate speciation and evolved from gamma-like transmembrane peptidases in ancestral vertebrates. *Journal of Immunology (Baltimore, Md. : 1950)*, *179*(9), 6072–9.

Van der Laan, M. J., & Gruber, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *The International Journal of Biostatistics*, *6*(1), Article 17. doi:10.2202/1557-4679.1181

Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, *6*, Article25. doi:10.2202/1544-6115.1309

Vanderweele, T. J., & Arah, O. A. (2011). Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology (Cambridge, Mass.)*, *22*(1), 42–52. doi:10.1097/EDE.0b013e3181f74493

Varki, A. (2012). Nothing in medicine makes sense, except in the light of evolution. *Journal of Molecular Medicine (Berlin, Germany)*, *90*(5), 481–94. doi:10.1007/s00109-012-0900-5

Vujkovic-Cvijin, I., Dunham, R. M., Iwai, S., Maher, M. C., Albright, R. G., Broadhurst, M. J., … McCune, J. M. (2013). Dysbiosis of the gut microbiota is associated with HIV disease progression and tryptophan catabolism. *Science Translational Medicine*, *5*(193), 193ra91. doi:10.1126/scitranslmed.3006438

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164. doi:10.1093/nar/gkq603

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., … Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, *42*(Database issue), D1001–6. doi:10.1093/nar/gkt1229

Williamson, S., Fledel-Alon, A., & Bustamante, C. D. (2004). Population genetics of polymorphism and divergence for diploid selection models with arbitrary dominance. *Genetics*, *168*(1), 463–75. doi:10.1534/genetics.103.024745

Wolpert, D. H., & Macready, W. G. (1997). No Free Lunch Theorems For Optimization. *IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION*, *1*(1), 67–82.

Wu, J., Sinfield, J. L., Buchanan-Wollaston, V., & Feng, J. (2009). Impact of environmental inputs on reverse-engineering approach to network structures. *BMC Systems Biology*, *3*, 113. doi:10.1186/1752-0509-3-113

Yandell, M., & Ence, D. (2012). A beginner’s guide to eukaryotic genome annotation. *Nature Reviews. Genetics*, *13*(5), 329–42. doi:10.1038/nrg3174

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–91. doi:10.1093/molbev/msm088

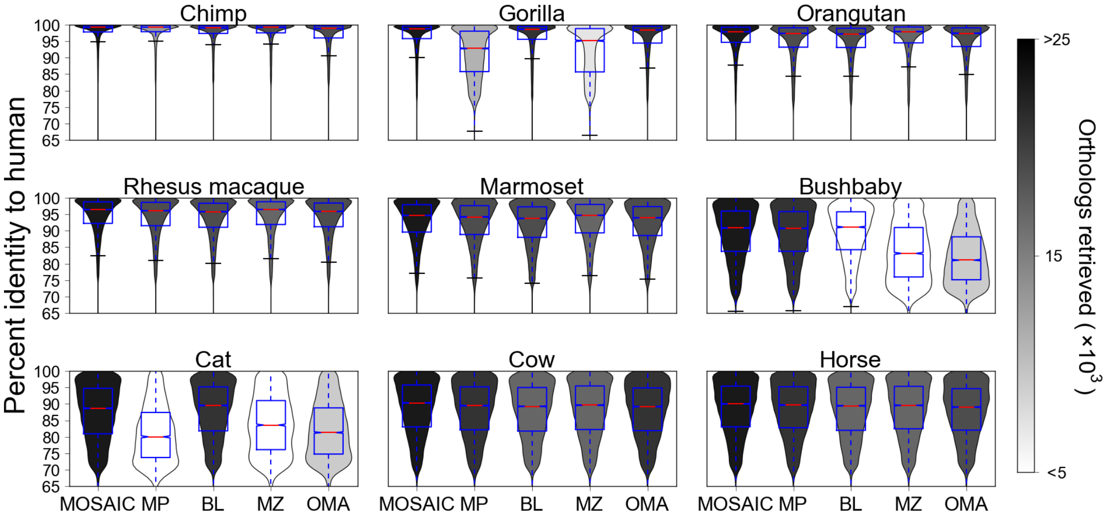
Yu, C., Zavaljevski, N., Desai, V., & Reifman, J. (2011). QuartetS: a fast and accurate algorithm for large-scale orthology detection. *Nucleic Acids Research*, *39*(13), e88. doi:10.1093/nar/gkr308

Zhao, S., Sakai, A., Zhang, X., Vetting, M. W., Kumar, R., Hillerich, B., … Jacobson, M. P. (2014). Prediction and characterization of enzymatic activities guided by sequence similarity and genome neighborhood networks. *eLife*, *3*. doi:10.7554/eLife.03275

Appendix A Supplemental Material to Chapter Two

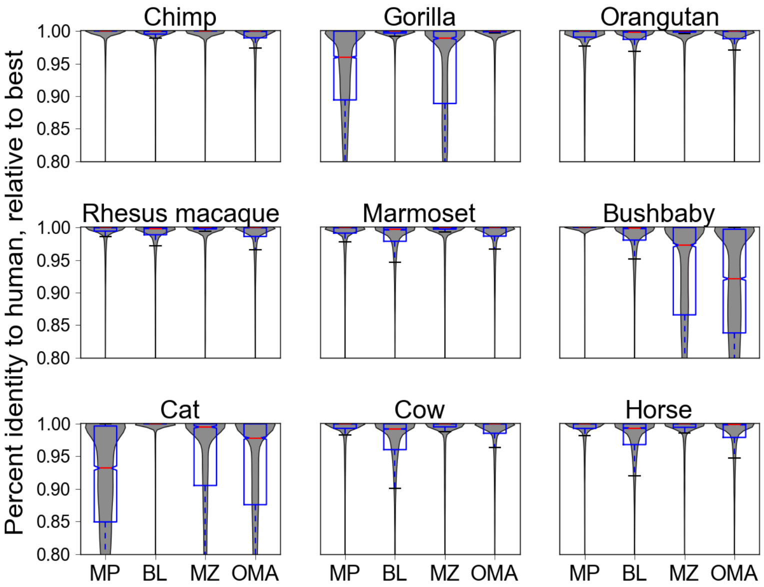
MOSAIC adds new sequences, maintains or increases average levels of sequence identity

Figure SA-1 demonstrates that, for each species, MOSAIC retrieves a much larger number of sequences than any method alone, while maintaining levels of percent identity comparable to those of the best performing method. It should be noted here that in our current examples, MOSAIC is designed to optimize the metric of sequence identity to human. Indeed, for a given putative ortholog, MOSAIC is guaranteed to improve or maintain percent identity compared to its constituent methods. Counter-intuitively, this provides no assurance that MOSAIC will provide gains in *average* levels of percent identity. For example, average levels of percent identity could decrease if MOSAIC ensures the inclusion of a greater number of species by pulling in poorly scoring sequences that were initially filtered out by the majority of component methods. However in Figure SA-1, we see that this is not the case.



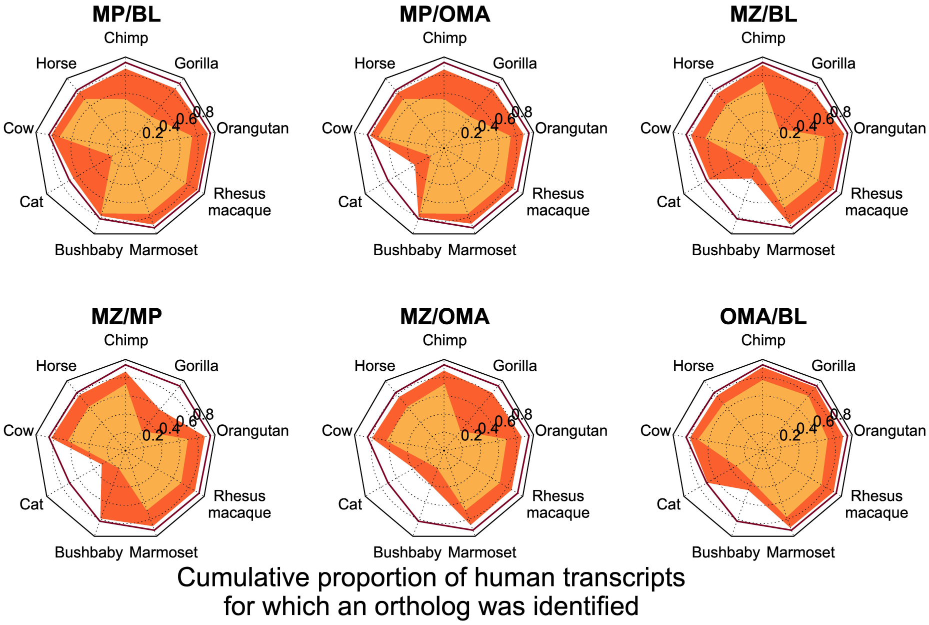
Supp. Figure A-1. Distributions of percent identity relative to the highest scoring ortholog, stratified by species. *This plot demonstrates how each method's performance compares to the best method. Each data point is a putative ortholog from a given species. Distributions are summarized by violinplots with boxplots overlaid.*

We next evaluated percent identity to human for each ortholog proposed by each method relative to the highest scoring ortholog from all methods. Figure SA-2 demonstrates that relative performance is species-specific. In particular, we note that the performance disparities across methods are much more pronounced for gorilla, bushbaby, and cat, both in terms of the number and quality of obtained orthologs.

