Title: Short branches lead to systematic artifacts when BLAST searches are used as surrogate for phylogenetic reconstruction.

Authors: Amanda A. Dick^{a*}, Tim J. Harlow^a, and J. Peter Gogarten^{a,b*}.

a Department of Molecular and Cell Biology, University of Connecticut, 91 North Eagleville Road Unit 3125, Storrs, Ct USA 06236-3125.

b Institute for Systems Genomics, University of Connecticut, Storrs, Ct USA 06236-3125

* Corresponding Authors:

Amanda Dick

Department of Molecular and Cell Biology

University of Connecticut

91 North Eagleville Rd, Unit 3125

Storrs, CT 06269

United States of America

Dick.amanda.a@gmail.com, Amanda.Dick@uconn.edu

Johann Peter Gogarten

Department of Molecular and Cell Biology

University of Connecticut

91 North Eagleville Rd, Unit 3125

Storrs, CT 06269

United States of America

jpgogarten@gmail.com , gogarten@uconn.edu

Phone: 1 860 486 4061

Abstract

Long Branch Attraction (LBA) is a well-known artifact in phylogenetic reconstruction when dealing with branch length heterogeneity. Here we show another phenomenon, Short Branch Attraction (SBA), which occurs when BLAST searches, a phenetic analysis, are used as a surrogate method for phylogenetic analysis. This error also results from branch length heterogeneity, but this time it is the short branches that are attracting. The SBA artifact is reciprocal and can be returned 100% of the time when multiple branches differ in length by a factor of more than two. SBA is an intended feature of BLAST searches, but becomes an issue, when top scoring BLAST hit analyses are used to infer Horizontal Gene Transfers (HGT)s, assign taxonomic category with environmental sequence data in phylotyping, or gather homologous sequences for building gene families. SBA can lead researchers to believe that there has been a HGT event when only vertical descent has occurred, cause slowly evolving taxa to be over-represented and quickly evolving taxa to be under-represented in phylotyping, or systematically exclude quickly evolving taxa from analyses. SBA also contributes to the changing results of top scoring BLAST hit analyses as the database grows, because more slowly evolving taxa, or short branches, are added over time, introducing more potential for SBA. SBA can be detected by examining reciprocal best BLAST hits among a larger group of taxa, including the known closest phylogenetic neighbors. Therefore, one should look for this phenomenon when conducting best BLAST hit analyses as a surrogate method to identify HGTs, in phylotyping, or when using BLAST to gather homologous sequences.

Key words: Short Branch Attraction, Long Branch Attraction, LBA, Top Scoring Blast Hit, Gene Transfer Detection, Branch Length Space, Phylotyping, HGT.

Background 1

BLAST stands for Basic Local Alignment Search Tool (Altschul et al., 1997) and is freely available from the NCBI. It is currently the most frequently used bioinformatics tool, because of its speed and ease of use. The user needs to supply a query sequence and choose a database and with a few simple clicks, BLAST will search the database for matches in sequence similarity to the query. This is incredibly useful for a number of purposes. For example, if an unknown gene is sequenced and a close match is present in the database, BLAST results will indicate what type of gene it is and what species it came from. If only distant relatives are in the database, BLAST can still provide information on the function of the sequence by identifying functional domains present. This works, to a point, even if the domains are used in a novel way. BLAST is also used in identifying genes in whole genome sequencing and is used in the most effective gene calling programs.

BLAST can do all of these things because it is a tool for detecting sequence similarity. If what a researcher wants to do is detect sequence similarity, BLAST is a great program to use. A problem only arises, when researchers overreach and try to use BLAST for more than identifying sequence similarity. In recent years, BLAST has been used with increasing frequency to inform on phylogeny. This first started by using BLAST to generate datasets with appropriate orthologs from various sequenced genomes, for further analysis with phylogenetic reconstruction programs. The unspoken assumption in this case was that the later appropriate phylogenetic reconstruction would erase any bias from using BLAST to choose the sequences included in the dataset. As sequencing ability increased, ability to analyze data lagged. In the beginning, due to the infeasibility of whole genome phylogenetic reconstruction, the easiest analysis to perform was to BLAST the entire genome. This led to BLAST being used to detect Horizontal Gene

Transfer, by assuming the top hit was phylogenetically informative. As analyses needed to be done on larger and larger scales, BLAST was put to new and unintended uses.

