Alignment Uncertainty and Genomic Analysis

Karen M. Wong,[1] Marc A. Suchard,[2] John P. Huelsenbeck[3]*

The statistical methods applied to the analysis of genomic data do not account for uncertainty in the sequence alignment. Indeed, the alignment is treated as an observation, and all of the subsequent inferences depend on the alignment being correct. This may not have been too problematic for many phylogenetic studies, in which the gene is carefully chosen for, among other things, ease of alignment. However, in a comparative genomics study, the same statistical methods are applied repeatedly on thousands of genes, many of which will be difficult to align. Using genomic data from seven yeast species, we show that uncertainty in the alignment can lead to several problems, including different alignment methods resulting in different conclusions.

A common theme in comparative genomics studies is a flow diagram, or chart, tracing the various steps and algorithms used during the analysis of a large number of genes. Flow charts can be quite sophisticated, with steps such as identifying orthologous gene sets, aligning the genes, and performing different statistical analyses on the resulting alignments. The key point, and a great practical difficulty in comparative genomics studies, is that the analyses must be repeated many times. The procedure, then, is largely automated, with scripting languages such as Perl or Python cobbling together individual programs that perform each step. In addition, many of the individual steps involve procedures originally developed in the evolutionary biology literature, to perform phylogeny estimation or to identify individual amino acid residues under the influence of positive selection (*1*). Statistical methods that until recently would have been applied to a single alignment, carefully constructed, are now applied to a large number of alignments, many of which may be of uncertain quality and cause the underlying assumptions of the methods to fail.

How might alignment uncertainty affect genomic studies? We performed a study designed to uncover the effect that alignment has on inferences of evolutionary parameters. We examined genomic data from seven yeast species (*Saccharomyces cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. bayanus*, *S. castellii*, and *S. kluyveri*). Earlier molecular evolution studies that included these species established the appropriateness of sequence comparisons between them (*2*–*4*), with estimated divergence dates from *S. cerevisiae* ranging from as little as 5 million years for *S. paradoxus* to about 100 million years for *S. kluyveri* and average pairwise sequence similarity ranging from 54 to 89%. The comparisons we carried out among the seven yeast species are, thus, reasonable and of the sort that any evolutionary biologist might make. Accurate inference of evolutionary processes from molecular sequences also relies on the compared sequences being orthologous. However, correct identification of orthologous sequences is not trivial because current alignment algorithms do not evaluate homology and will align sequences regardless of proper evolutionary relationships. We combined two earlier data sets of previously identified orthologous open reading frames (ORFs) from studies on the comparative genomics analysis of yeast (*3*, *4*). The orthologs identified from the Kellis *et al*. (*4*) study were used for species that overlapped between the two studies (*S. mikatae* and *S. bayanus*), and only those ORFs for which all seven species contained a detected orthologous sequence were included in the analysis. Overall, we considered a total of 1502 sets of orthologous gene sequences.

For each orthologous gene set, we applied seven different alignment programs—Clustal W, Muscle, T-Coffee, Dialign 2, Mafft, Dca, and ProbCons (*5*–*11*)—aligning data by amino acid sequence under default program settings and using the aligned amino acid sequences to construct nucleotide alignments. From this intensive undertaking, we produced a table of 1502 × 7 alignments. Alignments were then subjected to several statistical analyses of the sort that an evolutionary biologist might apply; specifically, we estimated the phylogeny using maximum likelihood under the GTR+Γ model of DNA substitution and the number of positively selected sites for each alignment (*1*).

Estimates of phylogeny and inferences of positive selection were sensitive to alignment treatment. Confirming previous studies showing that alignment method has a considerable effect on tree topology (*12*–*14*), we found that 46.2% of the 1502 ORFs had one or more differing trees depending on the alignment procedure used. The number of unique trees outputted for each ORF varied from one to six, and the average symmetric-difference distance (*15*) between trees for each ORF ranged from 0 to 6.67 (for trees of seven species, the maximum possible value is eight). Figure 1 shows a case in which alignments produced by the seven different alignment programs resulted in six different estimates of phylogeny. In general, phylogenies estimated from different alignments for an ORF were more concordant when the alignments were similar. Figure 2A shows a strong positive relation between a measure of variability in alignments across alignment treatments and the average topological distance between estimated trees (*15*). The support for the maximum-likelihood trees, measured by the nonparametric bootstrap, was generally lower when alignments were dissimilar across treatments (Fig. 2B). One does not usually find strongly supported, but conflicting, phylogenies produced by different alignment treatments.