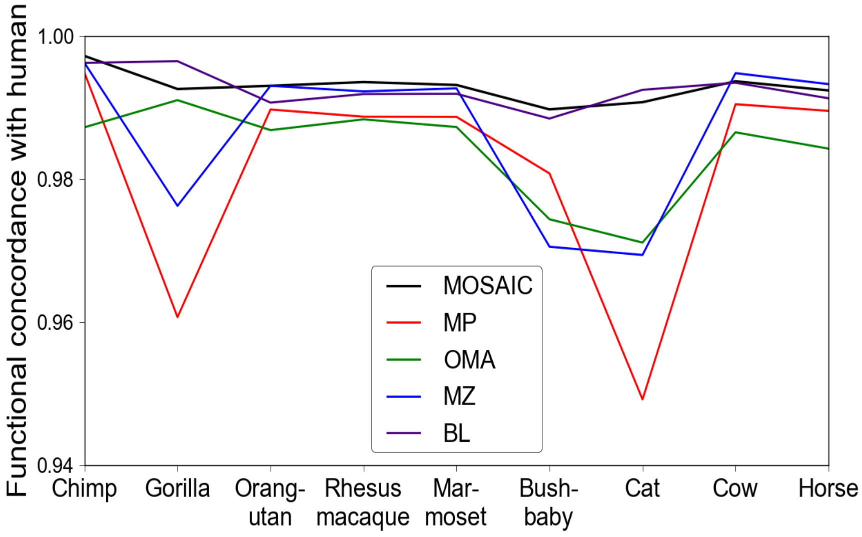


Supp. Figure A-2. **The effect of method integration on sequence identity.** *A comparison of the overall distributions of percent identity to human for MOSAIC and its component methods. Smoothed distributions underlying the boxplots are shaded according to the number of human transcripts for which an ortholog was proposed. White denotes 5000 sequences or less. Darker shades signify increasingly larger numbers of detected orthologs.*

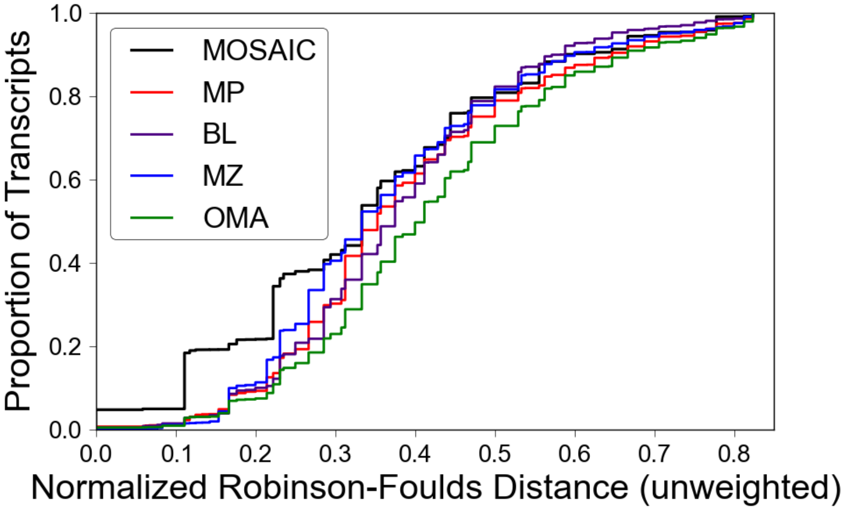
Examining each OD method in detail yields some hypotheses about the origin of these differences in performance. Errors in proteome prediction, both in terms of false-positives and false-negatives, are likely to have large effects on both MultiParanoid and OMA. Meanwhile, spurious syntenic information is expected to compromise the integrity of ortholog predictions produced by MultiZ. Finally, the lack of an assembled genome for bushbaby may negatively impact the quality of BLAT due to the segmentation of exon sets across multiple unordered scaffolds.



Supp. Figure A-3. **The cumulative proportion of transcripts for which an ortholog is identified.** *We show have all pairs of methods perform in retrieving orthologs for each species.*



Supp. Figure A-4. **The rate of concordance between functional annotations for proposal orthologs and human transcripts.**



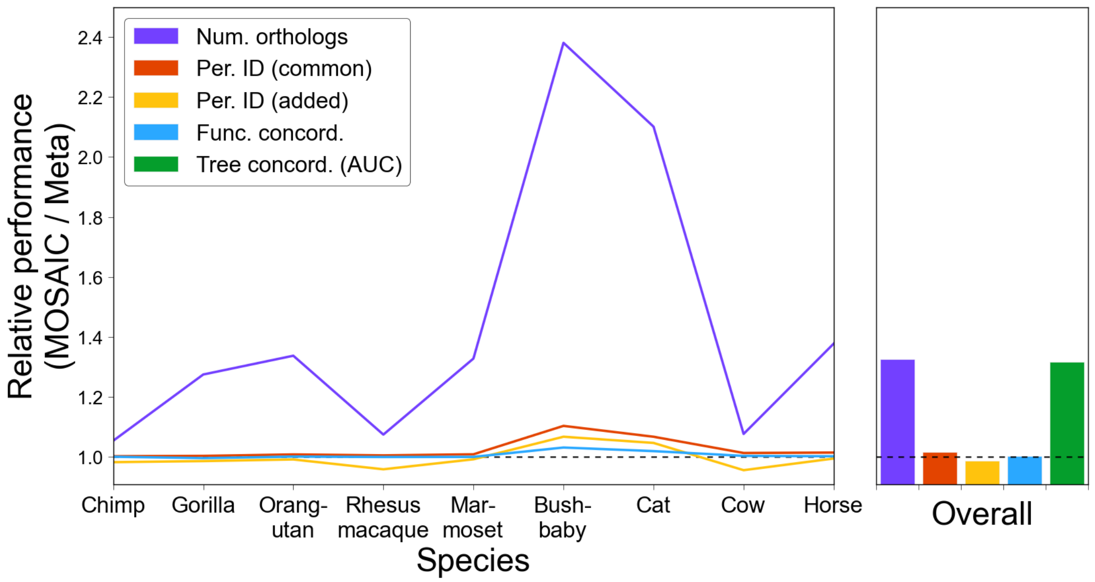
Supp. Figure A-5. **The cumulative proportion of human transcripts as a function of the maximum allowable Robinson-Foulds distance between the gene tree and the species tree.**

Figure SA-5 presents the cumulative proportion of alignments included as a function of the maximum allowable RF distance. Multiz is seen to perform the best of any individual method, likely due to its utilization of syntenic information. Surprisingly, the tree-based OD method, OMA, is seen to be the worst performing method according to this tree-based metric. Combining all methods using MOSAIC leads to a strong enrichment of highly concordant gene trees, while providing performance that is competitive with all component methods at more permissive RF distance cutoffs.

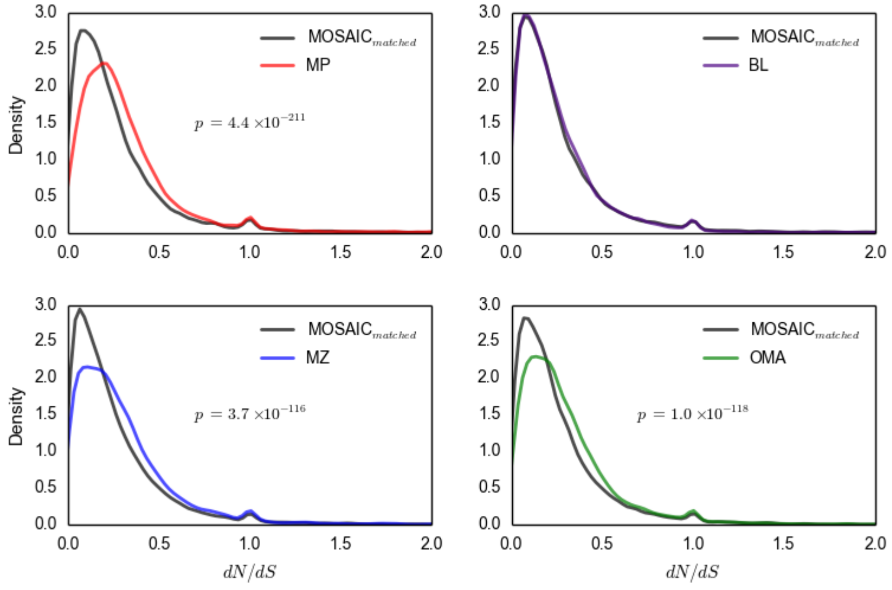
Comparison to a related method

We have shown that MOSAIC provides a large increase in the number of detected orthologs relative to its component methods, while simultaneously maintaining or improving functional-, phylogenetic-, and sequence identity-based measures of ortholog quality. Next, we sought to compare this method of OD integration to the only alternative of which we are aware: metaPhOrs (Pryszcz et al., 2011). Using an approach based on tree overlap, metaPhOrs integrates ortholog predictions using phylogenetic trees from seven databases: PhylomeDB, Ensembl, TreeFam, EggNOG, OrthoMCL, COG, and Fungal Orthogroups.

