Much ado about bacteria-to-vertebrate lateral gene transfer

Diane P. Genereux and John M. Logsdon Jr

Department of Biology, Graduate Program in Population Biology, Ecology and Evolution, Emory University, 1510 Clifton Road, Atlanta GA 303, USA

When the International Human Genome Sequencing Consortium (IHGSC) published its draft of the human genome in February 2001, several genes were identified as possible bacteria-to-vertebrate transfers (BVTs). These genes were identified by their highly significant sequence similarity to bacterial genes in BLAST searches, and by their lack of matches among non-vertebrate eukaryote genes. Many were later rejected as BVTs by several methods, including recovery of probable orthologs from the genomes of incompletely sequenced eukaryotes. Whereas the BVT issue has received considerable attention, there has been no compilation of all potential BVTs considered to date, nor any proposal of a single comprehensive method for rigorously establishing the veracity of a putative BVT. In reviewing the work to date, we list all of the proteins examined and propose systematic tests to investigate whether a vertebrate gene proposed as a BVT is indeed of bacterial origin. We use the proposed strategy to test – and reject – one of the BVTs from the original IHGSC list.

At least 113 'genes entered the vertebrate (or pre-vertebrate) lineage by horizontal transfer from bacteria' reported the International Human Genome Sequencing Consortium (IHGSC) [1]. It was clearly a claim that would not go unnoticed. Ponting, an author on the original IHGSC paper, immediately offered commentary [2]. Shortly after, Salzberg *et al.* [3] offered a reanalysis, accompanied by a comment by Andersson [4]. DeFilippis and Villarreal [5] and Roelofs and Van Haastert [6] investigated BVT again, using different techniques. The debate even made the pages of *The New York Times*, where it was dubbed 'a fresh skirmish in the genome wars'.

The idea of gene transfer from a prokaryote to the vertebrate lineage is fascinating. However, the significant mechanical barriers, as well as constraints to natural selection, urge skepticism when considering interdomain transfer. For a gene to be transferred into a vertebrate genome, it would have to cross both the cellular and nuclear membranes to reach a chromosome. Because vertebrates sequester cell lines early in development, transferred DNA would be heritable only if it found its way to a germ cell. Moreover, laterally transferred genes enter recipient populations at tiny frequencies, putting them at high risk of loss by genetic drift. Because selection is weak relative to drift in small populations, fixation of a laterally transferred gene in a characteristically small vertebrate population would require an exceptionally large fitness benefit (see also Refs [2–4]).

But some data could suggest that bacteria-to-vertebrate transfer (BVT) is not as improbable as we might assume. Ponting [2] notes that several of the claimed BVTs could impact phenotypes relevant to fitness, and might therefore be subject to natural selection. For instance, deficiencies in some of the BVT proteins are associated with abnormal maternal behavior and metabolic disorders in humans. However, this argument is problematic because it seems unlikely that proteins having key roles in the function of highly integrated metabolic pathways are of recent origin. The BVT hypothesis requires that the putatively transferred proteins are very young: no older than the vertebrate lineage or its immediate ancestor, which existed about 450 million years ago.

Here, we review the work to date on the BVT hypothesis. Figure 1 is a comprehensive representation of all of the proteins considered to date; names are given for proteins not yet rejected as potential BVTs (for other protein names, see Supplementary Data at http://archive.bmn.com/supp/tig/April2003-Geneveux.html). We then discuss methods that are specific enough to avoid indicating transfer when it has not occurred, and sensitive enough to detect BVT if and when it has.

Perilous BLASTs against a sparse dataset

The IHGSC's original list of BVTs consists of 113 human sequences that, in BLASTP searches, hit bacterial homologs with scores at least nine orders of magnitude better than the best eukaryotic hit [1]. The list of 113 proteins had been whittled down from an initial group of 223. The other 110 were excluded because they were only sparsely distributed among prokaryotes, leaving uncertain their status as characteristic bacterial proteins. The compilation of putative BVTs represents an ambitious effort to uncover an unseen chapter in our genomic history. However, as Koski and Golding [7] have recently shown, uncorrected heterogeneities in amino acid replacement rates make pair-wise sequence similarity – the parameter measured by BLASTP – an unreliable sole indicator of phylogenetic relationships. This problem is especially severe when BLAST searches involve query sequences that have few phylogenetically close homologs among existing sequence data, such as human proteins.