Previous studies on the effects produced by different alignment methods focused on tree topology. Yet, other commonly estimated evolutionary parameters, such as substitution rates and the frequency of positively selected sites, are also alignment dependent. To examine if variable alignments for an ORF affect the inference of these parameters, we estimated the synonymous (d_S) and nonsynonymous (d_N) substitution rates for each gene and inferred sites under positive selection using Paml, under the M2 model with (initially) a threshold of 0.5 for inferring a site to be under positive selection (*1*). Overall estimates of substitution rates did not differ significantly among alignment treatments (Kruskal-Wallis test: d_N, $P = 0.59$; d_S, $P = 0.08$; d_N/d_S, $P = 0.51$), and for most ORFs none of the sites were inferred as under positive selection, regardless of the alignment treatment (1032 ORFs). However, of the remaining 470 ORFs, only 44 showed a consistent number of positively selected sites. Thus, in 28.4% of the cases, we found that the inference of positively selected sites was also sensitive to the method of alignment. Raising the threshold for flagging sites as under the influence of positive natural selection to 0.95 reduced the number of conflicting ORFs (Fig. 3); in 14.8% of the cases, positive-selection inference was sensitive to alignment treatment. However, reducing conflict among alignment treatments comes at the cost of finding fewer sites under positive selection, and in many cases alignment treatments still produce discordant inferences of positive selection.

We hypothesize that the inconsistent inferences of alignments produced by the seven different alignment methods examined here is not necessarily a fault of the alignment procedures, but rather reflects underlying variability in the processes of substitution, insertion, and deletion that makes some ORFs inherently more difficult to align. We examined alignment variability by approximating the marginal posterior probability distribution of the alignment for each ORF, using the program BAli-Phy (*16*, *17*). BAli-Phy implements a stochastic model of insertion and deletion and explores posterior probability distributions of phylogenetic model parameters, such as the tree and branch lengths, as well as the

[1]Section of Ecology, Behavior and Evolution, University of California, San Diego, La Jolla, CA 92093, USA. [2]Department of Biomathematics, University of California, Los Angeles, Los Angeles, CA 90095, USA. [3]Department of Integrative Biology, University of California, Berkeley, Berkeley, CA 94720, USA.

*To whom correspondence should be addressed. E-mail: johnh@berkeley.edu

probability distribution of alignment by Markov chain Monte Carlo (MCMC). Quantifying the uncertainty of complex discrete random variables, such as alignments, is a formidable task. We developed a crude summary statistic that reflects variability of the alignments sampled with MCMC for each ORF; we calculated a distance between all pairs of sampled alignments and considered the mean of these pairwise distances as a measure of inherent alignment uncertainty for each ORF. To measure distances between alignments, we exploited the metric of Schwartz *et al*. (*18*). Effectively, this metric counts the number of pairwise homology statements upon which two alignments disagree. We found that alignment variability, as reflected by the marginal posterior probability distribution of alignments, was associated with the inconsistency of alignments produced by the seven different alignment methods (Fig. 2C) and with the number of estimated nonsynonymous substitutions for an ORF (Fig. 2D).

The problem of alignment uncertainty in genomic studies, identified here, is not a problem of sloppy analysis. Many comparative genomics studies are carefully performed and reasonable in design. However, even carefully designed and carried out analyses can suffer from these types of problems because the methods used in the analysis of the genomic data do not properly accommodate alignment uncertainty in the first place. Moreover, the genes that are of greatest interest to the evolutionary biologist probably suffer disproportionately. For example, in several studies, the genes of greatest interest were the ones that had diverged most in their nonsynonymous rate of substitution (*19*). But, these are the very genes that should be the most difficult to align in the first place. We also do not believe that the alignment uncertainty problem is one that can be resolved by simply throwing away genes, or portions of genes, for which alignment differs. Quality checks are common in comparative genomics studies, often referred to as "filters" in a flow diagram showing the analyses that were performed. The filters usually exclude