But has the usefulness of BLAST overreached when it comes to phylogenetic inference? What can a program built to determine sequence similarity tell the user about the underlying phylogeny of the sequences? What is the link between phylogeny and sequence similarity? In this paper, we propose that that link is total branch length. Furthermore, it is the total amount of substitution between any two homologous sequences that determines the strength of a match when it comes to sequence similarity. This has implications for branch length heterogeneity interfering with phylogenetic inference by sequence similarity detection algorithms, such as BLAST. The following experiments were done with the aim of illustrating total branch length as the link between sequence similarity and phylogeny and to clarify the situations in which sequence similarity can be used as a proxy for phylogeny. To explore how BLAST results vary with respect to branch length, we use simulated sequence evolution along a family of trees depicted in Figure 1 Panel H.

Many studies have screened for horizontally transferred genes based on top scoring BLAST hit analyses (McNulty et al., 2012; Nelson et al., 1999; Ruepp et al., 2000; Worning et al., 2000; Zhaxybayeva et al., 2009). When unidirectional analyses were shown to be phylogenetically uninformative (Koski and Golding, 2001), a switch was made to the more restrictive reciprocal top scoring BLAST hit approach (e.g., (Dagan et al., 2010)). Reciprocal Top Scoring BLAST hits are also used in gene family reconstruction to determine homologous gene families (Zhaxybayeva et al., 2006). In addition, BLAST is used in phylotyping or ribotyping analyses where environmental sequences are assigned a taxonomic affiliation based on top scoring BLAST hit (Basaran et al., 2001; Maloy et al., 2009). In doing these types of analyses one

assumes that the top scoring BLAST hit will be the sister taxon, especially if the hit is reciprocal. BLAST was developed to find regions of homology and pick out the most similar sequence from a database, not to inform on phylogeny (Altschul et al., 1990). It is well established that the top scoring BLAST hit does not necessarily represent the nearest phylogenetic neighbor (Eisen, 2000; Koski and Golding, 2001) and this is indeed expected, because the nearest neighbor is not always the most similar. However, the question of how useful BLAST searches might be in matters of phylogeny is still left unanswered. Specifically branch-length space has never been mapped with respect to the ability of BLAST to recapitulate phylogeny.

BLAST searches work by comparing two sequences at a time and looking for local alignments using the Maximal Segment Pair (MSP) score (Altschul et al., 1990). In a gapped BLAST search (Altschul et al., 1997), the current standard for BLAST, it first matches identical residues between two sequences and stores this information. Continuous regions with high scoring segments are identified, using a threshold score as a cutoff to speed up the process, and are then extended, using a similarity criterion. The highest scoring regions are joined using dynamic programming. The best match made in the pairwise alignments is reported as the top scoring BLAST hit and is a factor of the length and identity of the match. Longer matches are considered good measures of homology and these matches can include indels and small regions of mismatch. Fewer mismatched positions result in a higher percent identity and a stronger match, as measured by E-value, which gives the expected number of matches of a given quality due to chance alone. It has been shown that the phylogenetic nearest neighbor does not always produce the top scoring BLAST hit (Koski and Golding, 2001), but given the pairwise nature of the BLAST algorithm, might it be that the taxon with the shortest total evolutionary distance will be the top scoring hit?

With a normal BLAST search (blastn or blastp), an entire database is searched, but the other sequences in the database do not influence the results of an individual search, except that E-value is proportional to the size of the databank (Altschul et al., 1990). This differs from phylogenetic reconstruction programs, in which other sequences are used to give weight to conserved residues and calculate the evolutionary model when calculating the tree (Guindon and Gascuel, 2003; Ronquist and Huelsenbeck, 2003). Inclusion of an outgroup can allow for rooting the tree and identification of synapomorphies and symplesiomorphies. Without knowledge of the outgroup, it is impossible to distinguish synapomorphies from symplesiomorphies and the two categories of matches are therefore given equal weight in the BLAST search. The importance in using synapomorphies to the exclusion of symplesiomorphies in phylogenetic reconstruction traces back to Willi Hennig, who first expounded on their importance in forming monophyletic groups (Hennig, 1975, 1965). If monophyletic groups are not used, then there is increased ambiguity in detection of HGT. With phenetic programs, such as BLAST, where unrooted trees are used, there is no way of differentiating a synapomorphy from a symplesiomorphy. The remaining question is at which sets of branch lengths will this cause a problem?

