Horizontal gene transfer is not a hallmark of the human genome

Steven L. Salzberg

Abstract
Crisp et al. recently reported that 145 human genes have been horizontally transferred from distant species. Here, I re-analyze those genes listed by Crisp et al. as having the highest certainty of having been horizontally transferred, as well as 17 further genes from the 2001 human genome article, and find little or no evidence to support claims of horizontal gene transfer (HGT).

Introduction
A recent study by Crisp et al. [1] re-examined a claim, originally made in the landmark 2001 human genome paper, that bacteria had horizontally transferred 223 genes into a vertebrate ancestor of humans [2]. That claim was refuted soon after the original report [3, 4]. Using an alignment-based scoring scheme, the study by Crisp et al. [1] reported that 145 human genes, including 17 of those from the 2001 study, had been horizontally transferred from distant species. Here, I describe a re-analysis of these 17 genes and of the 28 highest-confidence genes newly claimed by Crisp et al. [1] to have been horizontally transferred, taking a more skeptical perspective, and find little or no evidence to support claims of horizontal gene transfer (HGT).

Hundreds of eukaryotic genomes and thousands of bacterial genomes have been sequenced in the 15 years since the human genome was published. In their recent report, Crisp et al. [1] argue that, with the availability of this far larger collection of genomes, the likelihood of false HGT findings that are actually the result of gene loss is now greatly reduced. Their reanalysis, which was based on a combination of BLAST searches and phylogenetic trees, identified hundreds of “foreign” genes in animals; this led them to claim that HGT “has occurred on a previously unsuspected scale in metazoans” and that it is a significant factor in animal evolution.

In this study, I re-examined the claims of Crisp et al. [1] focusing on the human genes. Instead of using a large-scale, automated analysis, which by its very nature could enrich the results for artifactual findings, I looked at each human gene individually to determine whether the evidence is sufficient to support the conclusion that HGT occurred. An important principal here is that extraordinary claims require extraordinary evidence: there is no doubt that the vast majority of human genes owe their presence in the human genome to the normal process of inheritance by vertical descent. Thus, if other, more mundane processes can explain the alignments of a human gene sequence, these explanations are far more likely than HGT.

Results
For my re-analysis, I re-aligned the 17 human genes that were originally reported as having undergone bacterial-vertebrate transfer (BVT), a finding that has been rejected by our work [3] and that of others [4, 5], but re-claimed by Crisp et al. [1] (Table 1). I found that the evidence does not support HGT for any of them. (One important point worth noting here is that Crisp et al. listed some of these genes as “confirmed” by Salzberg et al. [3]. This was not the case; our previous study invalidated most of the previously claimed HGT events, but was not able to dismiss all of them. Our study made it clear that we did not consider the presence of the remaining genes to be the result of HGT events.) Crisp et al. [1] reported a total of 145 human genes that they claimed to be the result of HGT; 39 of these are labeled in their highest confidence group, class A. Of these 39, seven are included in the first group of 17, leaving 32 newly claimed HGT events. I examined these 32 class A genes (Table 2) and again find no evidence for HGT. A detailed, gene-by-gene description of these analyses can be found in Additional file 1 and the
| Gene group number | Ensembl gene ID | Gene name | Best metazoan, non-chordate match | Best non-metazoan match | Explanation |
|-------------------|----------------|-----------|----------------------------------|-------------------------|-------------|
| 1                 | ENSG00000166743, ENSG00000183747, ENSG000001835187, ENSG00000183549 | Acyl-CoA synthetase medium-chain family members 1, 2A, 3, and 5 | Lingula anatina or Aplysia californica | Desulfotomaculum thermocisternum or D. kuznetsovii | Bacterial HGT No HGT |
| 2                 | ENSG00000047457 | Ceruloplasmin (ferroxidase) | Exaiptasia pallida (sea anemone) | Klebsormidium flaccidum (algae) | Plant HGT No HGT |
| 3                 | ENSG00000107618, ENSG00000265203 | Retinol-binding protein 3 | None | Stackebrandia nassauensis | Bacterial HGT Gene loss |
| 4                 | ENSG00000181019, ENSG00000124588 | NAD(P)H dehydrogenase, quinone 1 and 2 | L. anatina or Capitella teleta | Sinorhizobium fredii | Bacterial HGT No HGT |
| 5                 | ENSG00000132744 | Aspartoacylase (aminocyclase) 3 | Saccoglossus kowalevskii | Oscillatoria sp. PCC 10802 | Bacterial HGT No HGT |
| 6                 | ENSG00000170961, ENSG00000105509, ENSG00000103044 | Hyaluronan synthase 1, 2, and 3 | Papilio xuthus | Lichtheimia ramosa | Fungal HGT Rate variation |
| 7                 | ENSG00000175806 | Methionine sulfoxide reductase A | Lottia gigantea | Pleurocapsa sp. PCC 7319 | Bacterial HGT No HGT |
| 8                 | ENSG00000095596 | Cytochrome P450, family 26, subfamily A, polypeptide 1 | C. teleta | Gentileinema sp. PCC 7407 | Bacterial HGT No HGT |
| 9                 | ENSG00000113790 | Enol-CoA hydratase/3-hydroxyacyl-CoA dehydrogenase | L. anatina | Capsospora owczarzaki | Protist HGT No HGT |
| 10                | ENSG00000166532 | Ribosomal modification protein rimK-like family member B | A. californica | Kitasatospora cheerianensis | Archaean HGT Rate variation |
| 11                | ENSG00000172508 | Carnosine synthase 1 | Crassostrea gigas | Perkinsus marinus | Protist HGT No HGT |