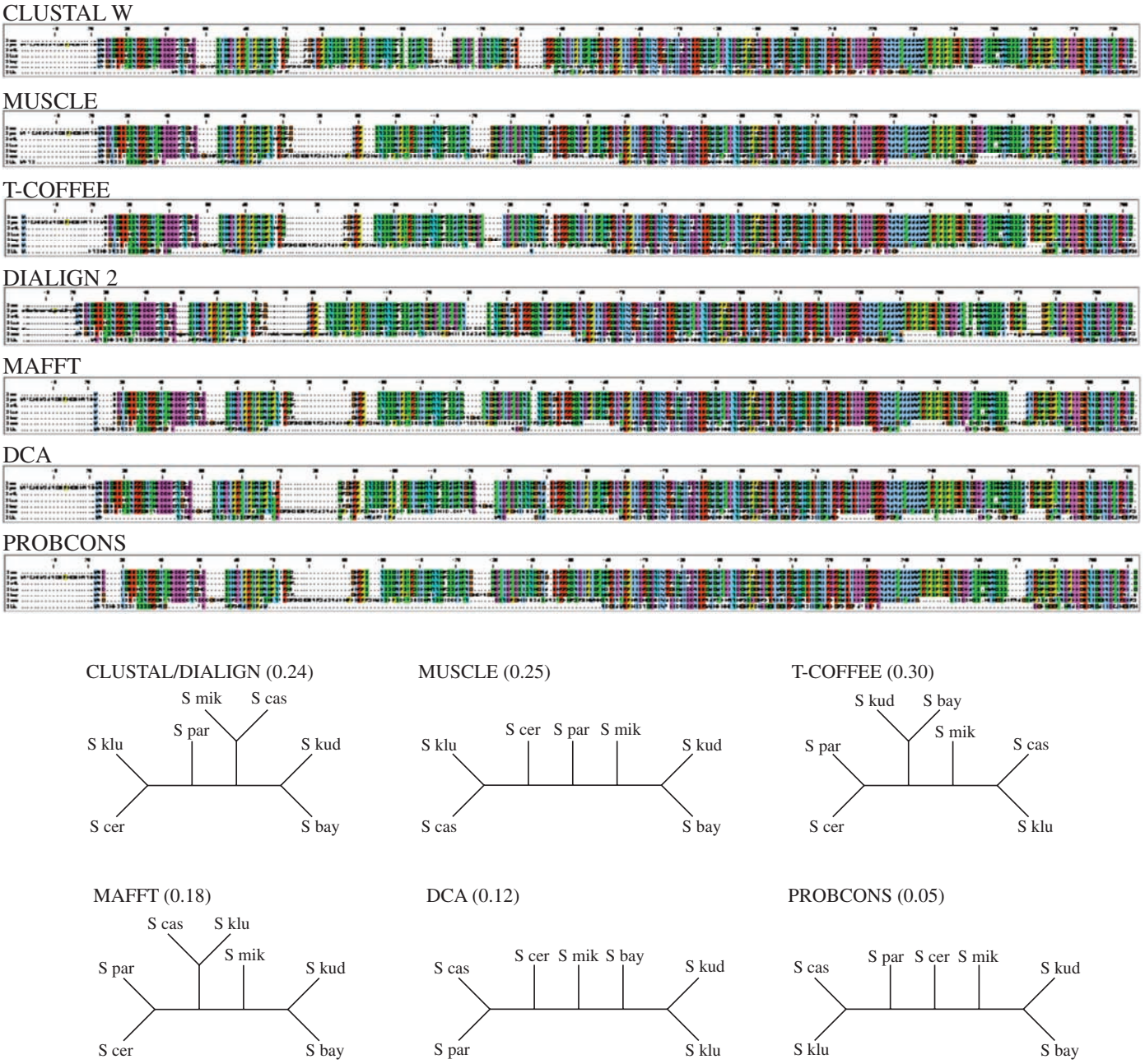

Fig. 1. An example, involving ORF YPL077C, in which alignments produced by seven different alignment methods produce six different estimated trees, albeit with low bootstrap support (bootstrap proportions shown parenthetically for each tree).