Corresponding author: John M. Logsdon Jr (jlogsdon@biology.emory.edu).

http://tigs.trends.com 0168-9525/03/$ - see front matter doi:10.1016/S0168-9525(03)00055-6

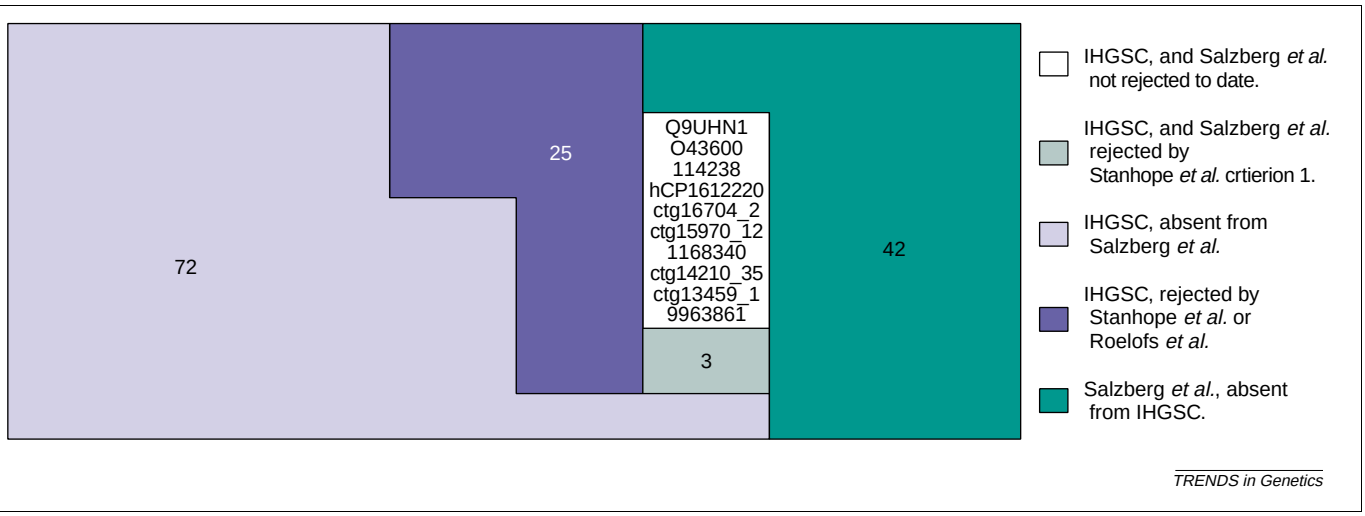

Fig. 1. A comprehensive set of genes considered to date as possible bacteria-to-vertebrate transfers (BVTs). Protein sequences identified independently by IHGSC [1] (110; three of the original 113 proteins could not be accounted for) and Salzberg *et al.* [3] (55) were compiled; 13 appeared in both sets, resulting in a comprehensive, nonredundant set of 152 proteins. Each protein is represented by a box whose color and pattern indicate its status as a BVT. Because *bona fide* BVTs should be initially identifiable by both the IHGSC and Salzberg methods, proteins are unlikely BVTs if they were either unique to Salzberg (42) or to IHGSC (72). Some were both unique to IHGSC *and* directly rejected by either Roelofs and Van Haastert [6] or Stanhope *et al.* [11] (25). Of the 13 remaining, 3 were rejected by Stanhope *et al.* A set of 10 proteins shared by the two sets, and not rejected by any method applied to date, is given (one of these proteins, hCP1612220, was actually unavailable for consideration in the Stanhope *et al.* analysis). From this remaining set of 10 candidates, Q9UHN1 was chosen as an exemplar case to illustrate the phylogenetic analysis required for rigorous testing of the BVT hypothesis (see Fig. 2). For protein accession numbers and more detailed descriptions of methods and observations see Supplementary Data at http://archive.bmn.com/supp/tig/April2003-Geneveux.html.