If time that passed is proportional to branch lengths, then we expect that the length of interior branches should ensure that the shortest total branch length is between sister taxa, as shown in Figure 1 panel A. If some branch lengths are shorter, the shortest total branch length may be between the two short branches. For the example tree given in Fig 1 B, the condition for non-sister taxa to be connected by the shortest overall branch is given by Equation 1.

$$\text{Equation 1: } Y + X > Z + 2X$$

Where Y is the length of the invariable branches (0.1 substitutions per site in Figure 1 panels A, B, and H), Z is the length of the central branch, and X is the length of the variable branches.

Equation 1 reduces to Equation 2. Because the relationship between sequence similarity and phylogenetic reconstruction using BLAST is dependent on total branch lengths, these equations mean that Short Branch Attraction is likely to occur, if the combined lengths of one of the short branches and the central branch are shorter than the length of the invariable branch, as seen in Figure 1 panel B.

Equation 2: $Y > Z + X$

Short Branch Attraction differs from the well-known Long Branch Attraction (LBA), because Long Branch Attraction is a phenomenon of phylogenetic reconstruction. Short Branch Attraction is a phenomenon of phenetic reconstruction. In cases of Short Branch Attraction, it is the lack of evolution between slowly evolving taxa that causes the artifact. Conversely, LBA has been demonstrated in phylogenetic reconstruction and is caused by numerous convergent apomorphies resulting in homoplasies and the attraction of long branches, while the short branches behave normally (Felsenstein, 1978; Huelsenbeck and Hillis, 1993; Swofford et al., 2001).

As the number of taxa included in the database grows, and the phylogenetic relationship between them increases in complexity, the behavior of the BLAST search becomes more difficult to predict. If only one taxon is on a short branch, as shown in Figure 1 panel C, then the shortest pairwise distances are still between sister taxa. In these cases top scoring BLAST hit analyses will be phylogenetically informative and should minimally return a top scoring BLAST hit to a taxon within the same domain as the query. Then top scoring BLAST hits to another domain really would represent Horizontal Gene Transfers (HGTs), or even ancestral HGTs, as shown in figure 1 panel D. In the case of ancestral HGTs, a closely related group of organisms share, through vertical inheritance, a HGT. BLAST would then return the other members of this

clade before returning members of the domain from where the transferred gene originated. Problems arise when a second clade is evolving as slowly as the one of interest, as shown in Figure 1 panel E with the Archaea on short branches. Here the top scoring BLAST hit will return the Archaea when given a slowly evolving bacterial query, even in cases where there is no HGT. In this tree, the red clade on the left is still the sister clade, but the shortest total branch lengths are to the Archaea, thus the Archaea will be returned as the top scoring BLAST hit.

Short Branch Attraction explains some instances of why BLAST searches return different results, in terms of HGTs, as the non-redundant database grows. The apparent HGTs observed from the short bacterial branch attracting to the short archaeal domain, is no longer present when other slowly evolving Bacteria are sequenced, as shown in Figure 1 panel F. This is because the shortest total branch lengths are now between the two slowly evolving Bacteria, creating a new stronger Short Branch Attraction. This phenomenon can also work the other way around, if before there were no slowly evolving taxa from other domains sequenced and then one is sequenced and added to the database, as shown in Figure 1 panel G. Here Short Branch Attraction is created when before there was none.

With gene family reconstruction, Reciprocal Top Scoring BLAST Hit analysis is normally a stepping stone to further phylogenetic analysis and not a proxy for phylogenetic analysis itself. Since homology is an all or nothing trait, technically there are no degrees of homology and any homolog included into the family is appropriate for gene family reconstruction. Inappropriate results are only possible when homologs are excluded. A bias in exclusion introduced at this step will likely be perpetuated further down the line, with certain homologs being left out of their appropriate larger gene family. There is currently no way to measure the rate of false negatives with a BLAST search and if these false negatives were

subject to a systematic bias, then it would be very difficult to determine that it was happening at all. Similarities shared with an ancestral sequence, due to Short Branch Attraction, could reasonably account for a systematic bias in gene family reconstruction.