ambiguous alignment regions according to some criterion. Discarding information from alignments is inadvisable for at least two reasons. First, one may end up discarding considerable portions of the primary data, some of which may be informative. In some cases, insertion and deletion events themselves are informative for phylogeny estimation (*20*). In other cases, excluding a gapped position leads to excluding substitutions that occur elsewhere in the tree at that site and are informative (*21*). Moreover, excluding data does not necessarily result in more concordant inferences. Figure 2E shows results of phylogenetic analyses in which gapped sites were excluded from the alignments. One still finds many genes for which phylogenetic inferences differ among alignment treatments. Second, when an appropriate statistical method of analysis is applied, one may be able to make conclusions even in the face of alignment uncertainty. For example, it might be that the number and identity of positively selected sites differ among alignment treatments. However, when the alignment uncertainty is properly accounted for, one may still be able to pick out some sites that are consistently under positive selection.

The common statistical procedure for accounting for parameter uncertainty is to treat the parameter as a random variable and sum or integrate over the uncertainty, weighting each possible value of the parameter by its prior probability. In a comparative genomics study, we advocate that alignment be treated as a random variable, and inferences of parameters of interest to the genomicist, such as the amount of nonsynonymous divergence or the phylogeny, consider the different possible alignments in proportion to their probability. Considering alignment as a random variable is innate to the statistical alignment pro-