A cursory analysis of the IHGSC list only heightened our skepticism. When we compared the putative BVTs to themselves using BLASTP, we found that one of the proteins (gi27803) bears more than 50% amino acid identity to five other BVTs, suggesting all six as paralogs. The production of six divergent paralogs from a single transferred gene would require multiple gene duplications and fixations, and significant time for accumulation of sequence differences. There are several duplicate chromosome regions unique to the human genome [8]; however, their production would require only a single duplication event, followed by selection for just one of the genes in that region. By contrast, the production of six paralogs from a laterally transferred gene would require a minimum of three independent duplication events, followed by selection favoring each new duplicate. This series of events seems especially improbable, given the relatively recent origin of vertebrates. In any case, subsequent analyses have clearly rejected these paralogs as BVTs (Fig. 1).

With regard to the BVT hypothesis, the problems of BLAST are especially severe. The number of fully sequenced prokaryotic genomes far exceeds the number of fully sequenced eukaryotic genomes. At last count (July 2002), 82 eubacterial and 16 archaeal, compared with only nine eukaryotic complete genomes, including human, were available at NCBI. BLAST is most successful at finding existing homologs in genomes that have been sequenced in their entirety, because genome completion generally results in gene and protein annotation. The observed pattern of broad representation among prokaryotes and sparse representation among eukaryotes for many of the BVTs is therefore consistent with expectations given sampling bias alone.

Finally, the structural complexity of eukaryotic genes further diminishes the probability of locating potential eukaryotic orthologs via BLASTP, which searches proteins predicted from nucleotide data. Prokaryotic genes are arranged in continuous open reading frames (ORFs), so bacterial protein prediction is fairly simple. By contrast, eukaryotic genes are often riddled with introns, making it difficult to predict the amino acid sequences they encode. Consequently, the number of annotated eukaryotic proteins is smaller, even for genomes well represented at the nucleotide sequence level.

Refining the list of BVTs: approaches to date

Shortly after publication of the IHGSC [1], several investigators acknowledged the shortcomings of the original analysis and applied more rigorous methods to test the BVT hypothesis. Salzberg *et al.* [3] were first to respond. They used the original set of all sequenced human proteins as queries in a BLASTP search of all complete prokaryotic and eukaryotic genomes. From the thousands of human proteins that matched bacterial proteins with expect (E) values of 10^{-10} or better (Box 1), they identified 41 proteins from the IHGSC set, and 46 from the Celera [9] set for which bacterial proteins were the best match as potential BVTs. Salzberg *et al.* produced two main findings. First, increasing the numbers of eukaryotic genomes used in the BLASTP comparisons quickly reduced the number of putative interdomain transfers, suggesting that apparent BVT might actually be the result of sampling bias. Second, using a simple calculation based on estimates of average genome size and the average rate of gene loss, Salzberg *et al.* found that random independent gene loss in several different lineages can by itself account for the lack of non-vertebrate eukaryotic homologs of many human proteins. Together, these findings led Salzberg *et al.* [3] to suggest gene loss from some eukaryotes as a more tenable explanation for the IHGSC's BLAST results. DeFilippis and Villarreal [5] added that every pattern consistent with BVT could alternately be explained by