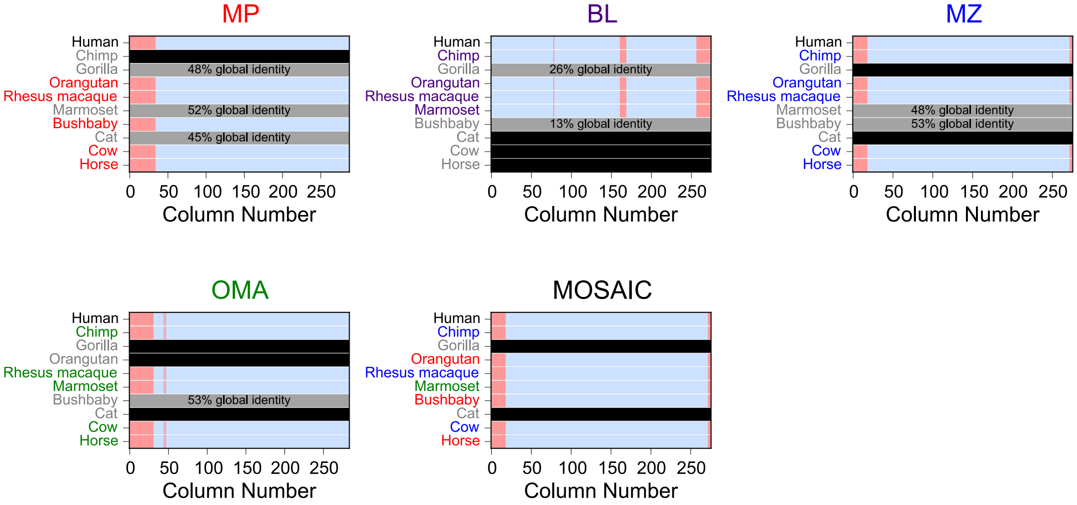
While MOSAIC is able to integrate an arbitrary number of OD methods of any time, metaPhOrs can only integrate tree-based methods. Since only pre-computed metaPhOrs data is available, we can also only examine the results of integrating the seven methods named above. This is then skewed comparison because MOSAIC only integrates four methods. Nevertheless, we compared MOSAIC and metaPhOrs based on the number of retrieved orthologs, average differences in sequence identity, and comparative levels of functional and phylogenetic concordance. We observe that MOSAIC provides large increases in the number of retrieved orthologs, while providing slight improvements in sequence identity for those cases where proposal orthologs are available from both methods (fig. SA-6). For the cases where MOSAIC predicted an ortholog but metaPhOrs did not, we examined the level of sequence identity in these sequences compared to the species-specific average returned by metaPhOrs. We find that these additional sequences display levels of sequence identity comparable to those provided by metaPhOrs. Finally, we observe that MOSAIC yields a slight increase in functional concordance, as well as a 40% increase in tree concordance, measured as the area under the curve below an RF distance of 0.5. A 0.5 threshold was chosen because there is little differentiation between methods after this point.



Supp. Figure A-6. **A comparison between MOSAIC and metaPhOrs**. *The relative performance between MOSAIC and metaPhOrs according to five metrics: 1.) the number of orthologs detected (purple); 2.) the percent identity to human for orthologs present in both (red); 3.) the percent identity to human for orthologs unique to MOSAIC compared to metaPhOrs species-specific average (yellow); 4.) rate of functional concordance between proposal orthologs and human transcripts (blue); and 5.) concordance between gene and species trees, as measured by a normalized, unweighted Robinson-Foulds distance (green). A.) The breakdown of relative performance by species. B.) Relative performance averaged across species. Scale is matched to panel A. Note that tree concordance is only included in panel B because it is calculated based upon full sequence alignments.*



Supp. Figure A-7. **The distribution of gene-level conservation (measured by dN/dS) for each component method versus MOSAIC$_{\mathbf{matched}}$.**



Supp. Figure A-8. A representation of the alignments returned by each method for TPSAB1.

```
CCDS10431.1     MLNLLLLALPVLASRAYAAPAPGQALQRVGIVGGQEAPRSKWPWQVSLRVHGPYWMHFCGGSLIHPQWVLTAAHCVGPDV 80
Pan [[multiz]]  MLSLLLLALPILASPAYAAPAPGQALQRAGIVGGQEAPRSKWPWQVSLRVRDRYWMHFCGGSLIHPQWVLTAAHCVGPDF 80
Pon [[inpara]]  MLSLLLLALPVLASPAYAAPAPGQALQRVGIVGGQEAPRSKWPWQVSLRVHGQYWMHFCGGSLIHPQWVLTAAHCVGPDV 80
Mac [[multiz]]  MLNLLLLALPVLVSPAHAAPAPGQALQRVGIVGGQEAPRSKWPWQVSLRLHGQYWMHFCGGSLIHPQWVLTAAHCVGPDV 80
Cal [[OMA]]     MLSLLLLVLPVLVSLAHSAPAPGQALPRAGIVGGQEAPGSRWPWQVSLRFHSQFWMHFCGGSLIHPQWVLTAAHCLGPDV 80
Oto [[inpara]]  MLSLLVLALPILGSRVHAAPAPGQASERAGIVGGQEAPESKWPWQVSLRQHTHFWMHICGGSLIHPQWVLTAAHCVGPEV 80
Bos [[multiz]]  MLHL--LALALLLSLVSAAPAPGQALQRAGIVGGQEAPGSRWPWQVSLRVSHQYWRHHCGGSLIHPQWVLTAAHCVGPEV 78
Equ [[inpara]]  MPNLLVLALALLVNLGHAAPAPGQALEREGIVGGQEASGSKWPWQVSLRKNTEYWKHFCGGSLIHPQWVLTAAHCVGPDI 80

CCDS10431.1     KDLAALRVQLREQHLYYQDQLLPVSRIIVHPQFYTAQIGADIALLELEEPVNVSSHVHTVTLPPASETFPPGMPCWVTGW 160
Pan [[multiz]]  KDLATLRVQLQEQHLYYQDQLLPVSRIIVHPQFYIIQTGADIALLELEEPVNVSSRVHTVTLPPASETFPPGMPCWVTGW 160
Pon [[inpara]]  KDLAALRVQLREQHLYYQDQLLPVGRIIVHPQFYTAQTGADIALLELEEPVNISSHVHTVTLPPASETFPPGMPCWVTGW 160
Mac [[multiz]]  KDLADLRVQLREQHLYYQDQLLPVSRIIVHPQFYAVQIGADIALLELEEPVNVSSHVHTVTLPPASETFPPGTPCWVTGW 160
Cal [[OMA]]     MDLANLRVQLREQHLYYKDRLLPVSRLIVHPQFYIVQTGADIALLELEEPVNVSSHVRTVTLPPASETFPAGTPCWVTGW 160
Oto [[inpara]]  QDLADFRVQLREQHLYYHDKLLPVSRIIPHPGFYMATTGADIALLELEEPVNISHSVHTITLPPASETFPPGTPCWVTGW 160
Bos [[multiz]]  HGPSYFRVQLREQHLYYQDQLLPISRIIPHPNYYSVENGADIALLELDEPVSISCHVQPVTLPPESETFPPGTQCWVTGW 158
Equ [[inpara]]  EDFRDIRVQLREQHLYYRDQLLPVSRILPHPYYYTVENGADIALLELQDPVNISSHVQVVTLPPASETFPPGTPCWVTGW 160
                       ↓                         ↓
CCDS10431.1     GDVDNDERLPPPFPLKQVKVPIMENHICDAKYHLGAYTGDDVRIVRDDMLCAGNTRRDSCQGDSGGPLVCKVNGTWLQAG 240
Pan [[multiz]]  GDVDNDEPLPPPFPLKQVKVPIMENHICDAKYHLGAYTGDDVRIIRDDMLCAGNTRRDSCQGDSGGPLVCKVNGTWLQAG 240
Pon [[inpara]]  GDVDNDEHLPPPFPLKQVKVPIMENHICDAKYHLGLYTGDDVRIVRDDMLCAGNSRRDSCQGDSGGPLVCKVNGTWLQAG 240
Mac [[multiz]]  GDVDNDVPLPPPFPLKQVKVPIMENHICDAKYHSGLYTGDDVRIIRDDMLCAGNSRRDTCQGDSGGPLVCKVNGTWLQAG 240
Cal [[OMA]]     GDVNTGEPLPPPFPLKQVKVPIVENQVCDMKYHAGLYTGDAVHIVRDDMLCAGNSRRDSCQGDSGGPLVCKVNDTWLQAG 240
Oto [[inpara]]  GDVDNDVGLPPPFPLKQVKVPIVENHICDAKYHMGLYTGDNVHIVGDNMLCAGNTRKDSCQGDSGGPLVCKVNGTWLQAG 240
Bos [[multiz]]  GNVDNGRRLPPPFPLKQVKVPVVENSVCDRKYHSGLSTGDNVPIVQEDNLCAGDSGRDSCQGDSGGPLVCKVNGTWLQAG 238
Equ [[inpara]]  GDVDNGVSLPPPFPLKEVKVPIVENSVCDRKYHTGVSTGDNIRIVQADMLCAGNRRHDSCQGDSGGPLVCKVKGTWLQAG 240

CCDS10431.1     VVSWGEGCAQPNRPGIYTRVTYYLDWIHHYVPKKP- 275
Pan [[multiz]]  VVSWDEGCAQPNRPGIYTRVTYYLDWIHHYVPKKHX 276
Pon [[inpara]]  VVSWGEGCAQPNRPGIYTRVTYYLDWIHRYVPKKP- 275
Mac [[multiz]]  VVSWDEGCAQPYRPGIYTRITYYLDWIHRYVPEKPX 276
Cal [[OMA]]     VVSWGEGCALPNRPGIYTRVTYYLDWIHQYVPKKP- 275
Oto [[inpara]]  VVSWGDGCAQPNRPGIYTRVTHYLDWIHHYVPKEP- 275
Bos [[multiz]]  VVSWGDGCAKPNRPGIYTRVTSYLDWIHQYVPQGPX 274
Equ [[inpara]]  VVSWANSCAQPNRPGIYTRVTYYLDWIYQYVPKDS- 275
```

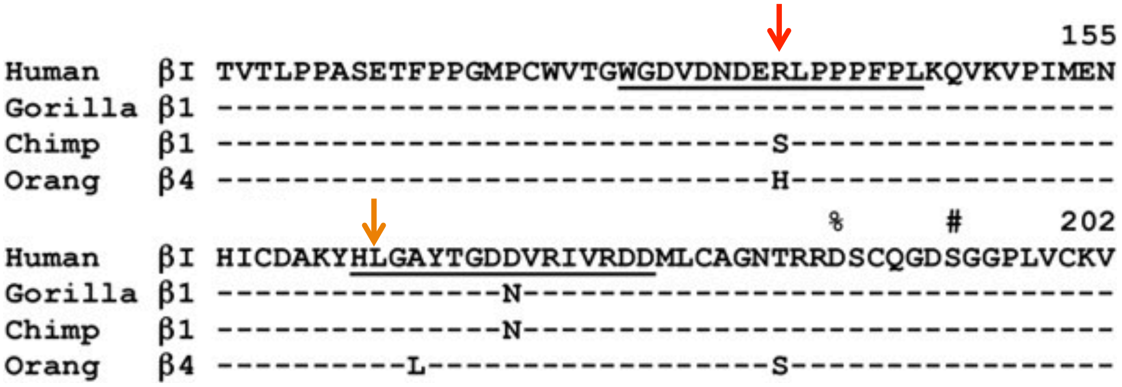
Supp. Figure A-9. **The MOSAIC alignment of TPSAB1.** *The MOSAIC-specific positively selected site is illustrated with the red arrow, while the site detected by several methods, including MOSAIC, is indicated in gold.*