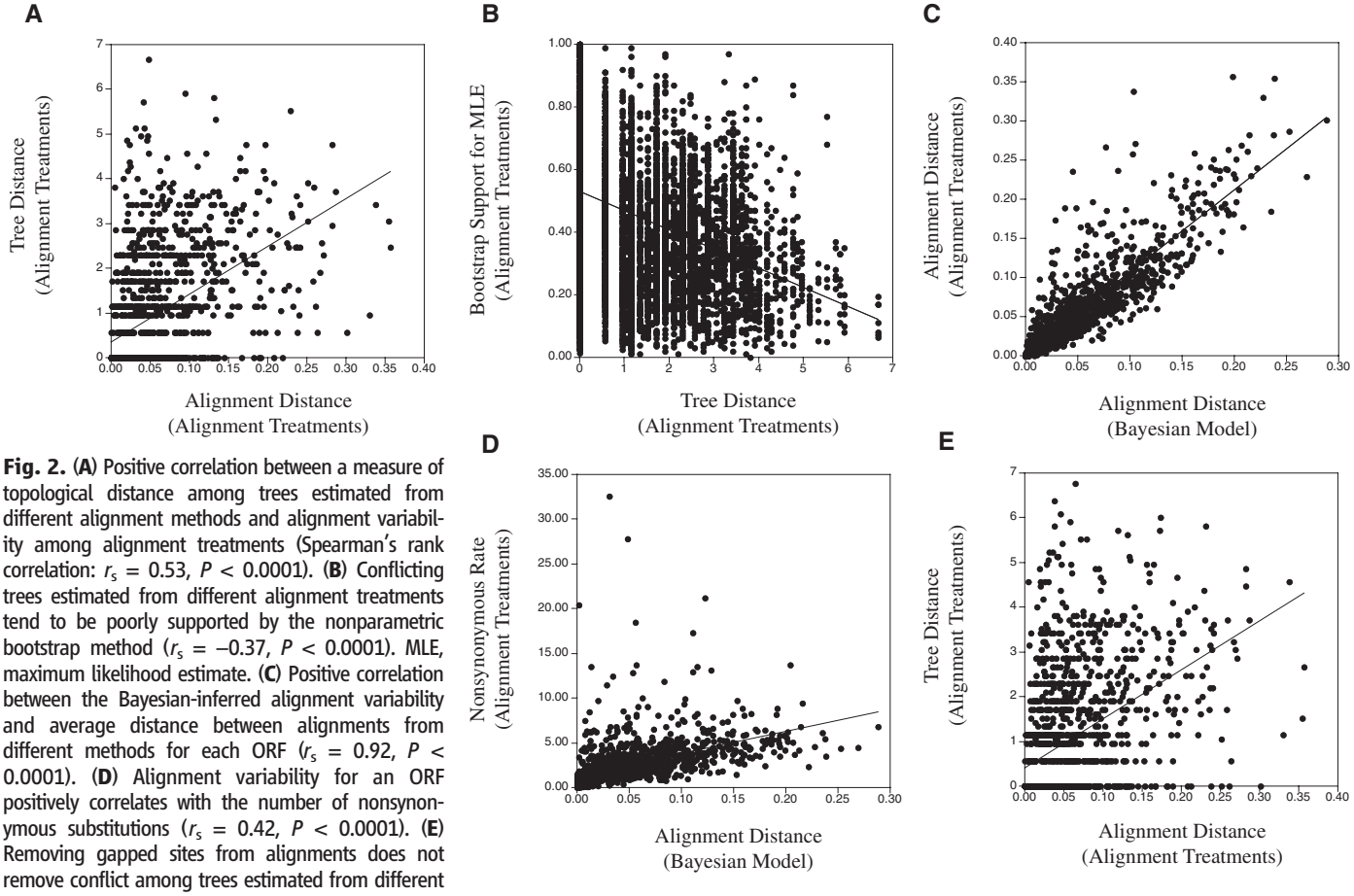

Fig. 2. (**A**) Positive correlation between a measure of topological distance among trees estimated from different alignment methods and alignment variability among alignment treatments (Spearman's rank correlation: $r_s = 0.53$, $P < 0.0001$). (**B**) Conflicting trees estimated from different alignment treatments tend to be poorly supported by the nonparametric bootstrap method ($r_s = -0.37$, $P < 0.0001$). MLE, maximum likelihood estimate. (**C**) Positive correlation between the Bayesian-inferred alignment variability and average distance between alignments from different methods for each ORF ($r_s = 0.92$, $P < 0.0001$). (**D**) Alignment variability for an ORF positively correlates with the number of nonsynonymous substitutions ($r_s = 0.42$, $P < 0.0001$). (**E**) Removing gapped sites from alignments does not remove conflict among trees estimated from different alignment treatments ($r_s = 0.52$, $P < 0.0001$).

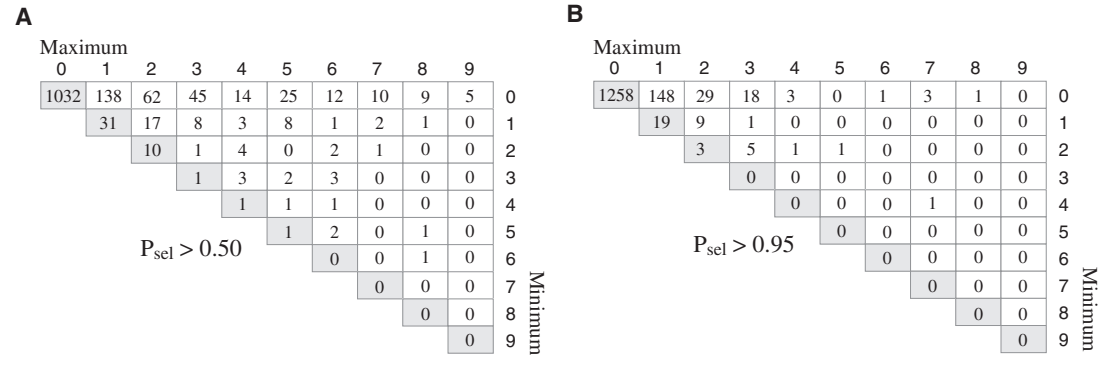

Fig. 3. (**A**) The range in the number of positively selected sites for each ORF. Inferences of positive selection for an ORF are consistent across alignment treatments when the minimum and maximum number of positively selected sites are equal. In many cases (426 of 1502 ORFs), inferences of positive selection varied depending upon the alignment treatment. (**B**) Increasing stringency for inferring positive selection to 0.95 decreases the number of sites inferred to be under positive selection; there remain many cases (222 of 1502 ORFs) in which inferences of positive selection differ according to alignment treatment.