Box 1. Many BLASTs for BVTs

BLASTP
This uses a query protein to find similar sequences in a database of protein sequences. It calculates pairwise sequence similarity, and returns a list of similar proteins, or 'hits'. The proteins are ranked by the 'expect value', or E value, an index of the statistical significance of observed similarity. Sequences with E values less than 10^{-11} are generally assumed to be homologs.
IHGSC
The IHGSC [1] used proteins predicted from the human genome as BLASTP queries against GenBank to compile the original BVT list. Human sequences with high-scoring 'hits' to a phylogenetically diverse set of bacteria, and no non-vertebrate eukaryotes with scores within nine orders of magnitude of the best bacterial hit were included in the BVT list.
Salzberg et al.
These authors [3] used the complete set of human proteins from both IHGSC and Celera as BLASTP queries against all available complete genome sequences. They identified a total of 87 proteins from the two human genome sets for which top hits were from bacterial genomes. They found an inverse relationship between the number of apparent BVTs and the number of complete genomes queried.
Stanhope et al.
These authors [11] used the 28 putative BVTs confirmed by PCR to query the nonredundant proteins and EST-others databases, and the complete genome sequences of *Drosophila melanogaster* and *Caenorhabditis elegans*.
TBLASTN
This uses a query protein sequence to search nucleotide sequence databases that have been translated in all possible reading frames. Searching nucleotide databases directly removes the need for proteins to be annotated.
Roelofs and Van Haasert
These authors used TBLASTN with human protein sequence as query against the partially complete nucleotide sequence database of the eukaryotic slime mold *Dictyostelium discoideum* [6]. Sequences with E values $<10^{-10}$ were assumed to be homologs.
Stanhope et al.
These authors [11] used 28 putative BVTs confirmed by PCR to query incomplete bacterial nucleotide databases.

transfer from a viral colonist of either an ancestral or more recently diverged vertebrate lineage. In their response, Salzberg and Eisen [10] concurred, but added that the viral hypothesis would be testable only with a larger set of viral genome sequence data.

Taking a different tack, Stanhope *et al.* [11] started not from the raw dataset but from a subset of the 113-member BVT list: those 28 genes whose occurrence in the human genome had been confirmed by PCR. Using either BLASTP or TBLASTN against the nonredundant proteins database, EST-others database, unfinished microbial genomes nucleotides database, and the complete genome sequences of *Drosophila melanogaster* and *Caenorhabditis elegans* (see Box 1 for a description of these methods), they identified probable homologs, aligned the amino acid sequences manually, and constructed phylogenetic trees for each protein.

The findings of Stanhope *et al.* [11] highlight two of the key problems with previous analyses: (1) using BLASTP scores as a marker of phylogenetic relatedness fails to discern the relationships revealed by appropriate sequence alignment and resulting trees, and (2) BLASTing only against complete genomes excludes information available from organisms for which sequencing and/or annotation are not yet complete. Not considering such nucleotide databases would further bias results in favor of the BVT hypothesis: eukaryotes are underrepresented among the completely sequenced genomes, but they are significant targets of partial genome sequencing efforts. Although Stanhope *et al.* [11] analyzed only 28 proteins from the original BVT list, their trees explicitly refute the BVT hypothesis for 16, suggest one vertebrate-to-bacteria transfer, and indicate that none of the remaining proteins are clear cases of BVT.

Later, using the original 113-protein list, Roelofs and Van Haastert [6] used TBLASTN against the partially complete sequence database of the eukaryotic slime mold, *Dictyostelium discoideum*. Using the conservative, but not entirely rigorous, criterion that a gene is an ortholog if a BLAST expectation value of 1010^{-10} or better is achieved, Roelofs and Van Haastert found 11 proteins from the IHGSC list with a clear homolog in the slime mold and 17 with probable homologs. Their phylogenetic tree for one of the putative BVTs, monoamine oxidase (MAO), displays the topology characteristic of a vertically transmitted gene: MAO from *Dictyostelium*, a non-vertebrate eukaryote, falls between prokaryotes and vertebrates, which each form monophyletic groups.

BVT analyses to date have helped both to reveal the pitfalls of the BLASTP approach, and to identify better methods. However, there has been no attempt to construct phylogenetic trees for all of the putative BVTs. As demonstrated, thorough, objective phylogenetic analysis is imperative, because each of the individual methods applied to date imposes its own biases. Using BLAST methods to generate the initial list of BVTs is an obvious first step, but their pitfalls as indicators of phylogeny have been discussed extensively [7,12–14]. TBLASTN is advantageous in not requiring that genes and proteins be predicted, although the complexity of eukaryotic genes predisposes them to scores lower than those for uninterrupted bacterial genes. The decision of Stanhope *et al.* [11] to BLAST against the EST-others database was particularly wise. The results of their analysis reveal the power of proper alignment and tree-building to address central questions of genome evolution.