To study the effect of Short Branch Attraction on BLAST searches, simulations of nucleotide sequence evolution were performed using various branch lengths to determine where in branch length space BLAST searches are subject to Short Branch Attraction and where the top scoring BLAST hit represents the sister taxon. Three sequence lengths, 10000, 1000, and 200 nucleotides, were simulated, to demonstrate the effect of sequence length on branch length space. With increasing sequence length, there is more data available for the program to use when deciding between two matches and the transition zone between returning the phylogenetically informative results and the Short Branch Attraction results is expected to be sharper. 1,000 nucleotides is a common sequence length and should demonstrate what the BLAST program can be expected to do in the case of typical real gene sequence data. 200 nucleotides is the length of typical environmental sequences and will be applicable to phylotyping or ribotyping analyses. A short sequence length should provide less data and therefore less information when choosing between the various possible matches. Thus the shorter sequence length is expected to show a more gradual transition between the Short Branch Attraction and the phylogenetically informative states.

Methods 2

Nucleotide datasets were simulated using Seq-gen (Rambaut and Grassly, 1997) using four taxa trees, as shown in Figure 1 panel H. Central branch lengths vary from 0.004 to 0.1 substitutions per site and variable branch lengths varying from 0.0001 to 0.1 substitutions per site. The lengths of the 2 invariable branches were held at 0.1 substitutions per site for all

simulations. These trees were rooted midway along the central branch, as indicated by the ^, to ensure symmetry. Three replicates of 100 simulated datasets were performed for each tree. Each dataset contained four sequences and was generated using the Jukes Cantor 69 substitution model, because of its simplicity (Jukes and Cantor, 1969). Three analyses were performed that vary with respect to sequence length: 10000, 1000, and 200 nucleotides. Each of the four taxa of each dataset were used as query against a database comprised of the other three taxa in the dataset to determine the top scoring BLAST hit in all four possible directions. Classic nucleotide BLAST, as implemented in blastall 2.2.19 (Altschul et al., 1997), was used. The results for the 100 datasets in a replicate were totaled to produce the percent of time the top scoring BLAST hit returned the sister taxa, which was graphed using rgl, a 3-dimensional graphing package, in R (Adler and Murdoch, 2012; R Developmental Core Team, 2008). Perl scripts were used to perform repetitive tasks en masse.

Results and Discussion 3

Figure 2 Panel A shows the results of the top scoring BLAST hit analysis using the simulated datasets of 10000 nucleotides in length and the mapping of branch-length space with respect to BLAST result. This graph is a 2-dimensional projection of a 3-dimensional plot; to view a short movie of this plot being rotated in 3 dimensions, see Supplemental Movie 1. There are three taxa in each database for the query to possibly match, but BLAST almost always returns either the sister taxa or the Short Branch Attraction taxa. Whenever the sequence belonging to the sister taxon is not returned, the sequence belonging to the slowly evolving taxa

is returned. When the sister taxon is returned in 0% of the simulations, then the Short Branch Attraction artifact occurs in 100% of the simulations.

Branch lengths are equal in the rear right corner of Figure 2 Panel A and as predicted, the top scoring BLAST hit did return the sister taxon with 100% success in these simulations. As branch lengths shorten, either along the central branch or along the peripheral branches, Short Branch Attraction plays an increasing role until the Short Branch Attraction taxon is returned in 100% of cases. There is a very sharp transition between the two states, as expected with such long sequences. Short Branch Attraction is a major problem when either the central or peripheral branches are one-tenth the length of the invariable branches. It is interesting to note that the effect of short branches is additive so that the effect of decreasing the length of the central and variable branches together is combined (*cf.* equation 1). When the central and variable branches are both half the length of the invariable branches, indicated in the graph with an asterisk, the Short Branch Attraction taxon is already returned in approximately 50% of simulations; branches that are the smallest amount shorter return the Short Branch Attraction artifact 100% of the time. This is cause for concern, because branches that differ by a factor of 2 are not normally considered so heterogeneous as to be a problem. Slightly heterogeneous branch lengths like these are commonplace in real data where both evolutionary rate and time between nodes, can vary.