>gi|146150402|gb|ABQ02500.1|:1-275 beta 1 tryptase [Gorilla gorilla]
MLNLLLLALPVLASPAYAAPAPGQALQRAGIVGGQEAPRSKWPWQVSLRVRGQ
YWMHFCGGSLIHPQWVLTAAHCVGPDVKDLAALRVQLREQHLYYQDQLLPVS
RIIVHPQFYTAQIGADIALLELEEPVNVSSHVHTVTLPPASETFPPGMPCWVTGWG
DVDNDE**R**LPPPFPLKQVKVPIMENHICDAKYH**L**GAYTGDNVRIVRDDMLCAGN
TRRDSCQGDSGGPLVCKVNGTWLQAGVVSWGEGCAQPNRPGIYTRVTYYLDWI
HHYVPKKP

Supp. Figure A-10. **The *Gorilla gorilla* sequence that is orthologous to TPSAB1.** *A* Gorilla gorilla gorilla *sequence was not present, presumably due to genome quality issues. For the* Gorilla gorilla *sequence, we highlight the residues of the positively selected sites indicated in Figure SA-9.*

Query species	Best match	% ID	% Similarity	Alignment length	Mismatches	E-value
Chimp	TPSAB1	94	95	262	15	0
Orangutan	TPSAB1	96	97	275	10	0
Rhesus Mac.	TPSAB1	92	95	263	21	2.0E-180
Marmoset	TPSAB1	85	90	262	39	3.0E-166
Bushbaby	TPSAB1	84	90	263	41	5.0E-167
Cow	TPSAB1	77	86	262	60	1.0E-148
Horse	TPSAB1	79	87	258	54	2.0E-153

Supp. Table A-1. SwissProt database BLAST results for each of the putative orthologs of TPSAB1.



```
                                                         155
Human    βI TVTLPPASETFPPGMPCWVTGWGDVDNDERLPPPFPLKQVKVPIMEN
Gorilla  β1 -----------------------------------------------
Chimp    β1 -----------------------------S-----------------
Orang    β4 -----------------------------H-----------------

                                         %     #         202
Human    βI HICDAKYHLGAYTGDDVRIVRDDMLCAGNTRRDSCQGDSGGPLVCKV
Gorilla  β1 ---------------N-------------------------------
Chimp    β1 ---------------N-------------------------------
Orang    β4 ----------L------------------S-----------------
```

Supp. Figure A-11. **Manually derived alignments of TPSAB1, reproduced from Trivedi et al. 2007.** *As above, The MOSAIC-specific positively selected site is illustrated with the red arrow, while the site detected by several methods, including MOSAIC, is indicated in gold.*

Appendix B Supplemental Material to Chapter Three

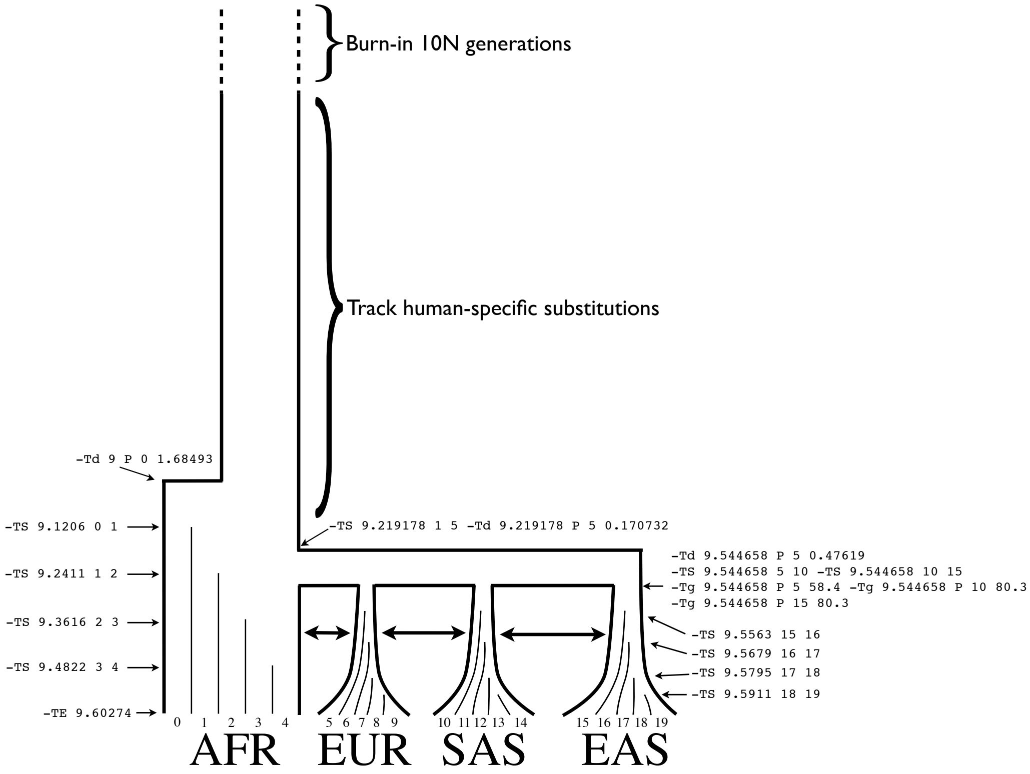
SnIPRE is robust to demography and models of selection

To test the robustness of SnIPRE to demography and models of selection, we used SFS_CODE (Hernandez, 2008) with the command line arguments presented below. These options are further illustrated in Fig. SB-1. In brief, we simulated divergent and polymorphic mutations using 6470 genes, a 20 population demographic model, and varying probabilities of selection ranging from 0 to 100% per site. Population-scaled selection coefficients for beneficial alleles ranged from slightly above 0 to approximately 30. Deleterious mutations were also modeled. SnIPRE analysis was then performed in the same way described for the real data.