Columns 5 and 7 contain the bitscores of the best BLAST alignment to a gene from the species in columns 4 and 6. Gene group number (column 1) refers to the number used in the main text.
| Gene ID   | Gene name                              | Results from Crisp et al. [1] study | Results from this study |
|----------|----------------------------------------|-------------------------------------|-------------------------|
|          | Gene ID                               | HGT index | Source HGT taxon | Best non-chordate meta-
|          |                                       |           |                | zoonan match | Bit score |
|          |                                       |           |                |              | Best non-chordate meta-
|          |                                       |           |                | zoonan match | Matching protein ID |
|          |                                       |           |                |              | Bit score | New HGT index |
| ENSG00000070269 | C14orf101             | 337        | Protist        | B4LPG6       | 37       | Lingula anatina | XP_013409033 | 566 | -192 |
| ENSG00000102805 | Ceroid-lipofuscinosis, neuronal 5  | 84         | Protist        | E3MUJ2       | 39       | Lottia gigantea | XP_000059770 | 338 | -215 |
| ENSG00000116721 | PRAME family member 1               | 72         | Protist        | A8MV52       | 44       | No hits          |                |      |      |
| ENSG00000157358 | PRAME family member 15              | 45         | Protist        | F1SLUY5      | 89       | No hits          |                |      |      |
| ENSG00000232423 | PRAME family member 6               | 45         | Protist        | F1SLUY5      | 92       | No hits          |                |      |      |
| ENSG00000117115 | Protein-arginine deiminase, types 2,3,1,5 | 464        | Bacteria       | B4IS81       | 45       | Priapulus caudatus | XP_014670176 | 560 | -51  |
| ENSG00000125458 | 5',3'-nucleotidase, cytosolic       | 135        | Bacteria       | E3LP71       | 39       | Saccoglossus kowalevskii | XP_006825056 | 222 | -48  |
| ENSG00000205309 | 5',3'-nucleotidase, mitochondrial   | 108        | Bacteria       | B5DJ4        | 41       | L. anatina       | XP_013404549 | 237 | -88  |
| ENSG00000133561 | GTPase, IMAP family members 6, 4, 7, 5, and 1 | 51         | Plant          | F1QNI4       | 94       | Crassostrea gigas | XP_011437597 | 208 | -63  |
| ENSG000001196329 |                                   |           |                |              |          |                  |                |      |      |
| ENSG00000213203 |                                   |           |                |              |          |                  |                |      |      |
| ENSG00000136153 | LIM domain 7                        | 47         | Protist        | H9JLZ2       | 86       | S. kowalevskii   | XP_006813620 | 268 | -135 |
| ENSG00000136830 | Family with sequence similarity 129, member B | 74         | Protist        | G6OB6       | 44       | Acetobacter digitafera | XP_015774462 | 115 | 3    |
| ENSG00000140718 | Fat mass and obesity associated      | 97         | Plant          | C9J4C3       | 92       | S. kowalevskii   | ALR88588     | 342 | -153 |
| ENSG00000148288 | Globo-side alpha-1,3-N-acetyl/galactosaminyl-
| transerase 1                              | 91         | Bacteria       | A8QEB7       | 39       | Cimex lectularis | XP_014240367 | 37  | 93   |
| ENSG00000154122 | Ankylosis, progressive homolog (mouse) | 129        | Protist        | F5GXN7       | 90       | S. kowalevskii   | XP_006818212 | 388 | -169 |
| ENSG00000172757 | Coflin 1                              | 31         | Fungi          | B5SO8       | 48       | Helobdella robusta | XP_00091217 | 108 | 3    |
| ENSG00000175573 | Chromosome 11 open reading frame 68   | 60         | Fungi          | D6WQ69       | 48       | S. kowalevskii   | XP_002740403 | 157 | -49  |
| ENSG00000177181 | Ribosomal modification protein rimiK-like family member A | 74         | Archaea        | E1GJH3       | 52       | Aplysia californica | XP_012936156 | 124 | 2    |
| ENSG00000212907 | Mitochondrially encoded NADH dehydrogenase 4 L | 69         | Protist        | P15554       | 68       | Xenoturbella bocki | YP_850984 | 67  | 71   |
| ENSG00000216937 | Coiled-coil domain containing protein 7 | 48         | Protist        | A75DV0       | 64       | No hits          |                |      |      |
| ENSG00000224226 | Retrotransposon-derived protein PEG10  | 52         | Fungi          | G0MWG3       | 49       | No hits          |                |      |      |
| ENSG000002256062 | ABO blood group (Histo-blood group ABO system transferase) | 84         | Bacteria       | H3JNK7       | 37       | S. kowalevskii   | XM_006825840 | 218 | -97  |