The top scoring BLAST hit analyses were reciprocal, down to the individual simulations, meaning that either Taxa B and D both returned the phylogenetic sister or they both return the Short Branch Attraction taxon. The observed reciprocity results in the curve of the planes for the two analyses having identical shapes and colocalization in branch length space, as shown in figure 2 Panel A.

When the nucleotide sequences are average in length, 1000 nucleotides per sequence, as shown in Figure 2 Panel B (Supplemental Movie 2), the transition is less distinct between returning the closely related sequence and the slowly evolving sequence. There is still a region in the rear right of the graph, where branch lengths are equal and BLAST returns the sister taxon 100% of the time, however, this region occupies a slightly smaller portion of branch length space. There is an extended region where BLAST transitions between the two possible answers, where any simulation stands a chance of returning the Short Branch Attraction taxon or the sister taxon, but not the third possible choice. And there is an extensive region of branch length space where Short Branch Attraction reigns as the most probable top scoring BLAST hit result.

As sequence length decreases even further to 200 nucleotides, as shown in Figure 2 Panel C (Supplemental Movie 3), the transition between sister taxon and the slowly evolving taxon becomes even less distinct. There is no region within the coordinates of this graph where the sister taxon is returned in 100% of the simulations and the region where BLAST returns the sister taxa in the majority of simulations is quite small. There is still a region of branch length space where Short Branch Attraction is returned in 100% of simulations, but this region is less than that seen for the other sequence lengths. There is a much greater proportion of branch length space where it is possible to return either the slowly evolving taxa or the sister taxa as the top hit. This means that shorter sequence lengths are less likely to suffer from Short Branch Attraction, but are also less likely to return the sister taxa. Because of the minuscule amount of information contained in such short sequences, the top BLAST hit of short sequences are increasingly likely to be a result of random chance. The top BLAST hit cannot return any sequence, however, because Taxon C, the third possibility, does not occur even with this short sequence length, when BLAST results seem so intensely ruled by chance, as seen in Supplemental figure 1.

It is possible to provide a mathematic approximation to indicate where in branch length space Short Branch Attraction is likely to be a problem, as shown with the solid curved line in Figure 3, panel A and Equation 2, where Y is set at 0.1 substitutions per site. This mathematical description closely coincides with the observed simulation data in figure 2, as shown by the Xs. Light colored Xs indicate that the sister taxa is returned in greater than 80% of simulations, while dark Xs indicate that the Short Branch Attraction artifact is returned in greater than 80% of cases. This relationship between branch lengths can be used to alert one to the dangers of Short Branch Attraction. When the length of the variable branch plus the length of the central branch is less than the length of the invariable branch, then the shortest evolutionary distance in the tree is between the two short branches and Short Branch Attraction is likely to be a problem.

Figure 3 panel B graphs Equation 2 using values of the length of the invariable branch, Y , ranging from 0.01 to 0.15. The invariable branch is the sister to the variable branch, so when this value is small, the shortest evolutionary distance is to the sister taxon in much of branch length space. As the length of the invariable branch increases, the proportion of branch length space where the shortest total branch length is between the two variable branches, increases. Therefore, Short Branch Attraction dominates branch length space when the invariable branch is long, but not when the invariable branch is short.

LBA is a well-known phenomenon in which heterogeneous branch lengths introduce artifacts into phylogenetic reconstructions, but this is not what is happening here in this phenetic reconstruction. The longest branches are not saturated with substitutions and they are behaving as expected in analyses when used as query. In 100% of cases where Short Branch Attraction is shown grouping the two short branches together, the two longer invariable branches are not subject to this phenomenon and instead return the sister taxon as the top scoring BLAST hit (data

not shown). This confirms that LBA is not the cause of the phenomenon. This observation also suggests a means to detect Short Branch Attraction. In the four taxa tree shown in Figure 1, Panel H, the two taxa picking each other as top scoring BLAST hit as a result of Short Branch Attraction are each the top scoring BLAST hit of their true sister taxon, as shown in Figure 3 Panel C. Therefore, using all taxa included in the analysis as queries will result in non-reciprocal best BLAST hit relationships with the taxa giving rise to Short Branch Attraction. If the sister taxon also returns the taxon of interest as the top scoring BLAST hit, then this reveals possible Short Branch Attraction and a phylogenetic tree should be constructed to elucidate the evolutionary history of the gene of interest. It is interesting to note that the program DarkHorse is already attempting to promote matches to more closely related taxa above those to more distantly related ones (Podell and Gaasterland, 2007). Doing so will likely reduce the effect of Short Branch Attraction in BLAST-based analyses.