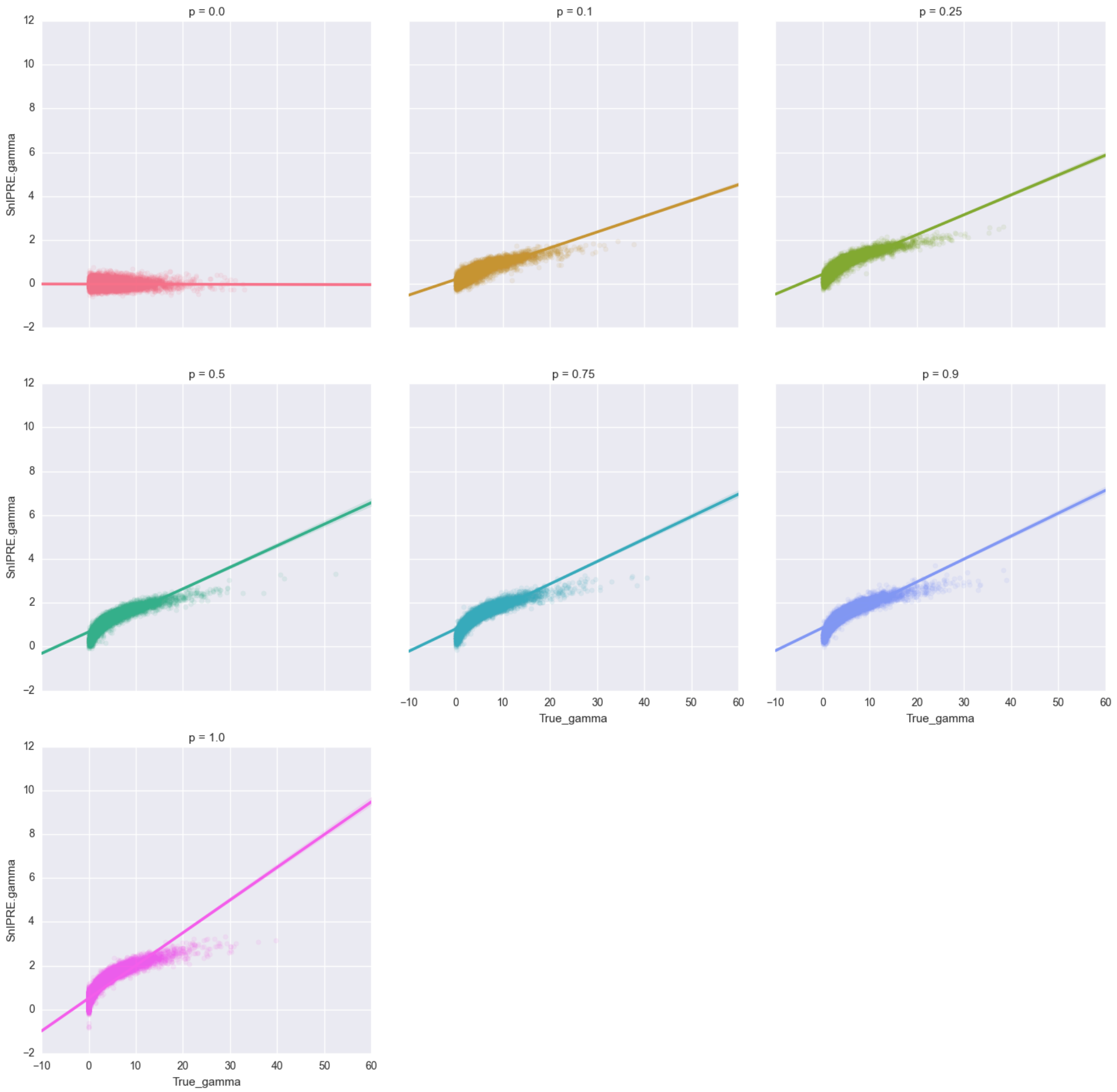
We show in Fig. SB-2 that SnIPRE demonstrates excellent rank correlation with true selection coefficients even in the face of demography, deleterious mutation, and heterogeneous selection across the gene. It should be noted however that the scale of the estimated selection coefficient is not accurate. SnIPRE estimates can then tell the researcher how selection compares across genes and whether there is positive selection within a particular gene, but it does not provide an accurate estimate of the scale of gamma.

```
./sfs_code 20 1 -TS 9.1206 0 1 -TS 9.2411 1 2 -TS 9.3616 2 3 -TS 9.4822 3 4
-TS 9.219178 1 5 -TS 9.544658 5 10 -TS 9.544658 10 15 -TS 9.5563 5 6 -TS
9.5679 6 7 -TS 9.5795 7 8 -TS 9.5911 8 9 -TS 9.5563 10 11 -TS 9.5679 11 12
-TS 9.5795 12 13 -TS 9.5911 13 14 -TS 9.5563 15 16 -TS 9.5679 16 17 -TS
9.5795 17 18 -TS 9.5911 18 19 -TE 0.60274 -Td 0 P 0 1.68493 -Td 9.219178 P
5 0.170732 -Td 9.544658 P 5 0.47619 -Tg 9.544658 P 5 58.4 -Tg 9.544658 P 6
58.4 -Tg 9.544658 P 7 58.4 -Tg 9.544658 P 8 58.4 -Tg 9.544658 P 9 58.4 -Tg
9.544658 P 10 80.3 -Tg 9.544658 P 11 80.3 -Tg 9.544658 P 12 80.3 -Tg
9.544658 P 13 80.3 -Tg 9.544658 P 14 80.3 -Tg 9.544658 P 15 80.3 -Tg
```

```
9.544658 P 16 80.3 -Tg 9.544658 P 17 80.3 -Tg 9.544658 P 18 80.3 -Tg
9.544658 P 19 80.3 -m L 1 1 1 1 0.738 0.738 0.738 0.738 0.738 0.4674 0.4674
0.4674 0.4674 0.4674 0.4674 0.4674 0.4674 0.4674 0.4674 1 1 1 1 0.738 0.738
0.738 0.738 0.738 0.4674 0.4674 0.4674 0.4674 0.4674 0.4674 0.4674 0.4674
0.4674 0.4674 1 1 1 1 0.738 0.738 0.738 0.738 0.738 0.4674 0.4674 0.4674
0.4674 0.4674 0.4674 0.4674 0.4674 0.4674 0.4674 1 1 1 1 0.738 0.738 0.738
0.738 0.738 0.4674 0.4674 0.4674 0.4674 0.4674 0.4674 0.4674 0.4674 0.4674
0.4674 1 1 1 1 0.738 0.738 0.738 0.738 0.738 0.4674 0.4674 0.4674 0.4674
0.4674 0.4674 0.4674 0.4674 0.4674 0.4674 0.06 0.06 0.06 0.06 0.06 1 1 1 1
0.192 0.192 0.192 0.192 0.192 0.192 0.192 0.192 0.192 0.192 0.06 0.06 0.06
0.06 0.06 1 1 1 1 0.192 0.192 0.192 0.192 0.192 0.192 0.192 0.192 0.192
0.192 0.06 0.06 0.06 0.06 0.06 1 1 1 1 0.192 0.192 0.192 0.192 0.192 0.192
0.192 0.192 0.192 0.192 0.06 0.06 0.06 0.06 0.06 1 1 1 1 0.192 0.192 0.192
0.192 0.192 0.192 0.192 0.192 0.192 0.192 0.06 0.06 0.06 0.06 0.06 1 1 1 1
0.192 0.192 0.192 0.192 0.192 0.192 0.192 0.192 0.192 0.192 0.01938 0.01938
0.01938 0.01938 0.01938 0.09792 0.09792 0.09792 0.09792 0.09792 1 1 1 1
0.192 0.192 0.192 0.192 0.192 0.01938 0.01938 0.01938 0.01938 0.01938
0.09792 0.09792 0.09792 0.09792 0.09792 1 1 1 1 0.192 0.192 0.192 0.192
0.192 0.01938 0.01938 0.01938 0.01938 0.01938 0.09792 0.09792 0.09792
0.09792 0.09792 1 1 1 1 0.192 0.192 0.192 0.192 0.192 0.01938 0.01938
0.01938 0.01938 0.01938 0.09792 0.09792 0.09792 0.09792 0.09792 1 1 1 1
0.192 0.192 0.192 0.192 0.192 0.01938 0.01938 0.01938 0.01938 0.01938
0.09792 0.09792 0.09792 0.09792 0.09792 1 1 1 1 0.192 0.192 0.192 0.192
0.192 0.01938 0.01938 0.01938 0.01938 0.01938 0.09792 0.09792 0.09792
0.09792 0.09792 0.192 0.192 0.192 0.192 0.192 1 1 1 1 0.01938 0.01938
0.01938 0.01938 0.01938 0.09792 0.09792 0.09792 0.09792 0.09792 0.192 0.192
0.192 0.192 0.192 1 1 1 1 0.01938 0.01938 0.01938 0.01938 0.01938 0.09792
0.09792 0.09792 0.09792 0.09792 0.192 0.192 0.192 0.192 0.192 1 1 1 1
0.01938 0.01938 0.01938 0.01938 0.01938 0.09792 0.09792 0.09792 0.09792
0.09792 0.192 0.192 0.192 0.192 0.192 1 1 1 1 0.01938 0.01938 0.01938
0.01938 0.01938 0.09792 0.09792 0.09792 0.09792 0.09792 0.192 0.192 0.192
0.192 0.192 1 1 1 1 --printGen -N 7300 -n 125
```



Supp. Figure B-1. A schematic of the SFS_code simulation scheme.



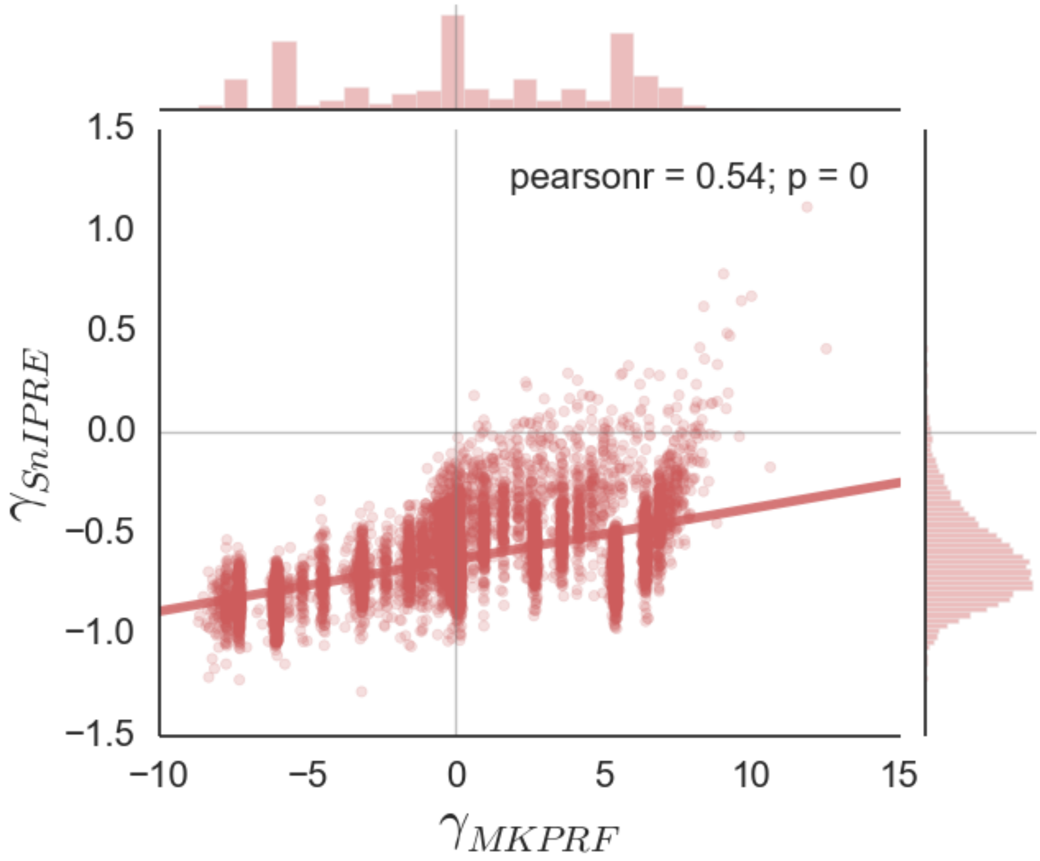
Supp. Figure B-2. Correlation between observed and inferred levels of selection for data simulated with demography, deleterious mutations, and heterogeneous selection.

Comparison to previous results

In 2005, Bustamante *et al*. calculated McDonald-Kreitman scores across the human genome using the McDonald-Kreitman Poisson Random Field model. This is a highly parametric statistical model derived from population genetic approximations (Williamson, Fledel-Alon, & Bustamante, 2004). The input data for this model was human and chimp exome sequences collected using an exome pulldown and sequencing approach with oligonucleotides designed for the human genome (Bustamante et al., 2005b). Human polymorphisms were calculated without frequency filtering. This is expected to lead to overly conservative estimates of selection due to the inclusion of deleterious polymorphism held at low frequency by natural selection. In addition, pulling down chimp sequences using human-directed oligonucleotides might be expected to introduce a slight bias towards more conserved genes. This is less of a factor due to the high levels of sequence identity between human and chimp.