Joining forces to assess the BVT hypothesis

The IHGSC's attempts to use nonphylogenetic methods to find BVTs are not, themselves, novel. There has been a longstanding effort in studies of molecular evolution to develop 'surrogate' (nonphylogenetic) methods that are effective in detecting laterally transferred genes. These methods include genomic

searches for bacterial ORFs with atypical nucleotide composition, Markov chain-based searches for atypical codon usage, and surveys for genes with unusual phylogenetic distribution. This last method used by the IHGSC to compile the original BVT list. The hope has been that these various surrogate methods would identify essentially the same sets of genes, thereby providing compelling results and reducing the need for time-consuming phylogenetic analysis. However, in a recent study of the *Escherichia coli* genome, Ragan [12] found that different surrogate methods located sets of potential lateral gene transfers (LGTs) whose overlap was even less than expected by chance. The true set of laterally transferred genes might well be the union, rather than just the intersection, of the genes independently identified by each method. On one

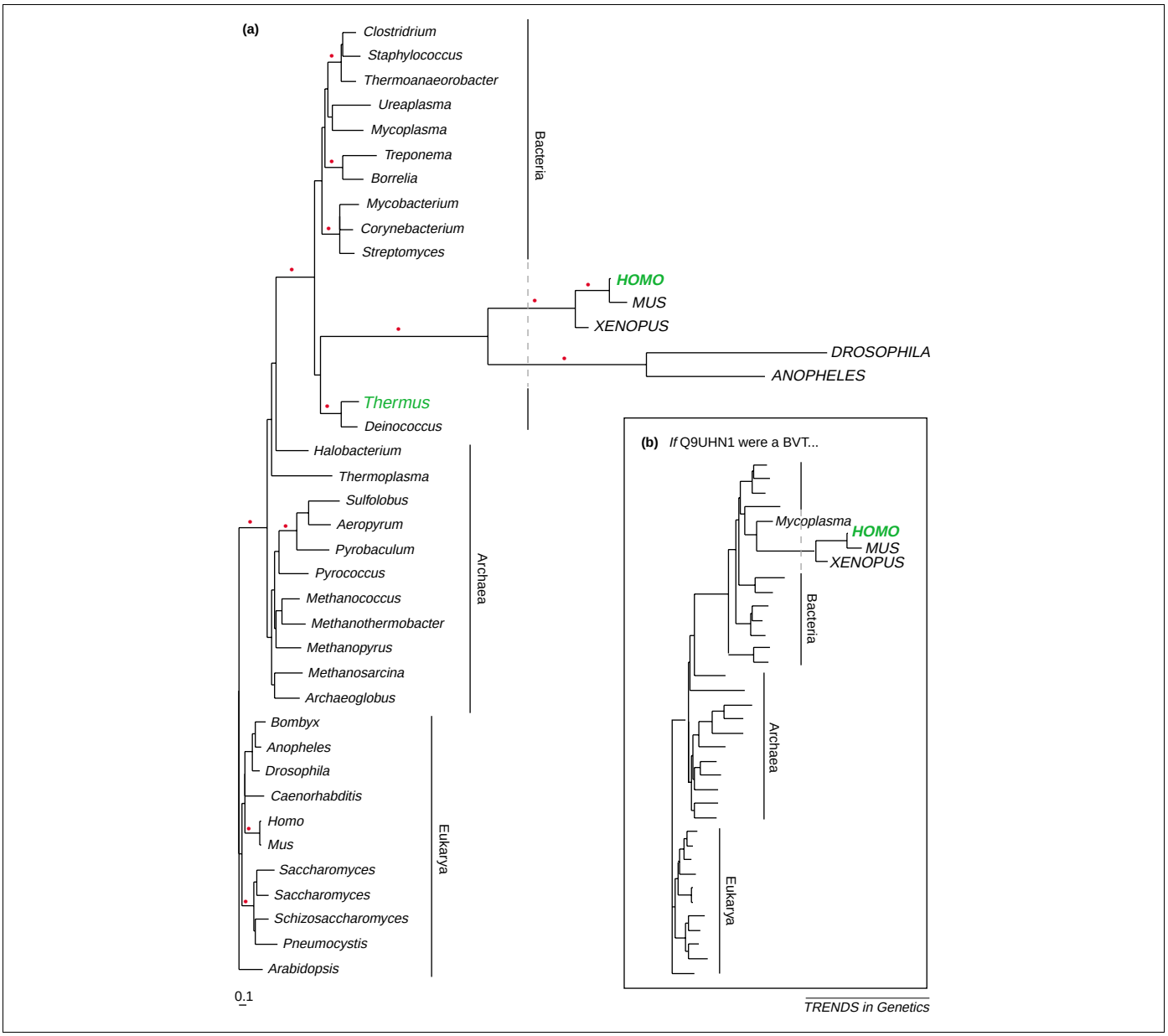

Fig. 2. Phylogenetic analysis and rejection of a candidate BVT. (a) Human protein Q9UHN1 was identified as a candidate BVT by both the IHGSC [1] and by Salzberg *et al.* [3] (Fig. 1). It is 485 amino acids long, and is described as 'DNA polymerase gamma subunit 2, mitochondrial precursor' in the NCBI ENTREZ database. Homologs of Q9UHN1 were collected through BLAST and PSI-BLAST searches of the NCBI nonredundant protein database (in this case, additional searches of the DNA databases using TBLASTN were not required to reject the BVT hypothesis). Taxonomically representative sequences were selected for further analysis including alignment by visual inspection and phylogenetic tree reconstruction. The alignment included 39 sequences and 341 aligned sites (available as Supplementary Data at http://archive.bmn.com/supp/tig/April2003-Geneveux.html). The tree shown was obtained using: (1) TREEPUZZLE 5.0 [15] to create a maximum likelihood distance matrix using the Jones, Taylor and Thornton (JTT) substitution matrix with eight gamma-distributed site rates and one invariable rate; and (2) PHYLIP 3.6 [16] (NEIGHBOR) to construct the tree. Bootstrap