Conclusion 4

If BLAST is used only for its original purposes of detecting homology and finding the most similar sequence from a database, then no issues result from returning a sequence that is not the most closely related one. When using BLAST, or any other measure of similarity, as a proxy for phylogeny, slowly evolving taxa attract each other as top scoring hits, creating Short Branch Attraction. Erroneous detection of HGT events or incorrect phylotyping can result from this phenomenon. Slowly evolving paralogs due to an increase in rate heterogeneity, but a similar overall rate of substitution, are more likely to be assembled into gene families. Paralogs with a more homogeneous substitution rate or an overall increase in substitution rate might be systematically left out of gene family assemblies.

Short Branch Attraction, a phenetic artifact, is different from LBA, a phylogenetic artifact and is the result of slowly evolving sequences retaining the same character state. Short Branch Attraction can occur with as little as a factor of two in terms of branch length heterogeneity and contributes to changing results in the detection of putative gene transfers as the non-redundant database grows. To test for Short Branch Attraction one should use all target sequences as queries, or at least the homolog from the expected sister taxon. If the sequence of interest is the best (non-reciprocal) BLAST hit from another homolog, further analyses need to be performed to elucidate the phylogenetic neighbor.

Abbreviations used

SBA: Short Branch Attraction

LBA: Long Branch Attraction

HGT: Horizontal Gene Transfer

MSP: Maximal Segment Pair

BLAST: Basic Local Alignment Search Tool

Indel: Insertion or deletion

Competing interests

The author(s) declare that they have no competing interests.

Authors' contributions

JPG supervises the lab in which the research was conducted, provided intellectual discussion on the topic, and participated in writing the manuscript. He noticed BLAST searches behaving oddly when looking for HGT and suggested it be looked into. Additionally, he graphed Equation 2 in Figure 3 panels A and B. AAD primarily wrote the manuscript, came up with the explanation of slowly evolving sequences causing an artifact, coined the term Short Branch Attraction, worked out Equations 1 and 2, devised the simulations to test whether short branches can cause a result like the one observed in real data, wrote or co-wrote perl scripts to carry out this work, and ran these scripts. TJH wrote or co-wrote perl scripts used to carry out this work and provided intellectual discussion on the topic.

Acknowledgements

David Williams provided critical intellectual discussion on devising the simulations. The UConn Bioinformatics Facility provided computing resources for the analyses reported in this manuscript. This research was supported in part through a grant from the National Science Foundation (DEB 0830024).