The HGT index, defined by Crisp et al. [1] as the difference in the best bitscore of a BLAST match to a non-metazoan and a metazoan species, is shown along with the bitscore of the best metazoan match. The recomputed HGT index (last column) is computed by subtracting the bitscore of best non-metazoan found by Crisp et al. [1] from that of the best non-chordate metazoans found by the new searches reported here. "No hits" means that no significant alignments were found to any non-chordate metazoans. For PRAME family members 1, 6, and 15, the protist alignment found by Crisp et al. [1] is a false positive caused by contamination. See main text for details.
sequences of the genes in Tables 1 and 2 can be found in Additional file 2.

Of the 17 genes from the original human genome paper that Crisp et al. [1] claim are true examples of HGT, my analysis finds that 12 genes fail to pass the authors’ own BLAST-based test for HGT, because their closest metazoan match has a bitscore that is greater than the best non-metazoan match (Table 1). Of the 28 genes representing new claims of HGT (Table 2), 26 fail the initial screen for HGT candidates, either because they fail the original BLAST bitscore test, because they represent contaminants in draft genomes, or because they are known mitochondrial or retrotransposed genes. The remaining seven genes (five from Table 1 and two from Table 2) include three close paralogs (HAS1–3) and thus represent four hypothesized HGT events. A combination of gene loss and evolutionary rate variation is more than adequate to explain these genes: among other reasons, the alignments and bitscores are the result of screening more than 20,000 human genes, and one might expect a few genes from this large set to be lost (or to have evolved slightly more rapidly) in the non-chordate genomes.

One reason that better BLAST results were found in the current study could well be that this study used data from May 2016, whereas Crisp et al.’s study used data from January 2013. A large number of additional genomes have been deposited in public archives during the three years between the two analyses. These species were not available to the previous study and thus the orthologous genes from these taxa were missed. Insofar as this explanation is correct, it strengthens the argument for gene loss as the explanation for the (very few) human genes that still have better BLAST matches in non-metazoans than in non-chordate metazoans.

Another factor is that because only non-chordates are considered, the alignments and bitscores between a human gene and these very distant relatives are necessarily quite weak. This distant relationship makes it more likely that some genes will not be found simply because the sequence has diverged too much for a pairwise alignment to detect it.