On the same data, SnIPRE predicts higher levels of conservation than MKPRF

As you can see in Figure SB-1, MKPRF predicts nearly a third of the genome to be under positive selection, approximately a third to be neutrally evolving, and the remaining third to be conserved. Given that the model is bayesian in nature, this likely points to an exaggerated influence of a uniform prior on selection class. Calculating selection using SnIPRE applied to the same data yield a strikingly different picture however. In this case, fewer than .5% of genes in the human genome are estimated to be under positive selection. The pearson correlation between the two result sets is 0.54.



Supp. Figure B-3. **A comparison of selection coefficents calculated using the same data using SnIPRE and MKPRF, respectively.** *Marginal distributions are plotted above and to th right. Neutrality (gamma=0) is indicated with a grey line.*

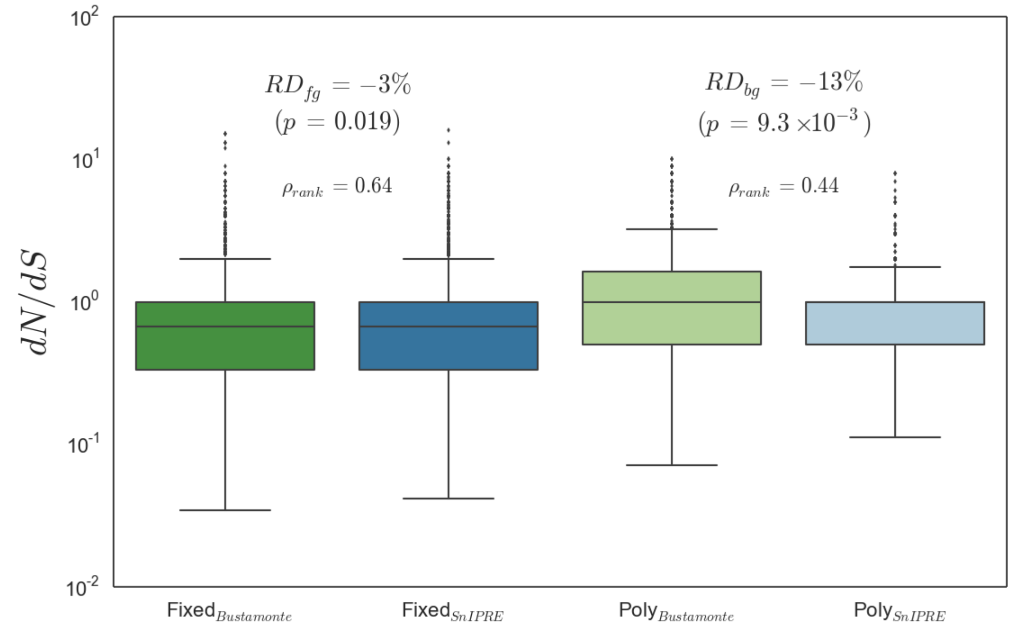
Updated input data addresses overly conservative neutrality estimate

The lack of positive selection seen using previously published data may stem mainly from an overly conservative null distribution. A classic test for selection is to compare non-synonymous and synonymous mutation rates (dN and dS, respectively). The assumption is that mutations that do not affect protein coding sequence will be evolutionarily neutral. Positive selection can then be declared is the non-synonymous mutation rate significantly outstrips that of synonymous mutations.

In the case of the McDonald-Kreitman test, the dN/dS ratio in polymorphism data is taken as the measure of neutral evolution within a particular gene, and the dN/dS ratio in divergence data is taken as the foreground for statistical comparison. This approach has the strength of making no assumptions about the neutrality of synonymous mutations—a common assumption that is known to be violated by the presence of transcription factor binding sites, RNA secondary structure, etc. This approach makes its own assumption however.

Specifically, it assumes that polymorphic mutations are evolutionary neutral, a presumption that falls apart for low frequency variants. The reason for this is two-fold. First, low frequency variants tend to be new, so evolution has not had the opportunity to purge deleterious variants. Second, older deleterious variants may have been driven to, or maintained at low frequency due to evolutionary selection.

In the data published by Bustamante *et al.,* all observed polymorphisms are included, regardless of frequency. We would expect this to inflate the dN/dS ratio in polymorphism relative to filtered data. In Figure SB-2 we see that this is indeed the case (p=9.3e-3). In addition to this observation, we also see that there is not a remarkably strong rank correlation between dN/dS ratios in the two sets of divergence data. Note that for a more equitable comparison, we did not perform ancestral sequence reconstruction in this case, but instead directly counted differences between modern-day sequences.

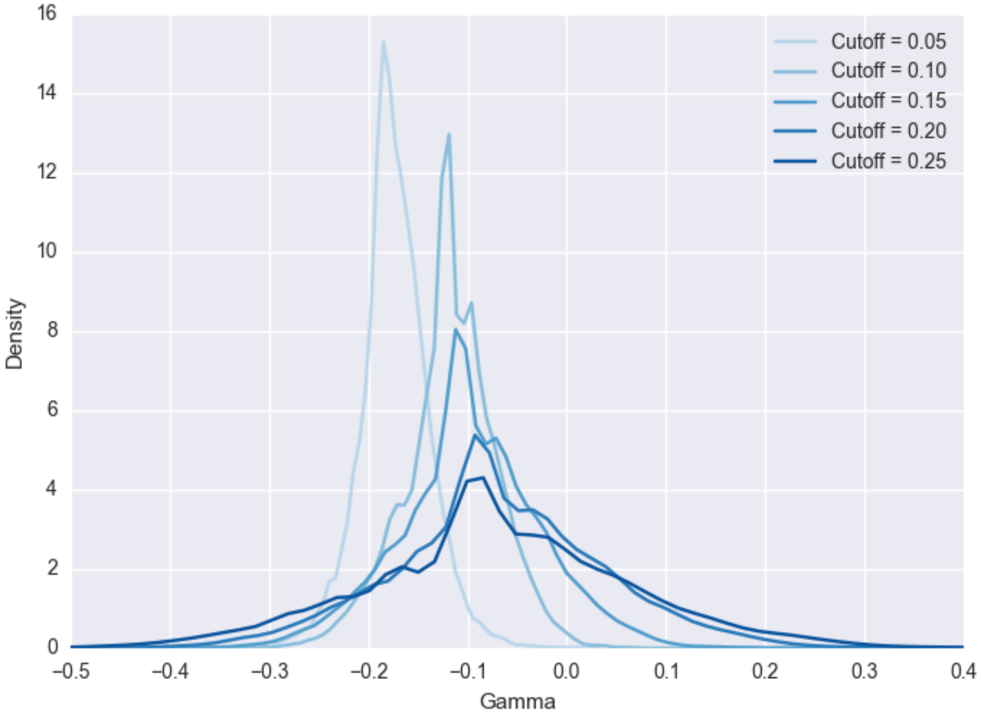


Supp. Figure B-4. **A comparison of dN/dS values between previously published (green) and newly calculated (blue) values.** *Distributions are separated according to fixed (dark) and polymorphic (light) mutations. For each comparison between methods, we show the rank correlation, the relative difference between means (RD), and the p-value of this comparison.*

Influence of allele frequency cutoff on estimated selection coefficient

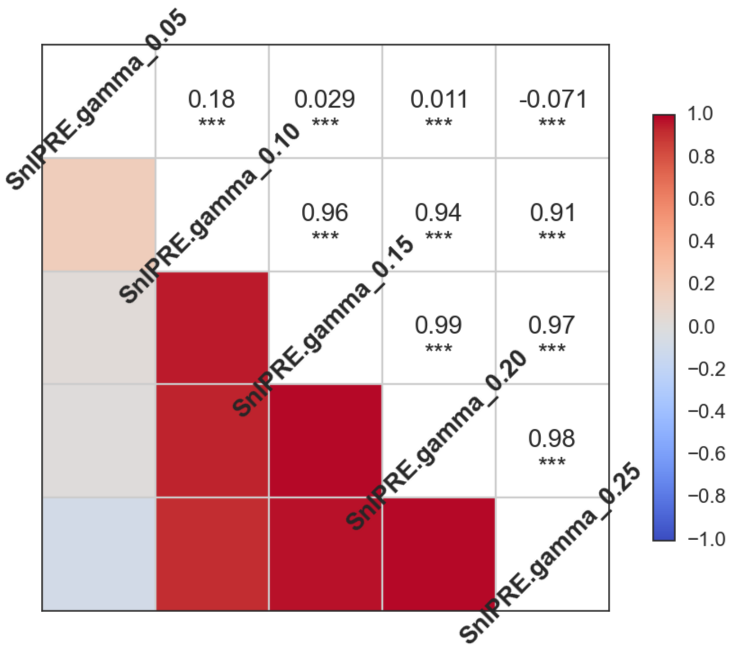
Consistent with theory and previous results (Messer & Petrov, 2013), we have shown that an allele frequency cutoff is important for removing excess deleterious, predominantly non-synonymous mutations from polymorphism data. We next sought to assess the sensitivity of downstream results to choice of cutoff.

In Figure SB-3 we show the influence of allele frequency cutoff on the overall distribution of gamma values. We see that at a 5% allele frequency, almost no genes are found to be positively selected. As this cutoff is raised, the both the mean and the variance of the gamma distribution increase. This is because we are removing a conservative bias in the null distribution, but we are also filtering out non-deleterious alleles and thus increasing the variability of our estimator. We believe that a cutoff of 15% strikes a good balance between bias and variance. However, since the genes that are considered positively selected are still sensitive to this cutoff, we calculate GO enrichment using a cutoff-free rank-based approach.