References

- Adler, D., Murdoch, D., 2012. RGL: 3D visualization device system (OpenGL).
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi:10.1016/S0022-2836(05)80360-2
- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.
- Basaran, P., Basaran, N., Cakir, I., 2001. Molecular differentiation of *Lactococcus lactis* subspecies *lactis* and *cremoris* strains by ribotyping and site specific-PCR. *Curr. Microbiol.* 42, 45–48.
- Dagan, T., Roettger, M., Bryant, D., Martin, W., 2010. Genome networks root the tree of life between prokaryotic domains. *Genome Biol. Evol.* 2, 379–392. doi:10.1093/gbe/evq025
- Eisen, J.A., 2000. Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. *Curr. Opin. Genet. Dev.* 10, 606–611.
- Felsenstein, J., 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* 27, 401–410.
- Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52, 696–704.
- Hennig, W., 1975. “Cladistic analysis of cladistic classification?”: A reply to Ernst Mayr. *Syst. Biol.* 24, 244–256.
- Hennig, W., 1965. *Phylogenetic systematics*. Univ. Ill. Press Urbana.
- Huelsensbeck, J., Hillis, D., 1993. Success of phylogenetic methods in the four-taxon case. *Syst. Biol.* 42, 247–264.
- Jukes, T., Cantor, C., 1969. Evolution of protein molecules. Munro HN Ed *Mamm. Protein Metab. Acad. Press N. Y.*
- Koski, L.B., Golding, G.B., 2001. The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol.* 52, 540–542. doi:10.1007/s002390010184
- Maloy, A., Culloty, S., Slater, J., 2009. Use of PCR-DGGE to investigate the trophic ecology of marine suspension feeding bivalves. *Mar Ecol Prog Ser* 381, 109–118.
- McNulty, S.N., Abubucker, S., Simon, G.M., Mitreva, M., McNulty, N.P., Fischer, K., Curtis, K.C., Brattig, N.W., Weil, G.J., Fischer, P.U., 2012. Transcriptomic and proteomic analyses of a *Wolbachia*-free filarial parasite provide evidence of trans-kingdom horizontal gene transfer. *PloS One* 7, e45777. doi:10.1371/journal.pone.0045777
- Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn, M.L., Dodson, R.J., Haft, D.H., Hickey, E.K., Peterson, J.D., Nelson, W.C., Ketchum, K.A., McDonald, L., Utterback, T.R., Malek, J.A., Linher, K.D., Garrett, M.M., Stewart, A.M., Cotton, M.D., Pratt, M.S., Phillips, C.A., Richardson, D., Heidelberg, J., Sutton, G.G., Fleischmann, R.D., Eisen, J.A., White, O., Salzberg, S.L., Smith, H.O., Venter, J.C., Fraser, C.M., 1999. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature* 399, 323–329. doi:10.1038/20601
- Podell, S., Gaasterland, T., 2007. DarkHorse: a method for genome-wide prediction of horizontal gene transfer. *Genome Biol.* 8, R16. doi:10.1186/gb-2007-8-2-r16
- R Developmental Core Team, 2008. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rambaut, A., Grassly, N.C., 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.* CABIOS 13, 235–238.

- Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinforma. Oxf. Engl.* 19, 1572–1574.
- Ruepp, A., Graml, W., Santos-Martinez, M.L., Koretke, K.K., Volker, C., Mewes, H.W., Frishman, D., Stocker, S., Lupas, A.N., Baumeister, W., 2000. The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. *Nature* 407, 508–513. doi:10.1038/35035069
- Swofford, D.L., Waddell, P.J., Huelsenbeck, J.P., Foster, P.G., Lewis, P.O., Rogers, J.S., 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Syst. Biol.* 50, 525–539.
- Worning, P., Jensen, L.J., Nelson, K.E., Brunak, S., Ussery, D.W., 2000. Structural analysis of DNA sequence: evidence for lateral gene transfer in *Thermotoga maritima*. *Nucleic Acids Res.* 28, 706–709.
- Zhaxybayeva, O., Gogarten, J.P., Charlebois, R.L., Doolittle, W.F., Papke, R.T., 2006. Phylogenetic analyses of cyanobacterial genomes: Quantification of horizontal gene transfer events. *Genome Res.* 16, 1099–1108. doi:10.1101/gr.5322306
- Zhaxybayeva, O., Swithers, K.S., Lapierre, P., Fournier, G.P., Bickhart, D.M., DeBoy, R.T., Nelson, K.E., Nesbø, C.L., Doolittle, W.F., Gogarten, J.P., Noll, K.M., 2009. On the chimeric nature, thermophilic origin, and phylogenetic placement of the *Thermotogales*. *Proc. Natl. Acad. Sci. U. S. A.* 106, 5865–5870. doi:10.1073/pnas.0901260106

Figure Legends

Figure 1: Relating branch lengths in phylogenetic trees to expected BLAST results. Panel A shows a homogeneous tree with the shortest total pairwise distances between sister taxa. The top scoring BLAST hit is expected to be the sister taxon. Panel B shows a heterogeneous tree where the shortest total pairwise distance is between the two slowly evolving taxa. BLAST is expected to return Short Branch Attraction in this case. Panel C shows a hypothetical three domain tree with only one slowly evolving clade on short branches, where the shortest pairwise distances are still between sister taxa. The top scoring hit is not expected to be out of domain. In cases where

the out of domain top hit has branches of normal length, as shown in Panel D, we can conclude that there really was a horizontal gene transfer event. But if the Archaea are all on short branches, as in Panel E, then the top scoring BLAST hit will return them even in cases where there is no HGT. Panel F shows how Short Branch Attraction might be the cause of discrepancies in top scoring BLAST hit analyses when the database grows, as more slowly evolving sequences are added, creating new, closer, Short Branch Attractions. Panel G shows how adding new slowly evolving species of Archaea to the database could introduce Short Branch Attraction when before there was none. Panel H shows the true tree used for the simulations. The root is indicated by the ^ and is at the midpoint along the central branch. Taxa B and D are on branches of length X. The length of the central branch is represented by Z.