cedure advocated by many (*22*–*24*). Statistical alignment, however, generally assumes that the phylogeny is known, a condition often violated in comparative genomics studies. Moreover, many biologists appear to take the position that when an alignment has been carefully constructed, incorporating uncertainty is unnecessary; in a phylogenetic study, for example, the phylogenetic marker is carefully selected because it is easy to align and has a substitution rate appropriate to the phylogenetic problem of interest (*25*), a selectivity that may help, but probably does not solve, the alignment uncertainty problem in many phylogenetic studies, especially those for anciently diverged species. In comparative genomics studies, however, the goal is to analyze all of the genes in the genome. As we have shown here, many of these genes will be difficult to align and result in highly variable evolutionary parameter estimates. Allowing for uncertainty in the alignment and, possibly, phylogeny simultaneously, through statistical phylo-alignment, should be of special importance in comparative genomics studies.

References and Notes

1. Z. Yang, R. Nielsen, N. Goldman, A. Pedersen, *Genetics* **155**, 431 (2000).
2. P. F. Cliften *et al*., *Genome Res.* **11**, 1175 (2001).
3. P. Cliften *et al*., *Science* **301**, 71 (2003).
4. M. Kellis, N. Patterson, M. Endrizzi, B. Birren, E. Lander, *Nature* **423**, 241 (2003).
5. J. D. Thompson, D. G. Higgins, T. J. Gibson, *Nucleic Acids Res.* **22**, 4673 (1994).
6. R. C. Edgar, *Nucleic Acids Res.* **32**, 1792 (2004).
7. C. Notredame, D. Higgins, J. Heringa, *J. Mol. Biol.* **302**, 205 (2000).
8. B. Morgenstern, *Bioinformatics* **15**, 211 (1999).
9. K. Katoh, K. Misawa, K. Kuma, T. Miyata, *Nucleic Acids Res.* **30**, 3059 (2002).
10. J. Stoye, *Gene* **211**, GC45 (1998).
11. C. B. Do, M. S. P. Mahabhashyam, M. Brudno, S. Batzoglou, *Genome Res.* **15**, 330 (2005).
12. J. A. Lake, *Mol. Biol. Evol.* **8**, 378 (1991).
13. D. A. Morrison, J. T. Ellis, *Mol. Biol. Evol.* **14**, 428 (1997).
14. N. B. Mugridge *et al*., *Mol. Biol. Evol.* **17**, 1842 (2000).
15. D. F. Robinson, L. R. Foulds, *Math. Biosci.* **53**, 131 (1981).
16. B. D. Redelings, M. A. Suchard, *Syst. Biol.* **54**, 401 (2005).
17. M. A. Suchard, B. D. Redelings, *Bioinformatics* **22**, 2047 (2006).
18. A. Schwartz, E. W. Myers, L. Pachter, http://arxiv.org/abs/q-bio.QM/0510052.
19. A. G. Clark *et al*., *Science* **302**, 1960 (2003).
20. B. D. Redelings, M. A. Suchard, *BMC Evol. Biol.* **7**, 40 (2007).
21. F. Lutzoni, P. Wagner, V. Reeb, S. Zoller, *Syst. Biol.* **49**, 628 (2000).
22. J. L. Thorne, H. Kishino, J. Felsenstein, *J. Mol. Evol.* **33**, 114 (1991).
23. I. Holmes, W. Bruno, *Bioinformatics* **17**, 803 (2001).
24. J. Hein, J. Jensen, C. Pedersen, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 14960 (2003).
25. A. Graybeal, *Syst. Biol.* **43**, 174 (1994).

26. This research was supported by NSF (DEB-0445453) and NIH (GM-069801) grants (J.P.H.) and an Alfred P. Sloan Research Fellowship (M.A.S.).