support values were determined with 200 replicates (PROTDIST and NEIGHBOR); those nodes with support >90% are marked by red circles. The group of eukaryotic proteins (likely orthologs) including Q9UHN1 is shown in all capitals; Q9UHN1 and the top non-vertebrate BLAST hit to it (from *Thermus thermophilus*) are shown in green. All of the other sequences, including bacterial, archaeal and eukaryal representatives, are glycyl-tRNA synthetases. Although it is possible that the eukaryotic Q9UHN1 orthologs arose by lateral gene transfer from bacteria, the event would have predated the invertebrate–vertebrate divergence. (b) Hypothetical depiction of a resulting phylogenetic tree had Q9UHN1 been transferred from *Mycoplasma* to a vertebrate ancestor.

hand, using any single method to scan for potential LGTs might therefore yield an unacceptably high number of false negatives. On the other hand, using a single method such as nucleotide content or codon usage to identify LGTs might yield an unacceptably high number of false positives, because genes under strong directional selection often acquire atypical characteristics. Ragan speculates that the disparities among the surrogate methods are most severe for older transfers, because, over time, mutational biases and selection for translational efficiency can drive LGTs to ameliorate to (acquire the sequence characteristics of) their new hosts. Given the effects of time since transfer on the extent of sequence amelioration, the disparities among the protein sets methods could be less severe for BVTs, which are necessarily no more than 450 million years old. However, for putative BVTs, as for all potential LGTs, only appropriate phylogenetic analysis can distinguish between real transfers and native genes that are atypical for other reasons.

A truly rigorous test of BVT as a hypothesis to explain the origin of some human proteins would first conduct a systematic survey of the human genome, using several different surrogate methods to compile a comprehensive list of potential LGTs. This would include comprehensive compositional analysis, and systematic TBLASTN and BLASTP searches against available databases of both complete and partial genomes. Each of the proteins compiled would next be subjected to careful manual sequence alignment, and phylogenetic analysis as already demonstrated by Stanhope *et al.* [11].