This study focuses only on human genes, but recent claims of high levels of HGT in other animals have also been reported. The most dramatic claim was the recent report that up to one-sixth of the genes in the tardigrade (Hypsibius dujardini) had been laterally transferred from other species [6], but that claim was quickly shown to be a false result due primarily to contamination of the genome assembly [7]. In Crisp et al. [1], contamination seems to be a likely explanation for the three human genes (PRAME family members 1, 6, and 15) reported as high-confidence HGT events, and a closer scrutiny of other automatically identified HGT candidates might reveal other cases. (Contamination has been reported to create false signals of HGT as far back as 2002 [8].) My re-examination here suggests that HGT is very rare rather than widespread in vertebrate genomes, and that every hypothesized HGT event needs to be subjected to careful scrutiny.

As we wrote in 2001 [3], “the argument for lateral gene transfer is essentially a statistical one, necessarily so because of the inherent impossibility of observing events that may have occurred in the distant past”. When searching a large set of genes against an even larger database, one must recognize that such large-scale, automated searches will inevitably find unusual results that include genes that were lost or evolved more rapidly in multiple lineages. Because HGT is such an unlikely event, the results of automated searches should be subjected to individual, close scrutiny with an eye toward explaining them through more mundane processes before concluding that these anomalies represent novel biological discoveries. As demonstrated here, a reanalysis using the latest genome databases shows that other than the well-known mitochondrial genome transfer and retrovirus-mediated events, no genes have been horizontally transferred into the human genome.

Methods
Ensembl identifiers for all genes proposed as examples of HGT were obtained from Crisp et al. [1] and validated by retrieving them from the Ensembl database (www.ensembl.org). Genomes and protein sequences were obtained from the National Center for Biotechnology Information (NCBI; www.ncbi.nlm.nih.gov) and UniProt (www.uniprot.org). Protein sequences were aligned individually using the blastp program and the non-redundant protein database, nr, available through the BLAST server at NCBI (https://blast.ncbi.nlm.nih.gov) or for direct download from the same source. To aid analysis, searches were run against the entire database and again with the phylum Chordata (taxon 7711) excluded from the results, which did not affect bitscores.

Additional files

| Additional file 1: | Gene-by-gene analysis of evidence for horizontal gene transfer for all genes in Tables 1 and 2. (DOCX 48 kb) |
| Additional file 2: | Protein sequences for Ensembl genes in Tables 1 and 2. (FA 24 kb) |

Abbreviations
BVT: Bacterial-to-vertebrate transfer; HGT: Horizontal gene transfer; NCBI: National Center for Biotechnology Information

Acknowledgements
This work was supported in part by the US National Institutes of Health under grants R01-GM083873 and R01-HG006677.
Competing interests
The author declares that he has no competing interests.

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Published online: 08 May 2017

References
1. Crisp A, Boschetti C, Perry M, Tunnacliffe A, Micklem G. Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. Genome Biol. 2015;16:50.
2. The International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.
3. Salzberg SL, White O, Peterson J, Eisen JA. Microbial genes in the human genome: lateral transfer or gene loss? Science. 2001;292:1903–6.
4. Stanhope MJ, Lupaš A, Italia MJ, Koreteke KK, Volker C, Brown JR. Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. Nature. 2001;411:940–4.
5. Genereux DP, Logsdon Jr JM. Much ado about bacteria-to-vertebrate lateral gene transfer. Trends Genet. 2003;19:191–5.
6. Boothby TC, Tenlen JR, Smith FW, Wang JR, Patanella KA, Osborne Nishimura E, et al. Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. Proc Natl Acad Sci U S A. 2015;112:15976–81.
7. Koutsovoulos G, Kumar S, Laetsch DR, Stevens L, Daub J, Conlon C, et al. No evidence for extensive horizontal gene transfer in the genome of the tardigrade Hypsibius dujardini. Proc Natl Acad Sci U S A. 2016;113:5053–8.
8. Willerslev E, Mourier T, Hansen A, Christensen B, Barnes L, Salzberg S. Contamination in the draft of the human genome masquerades as lateral gene transfer. DNA Seq. 2002;13:75–6.