Supporting Online Material

www.sciencemag.org/cgi/content/full/319/5862/473/DC1
SOM Text
References

9 October 2007; accepted 6 December 2007
10.1126/science.1151532

NFAT Binding and Regulation of T Cell Activation by the Cytoplasmic Scaffolding Homer Proteins

Guo N. Huang,[1,2]* David L. Huso,[3]† Samuel Bouyain,[4]† Jianchen Tu,[2]† Kelly A. McCorkell,[5]† Michael J. May,[5] Yuwen Zhu,[6] Michael Lutz,[7] Samuel Collins,[7] Marlin Dehoff,[2] Shin Kang,[2] Katharine Whartenby,[7] Jonathan Powell,[7] Daniel Leahy,[4] Paul F. Worley[2,8]‡

T cell receptor (TCR) and costimulatory receptor (CD28) signals cooperate in activating T cells, although understanding of how these pathways are themselves regulated is incomplete. We found that Homer2 and Homer3, members of the Homer family of cytoplasmic scaffolding proteins, are negative regulators of T cell activation. This is achieved through binding of nuclear factor of activated T cells (NFAT) and by competing with calcineurin. Homer-NFAT binding was also antagonized by active serine-threonine kinase AKT, thereby enhancing TCR signaling via calcineurin-dependent dephosphorylation of NFAT. This corresponded with changes in cytokine expression and an increase in effector-memory T cell populations in Homer-deficient mice, which also developed autoimmune-like pathology. These results demonstrate a further means by which costimulatory signals are regulated to control self-reactivity.

T cells are activated through the TCR and costimulatory pathways predominantly mediated by the cell surface receptor CD28. Although these pathways are relatively well defined, questions still remain about how costimulatory signals are regulated. The Homer family of cytoplasmic scaffolding proteins are known to function at the neuronal excitatory synapse (*1*, *2*), although their wide tissue distribution, including within the immune system, suggests that their functions may be relatively broad.

To investigate the in vivo functions of the Homer proteins, we generated mice in which the loci for each Homer gene were deleted (Homer1, 2, and 3). Of these, we noted that the Homer3-deficient mice (*3*) displayed lymphocyte infiltration of multiple organs and hyperplasia in lymph nodes by 10 weeks of age (fig. S1), which suggested that at least one of the family might possess some level of immune function. Because Homer proteins typically have redundant roles (*1*, *2*), we first assessed their possible role in T cell activation, by assaying interleukin-2 (IL-2) production in T cells lacking all three genes (TKO). IL-2 production was increased by a factor of 2 to 6 in anti-CD3–stimulated T cells from Homer TKO mice relative to wild-type controls (Fig. 1A). By contrast, when T cells were activated by costimulation of both CD3 and CD28, no measurable difference in IL-2 production was detected between wild-type and Homer-deficient mice (fig. S2).

To examine the potential role of Homer proteins in T cell activation in more detail, we used short hairpin RNAs (shRNAs) to knock down Homer gene expression in human Jurkat T cells (Fig. 1B). Knockdown of Homer2 or Homer3, but not Homer1, enhanced the expression of a luciferase reporter driven by the IL-2 promoter by a factor of 3 to 6 (Fig. 1C). Homer2 and Homer3 appeared to have redundant functions in these assays because overexpression of Homer2, but not Homer1, could rescue the loss of Homer3 (Fig. 1D). The IL-2 promoter integrates signals from the calcineurin-NFAT, MAPK-AP1, and NF-κB pathways (*4*, *5*); to identify which pathways might be regulated by Homer, we used luciferase reporter constructs under the control of multimerized binding elements for individual transcription factors. The calcineurin-NFAT pathway was preferentially enhanced in cells depleted of Homer2 or Homer3 (Fig. 1E). To respond to calcium signals, NFAT is first dephosphorylated by calcineurin (*4*), and in Jurkat T cells that expressed shRNAs targeting Homer3, enhanced dephosphorylation of the NFATc2 isoform was observed after activation but not under basal conditions (Fig. 1F). No difference

[1]Program in Biochemistry, Cellular and Molecular Biology, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA. [2]Solomon H. Snyder Department of Neuroscience, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA. [3]Department of Molecular and Comparative Pathobiology, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA. [4]Department of Biophysics and Biophysical Chemistry, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA. [5]Department of Animal Biology, University of Pennsylvania, Philadelphia, PA 19104, USA. [6]Department of Dermatology, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA. [7]Department of Oncology-Immunology/Hematopoiesis, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA. [8]Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA.

*Present address: Department of Molecular Biology, University of Texas Southwestern Medical Center, Dallas, TX 75235, USA.
†These authors contributed equally to this work.
‡To whom correspondence should be addressed. E-mail: pworley@jhmi.edu