Our objective here is not to re-do the entire BVT investigation as we believe it should be done, but instead to demonstrate the phylogenetic analysis. Rather than conduct a new comprehensive search for putative BVTs as described above, we provide a comprehensive list of all proteins considered to date as potential BVTs. We include both proteins definitely refuted already, and those that require further analysis (Fig. 1 and Supplementary Data at http://archive.bmn.com/supp/tig/April2003-Geneveux.html). Given Ragan's findings, we do not discount the possibility that there are some candidate BVTs that were omitted from the original lists of the IHGSC and Salzberg *et al.* However, we believe that listing all human genes for which simple BLAST searches reveal a high-scoring bacterial gene and no clear vertebrate ortholog strongly biases the set toward the inclusion of false positives, rather than the exclusion of false negatives. We demonstrate in Fig. 2a the required phylogenetic analysis of one protein not previously analyzed, and its rejection as a BVT. For comparison, we provide a hypothetical depiction in Fig. 2b of a phylogenetic tree indicating a *bona fide* BVT.

It is entirely possible that the presence of *all* putative BVTs in the human genome can be explained by phenomena other than lateral gene transfer. Additional studies might indicate that a *few* human genes are indeed of recent bacterial origin, but their numbers will almost certainly be nowhere near the 100 or more BVTs originally suggested by the IHGSC. Only appropriate phylogenetic analysis can distinguish between methodological artifacts, and compelling evidence that some bacterial genes have traveled a treacherous path of mechanical obstacles and selective improbability to take up permanent residence in the genomes of humans, or of our recent vertebrate ancestors.

Note added in proof

A detailing of the analysis performed as part of the IHGSC has been provided by Koonin *et al* [17]. These authors argue that the putative BVTs could be explained by bacterial transfer to more distant ancestors (such as the one shared by vertebrates and *Dictyostelium*) followed by extensive gene loss. Nonetheless, this is one of many possible scenarios that actually preclude definition of these cases as bacterial-to-vertebrate LGTs.

Acknowledgements

The authors thank Carl Bergstrom, Banoo Malik and Marilee Ramesh for their many helpful comments.

References

1 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
2 Ponting, C.P. (2001) Plagiarized bacterial genes in the human book of life. *Trends Genet.* 17, 235–237
3 Salzberg, S.L. *et al.* (2001) Microbial genes in the human genome: lateral transfer or gene loss? *Science* 292, 1903–1906
4 Andersson, J.O. *et al.* (2001) Genomics. Are there bugs in our genome? *Science* 292, 1848–1850
5 DeFilippis, V.R. and Villarreal, L.P. (2001) Lateral gene transfer or viral colonization? *Science* 293, 1048
6 Roelofs, J. and Van Haastert, P.J. (2001) Genes lost during evolution. *Nature* 411, 1013–1014
7 Koski, L.B. and Golding, G.B. (2001) The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol.* 52, 540–542
8 Bailey, J.A. *et al.* (2002) Recent segmental duplications in the human genome. *Science* 297, 1003–1007
9 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
10 Salzberg, S.L. and Eisen, J.A. (2001) Lateral gene transfer or viral colonization? *Science* 293, 1048
11 Stanhope, M.J. *et al.* (2001) Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature* 411, 940–944
12 Ragan, M.A. (2001) On surrogate methods for detecting lateral gene transfer. *FEMS Microbiol. Lett.* 201, 187–191
13 Logsdon, J.M. and Faguy, D.M. (1999) *Thermotoga* heats up lateral gene transfer. *Curr. Biol.* 9, R747–R751
14 Eisen, J.A. (2000) Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. *Curr. Opin. Genet. Dev.* 10, 606–611
15 Schmidt, H.A. *et al.* (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18, 502–504
16 Felsenstein, J. (1989) PHYLIP – Phylogeny Inference Package (version 3.2). *Cladistics* 5, 164–166
17 Koonin, E.V. *et al.* (2002) Horizontal gene transfer and its role in the evolution of prokaryotes. *Horizontal Gene Transfer*, 2nd edn, (Syvanen, M., Kado, C.I., *et al.* eds), pp. 277–304, Academic Press