Figure 2: Branch length versus the percent of time the top scoring BLAST hit recapitulates phylogeny. Panel A shows the results of the simulations of 10000 nucleotides in length. The percent of time the top scoring BLAST hit returns the sister taxon is graphed on the Y axis, with 100% meaning that the BLAST search always returns the sister taxon. At 0% the BLAST search always returns Short Branch Attraction; the BLAST search almost never returned the third possibility. Each dot represents the result of 100 simulations. The Z axis gives the length of the central branch, and the X axis the log base 10 of the length of the variable branches. Both the results of the analyses performed using Taxon B as query and those with Taxon D as query are shown, with Taxon B in the bright rainbow colors and D in the darker rainbow colors. The shape of the surface formed by the results is identical for the two analyses, showing that the BLAST hits are reciprocal. For the majority of branch length space, the BLAST search is subject to Short

Branch Attraction. It is only free from this phenomenon when the branches are relatively homogeneous in length. Panel B shows the result of using 1000 nucleotide long sequences in the simulations and demonstrates a more gradual transition between the two states and a smaller region in which the sister taxon is returned at a high frequency. Panel C shows the result of using sequences of 200 nucleotide in length. A much greater proportion of branch length space is now occupied by the transition zone, while the area where the sister taxon is returned the majority of the time is now virtually non-existent.

Figure 3: Ways to determine if Short Branch Attraction is a problem in a particular BLAST search. The solid line in Panel A plots Equation 2: $Y = Z + X$, where Y is set to 0.1 substitutions per site. The region filled with dark colored Xs shows where in branch length space Short Branch Attraction occurs when the invariable branches are of length 0.1 substitutions per site. The region filled with the light colored Xs shows where in branch length space the sister taxon is returned. The Largest Xs represent areas where all three simulations return the result greater than 80% of the time. The medium size Xs represent areas where only the simulations of 10000 and 1000 nucleotides in length return the result greater than 80% of the time. The smallest Xs represent areas where only the simulations of 10000 nucleotides in length return the result in greater than 80% of cases. If data falls into the region of branch length space with the dark colored Xs, top BLAST hit should not be used to inform on phylogeny. Panel B plots Equation 2 for values of Y between 0.01 and 0.15 substitutions per site. The area under the curve represents where Short Branch Attraction is expected with BLAST searches. The area above the curve represents where BLAST is expected to return the sister taxon as the top scoring hit. As the length of the invariable branch, Y , is increased, an increasing proportion of branch length space

is dominated by Short Branch Attraction. In Panel C, the sister taxon indicates when Short Branch Attraction might be occurring. In instances of Short Branch Attraction, both the sister taxon and the other slowly evolving taxon return the taxon of interest as the top scoring BLAST hit.

Supplemental Figure 1: Percent of time Short Branch Attraction is return versus branch length for sequences of 200 nucleotides in length. At 100%, Short Branch Attraction is always returned and occurs when branches are short. The sister is never returned in 100% of simulations, but when branch lengths are equal, the sister taxon is returned the majority of the time, illustrated by the minimum point on this graph. This graph is the inverse of the graph in Figure 2 Panel C, where the percent of time the sister taxa is returned, is graphed.

Supplemental Movie 1: The three dimensional graph of the percent of time BLAST recapitulates phylogeny versus branch length space for sequences of 10000 nucleotides.

Supplemental Movie 2: The three dimensional graph of the percent of time BLAST recapitulates phylogeny versus branch length space for sequences of 1000 nucleotides.

Supplemental Movie 3: The three dimensional graph of the percent of time BLAST recapitulates phylogeny versus branch length space for sequences of 200 nucleotides.

Figure 1

Figure 2

Branch Length versus % of Time Top Scoring Blast Hit Recapitulates Phylogeny

Panel A

Panel B

Panel C

Figure 3

Panel B

Panel C

Short Branch Attraction

Branch Length versus %
of Time Top Scoring Blast
Hit Recapitulates Phylogeny