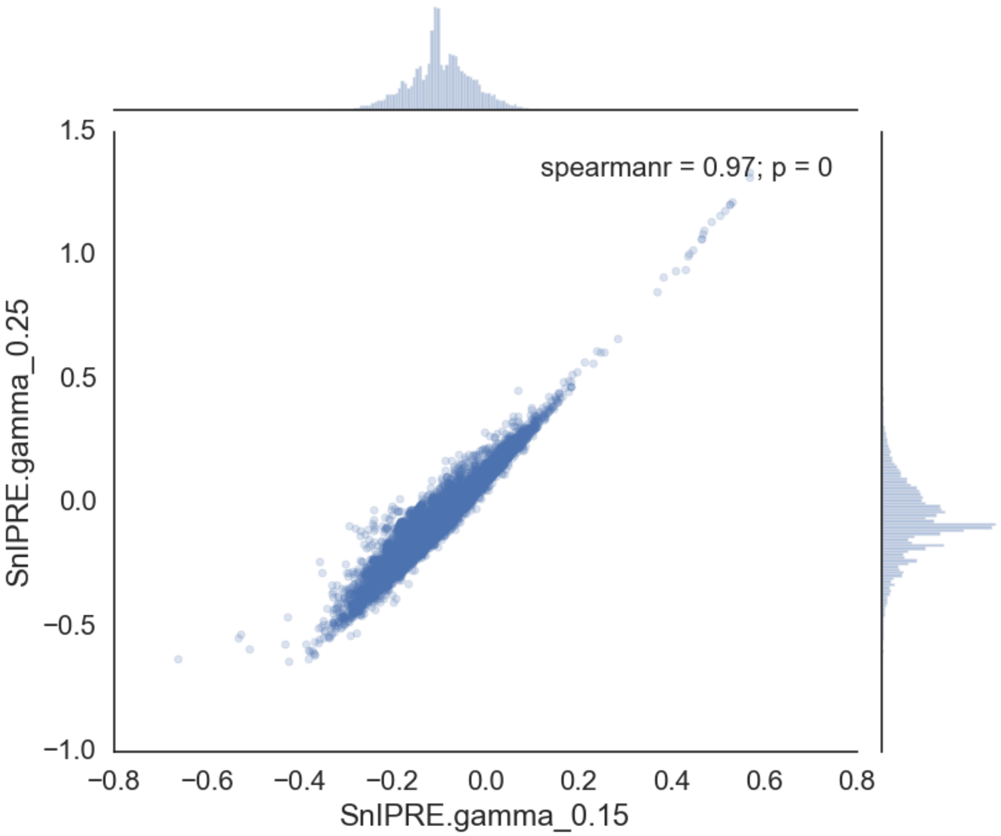


Supp. Figure B-5. **The distribution of SnIPRE-derived selection coefficients as a function of the allele frequency cutoff used as a criterion for including polymorphic mutations.**

Despite the changes to mean and variance in the distributions in (Figure SB-3), we find that, above an allele frequency cutoff of 5%, correlations between estimates range from 0.92 to 0.99 (Figure SB-4). This further convinces us that our results are unlikely to be unduly effected by this choice of cutoffs. A joint distribution plot of one of these pairwise comparisons is shown in Figure SB-5.



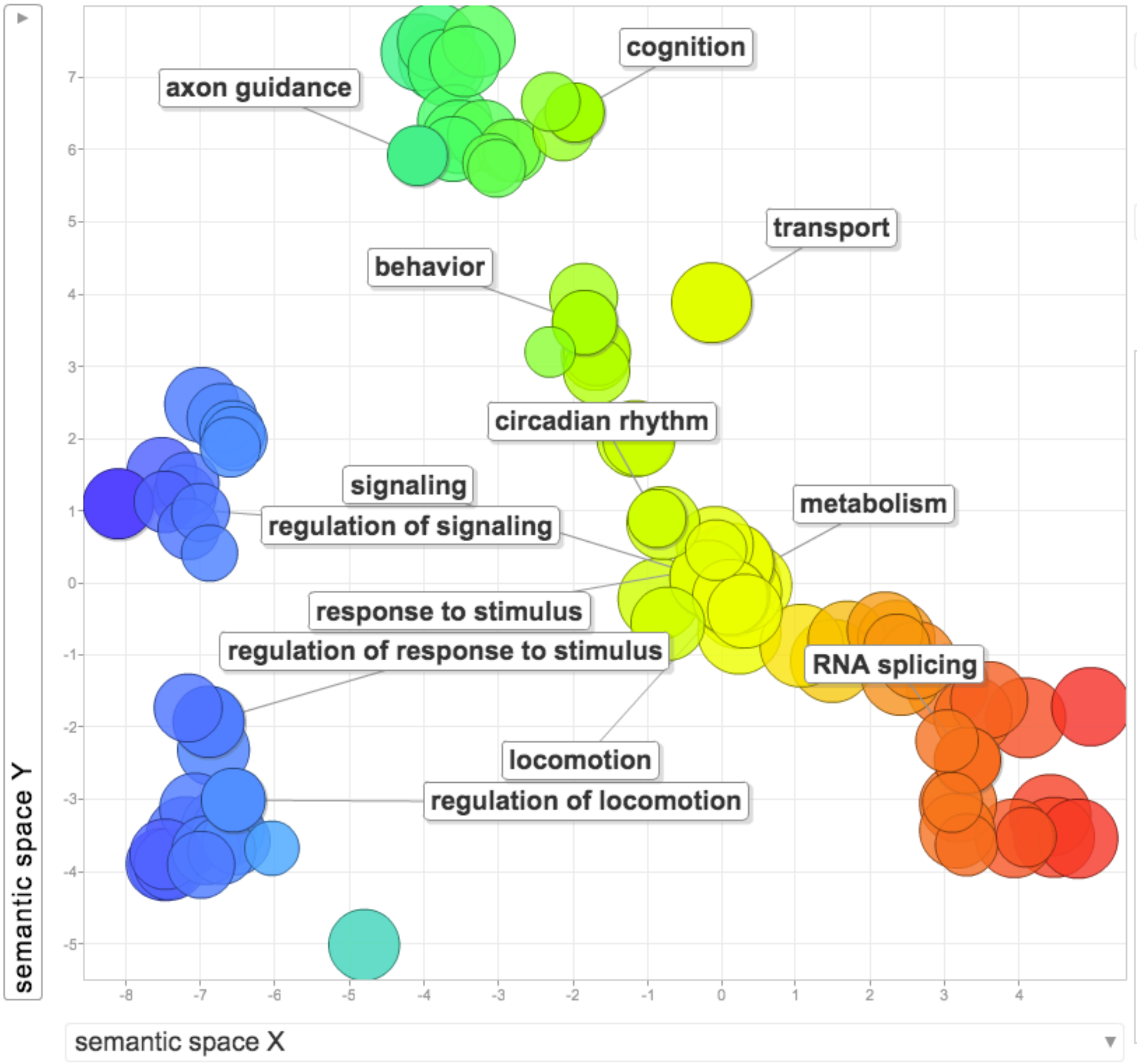
Supp. Figure B-6. A heatmap of rank correlations between selection coefficients calculated using varying allele frequency cutoffs.



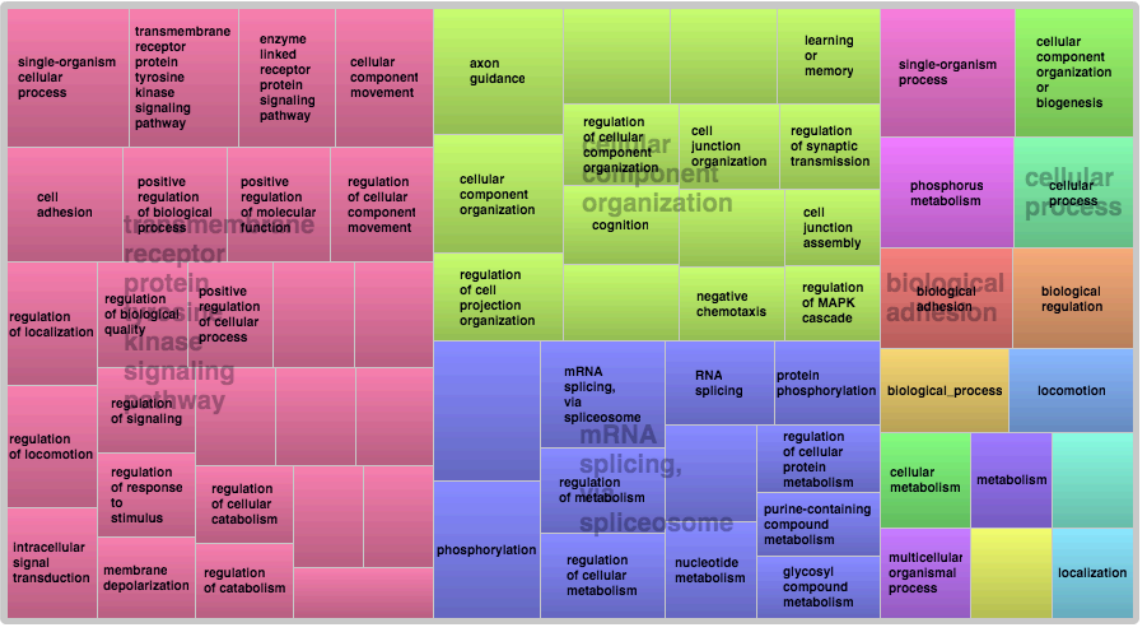
Supp. Figure B-7. **A plot of the joint distributions of SnIPRE-inferred gamma values using allele frequency cutoffs of 15% and 25%.**

Additional GO enrichment plots

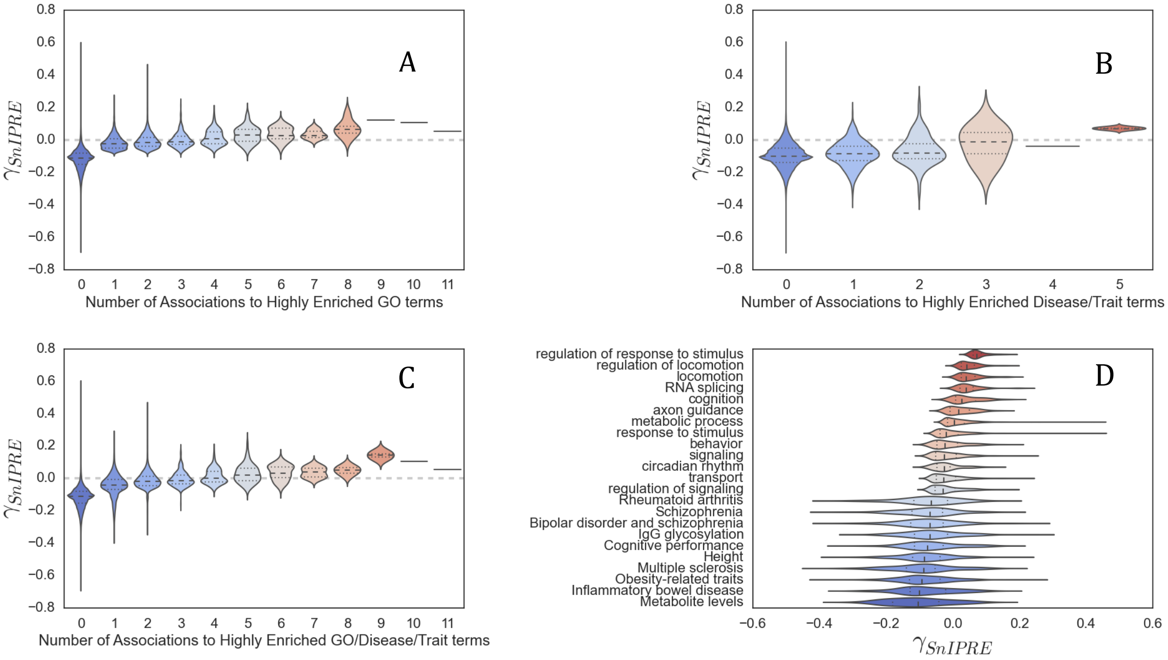
Since GO terms are organized in a hierarchical fashion, it is sometimes a challenge to understand at a glance how biologically distinct various GO categories may be. To assess the diversity of our hits at a glance, we used ReviGO (Supek, Bošnjak, Škunca, & Šmuc, 2011) to plot enriched GO terms by two axes of semantic similarity. These semantic similarity scores were derived using multi-dimensional scaling (MDS), a dimensionality reduction method conceptually similar to principal components analysis (PCA) (Figure SB-6). For convenience, we also summarized these terms using hierarchical clustering (Figure SB-7).



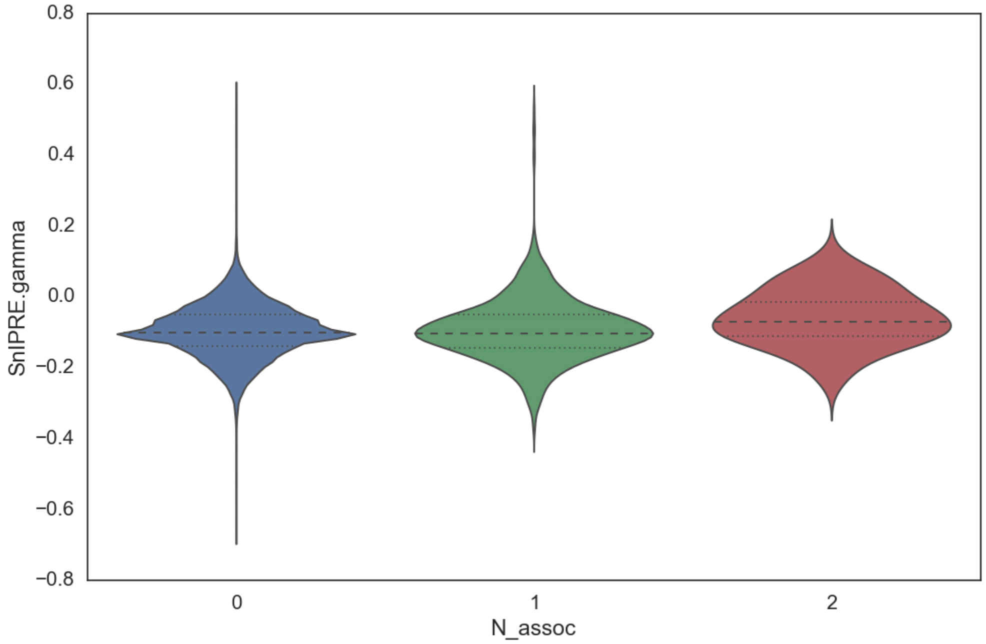
Supp. Figure B-8. **Enriched GO processes visualized using MDS-derived semantic similarity distances.**


single-organism cellular process
transmembrane receptor protein tyrosine kinase signaling pathway
enzyme linked receptor protein signaling pathway
cellular component movement
axon guidance
learning or memory
single-organism process
cellular component organization or biogenesis
cell adhesion
positive regulation of biological process
positive regulation of molecular function
regulation of cellular component movement
regulation of cellular component organization
cell junction organization
regulation of synaptic transmission
cellular component organization
cognition
cell junction assembly
phosphorus metabolism
cellular process
regulation of localization
regulation of biological quality
positive regulation of cellular process
regulation of cell projection organization
negative chemotaxis
regulation of MAPK cascade
biological adhesion
biological regulation
regulation of locomotion
regulation of signaling
mRNA splicing, via spliceosome
RNA splicing
protein phosphorylation
biological_process
locomotion
regulation of response to stimulus
regulation of cellular catabolism
regulation of metabolism
regulation of cellular protein metabolism
purine-containing compound metabolism
cellular metabolism
metabolism
intracellular signal transduction
membrane depolarization
regulation of catabolism
phosphorylation
regulation of cellular metabolism
nucleotide metabolism
glycosyl compound metabolism
multicellular organismal process
localization

Supp. Figure B-9. **Enriched GO categories clustered hierarchically.**



Supp. Figure B-10. **Distributions of postive selection scores within various groups of genes.** *A.) Positive selection distributions as a function of the number of enriched GO terms associated with each gene. B.) Positive selection as a function of the number of enriched disease/trait associations linked to particular genes. C.) Pool results pooling the categories from A and B. D.) Distribution of positive selection scores within the most enriched GO categories and disease/phenotype groups.*



Supp. Figure B-11. **For permuted selection scores, the distribution of positive selection scores as a function of the number of categories associated with a given gene.**

Appendix C Supplemental Materials to Chapter Four

Convergent cross mapping algorithm

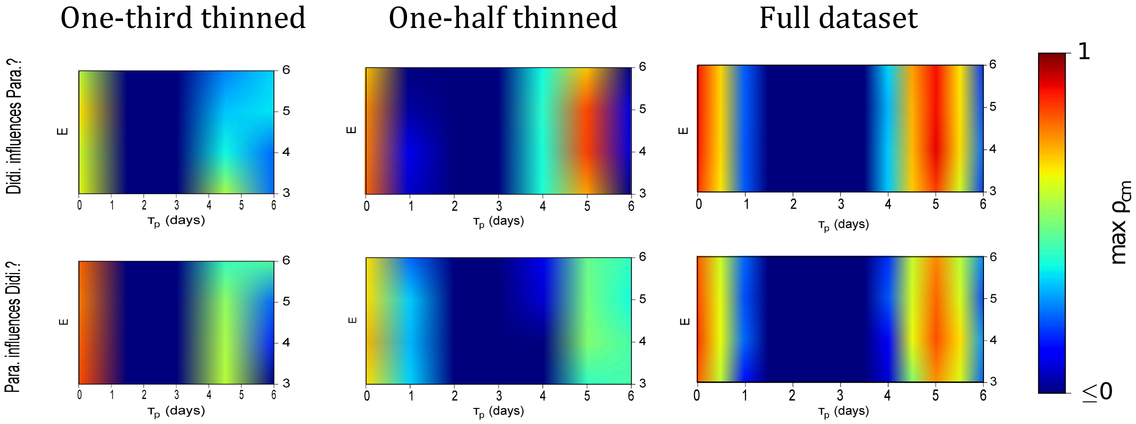
Consider time series of hypothetical variables *X* and *Y*. Convergent cross-mapping (CCM) employs state space reconstruction (SSR), thereby using time-lagged coordinates of each of these variables to produce shadow versions of their respective source manifolds. We will refer to these projection manifolds as M_x and M_y. To test whether *X* causes *Y*, CCM applies the following logic: Because manifold reconstruction preserves the Lyapunov exponents of the original system (Casdagli, Eubank, Farmer, & Gibson, 1991), if *X* causes *Y*, then time points that are close in M_y should also be close in M_x. Since M_x is constructed from lags of the observations of *X*, the points that are close in M_x will also have similar values in the corresponding time series. Therefore, if *X* causes *Y*, then M_y can tell us which observations of *X* should best predict a given point from *X*. Furthermore, predictability should increase with the number of manifold points that are considered.

To test whether *X* causes *Y*, M_y is used to infer the points in *X* that will best predict a given held-out point from X. We measure this performance using predictive skill, quantified by ρ_{ccm}. Intuitively, this procedure works as follows: A point is held out from *X*. We then use M_y to infer the points in M_x that will be closest to that point of interest. Using exponential weights derived from the relative pairwise distances of corresponding points in M_y, we predict the held-out point using other observations from *X*. Finally, ρ_{ccm} is calculated as the Pearson correlation between observed and predicted points, and so is a cross-validated measure. To examine whether the signal converges as expected for a causal relationship, these steps are repeated using increasing time series length (*L*).

Paramecium-Didinium system

Didinium is a free-living unicellular carnivore. Paramecium is its prey. More information about this system, as well as interactive graphs of time series and manifold constructions, can be found at:

http://cyrusmaher.github.io/CauseMap.jl/ParaDidiExample.html#paramecium-and-didinium



Supp. Figure C-1. **The maximal predictive skill as a function of E, tau p, and the number of included points.**

Fourier transform analysis

We calculated the characteristic frequencies of the paramecium and didinium time series by performing fourier transform analysis using the rfft function in the python module scipy.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature	10/31/14 Date