Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes

Alastair Crisp[1†], **Chiara Boschetti**[1†], **Malcolm Perry**[123], **Alan Tunnacliffe**[1*] and **Gos Micklem**[23*]

- *Corresponding authors: Alan Tunnacliffe at10004@cam.ac.uk - Gos Micklem gm263@cam.ac.uk
- † Equal contributors

Author Affiliations

[1]Department of Chemical Engineering and Biotechnology, University of Cambridge, New Museums Site, Pembroke Street, Cambridge CB2 3RA, UK

[2]Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK

[3]Cambridge Systems Biology Centre, University of Cambridge, Cambridge CB2 1QR, UK

For all author emails, please log on.

Genome Biology 2015, **16**:50 doi:10.1186/s13059-015-0607-3

Alastair Crisp and Chiara Boschetti contributed equally to this work.

The electronic version of this article is the complete one and can be found online at:http://genomebiology.com/2015/16/1/50

Received: 25 September 2014
Accepted: 4 February 2015
Published: 13 March 2015

Abstract

Background

A fundamental concept in biology is that heritable material, DNA, is passed from parent to offspring, a process called vertical gene transfer. An alternative mechanism of gene acquisition is through horizontal gene transfer (HGT), which involves movement of genetic material between different species. HGT is well-known in single-celled organisms such as bacteria, but its existence in higher organisms, including animals, is less well established, and is controversial in humans.

Results

We have taken advantage of the recent availability of a sufficient number of high-quality genomes and associated transcriptomes to carry out a detailed examination of HGT in 26 animal species (10 primates, 12 flies and four nematodes) and a simplified analysis in a further 14 vertebrates. Genome-wide comparative and phylogenetic analyses show that HGT in animals typically gives rise to tens or hundreds of active ‘foreign’ genes, largely concerned with metabolism. Our analyses suggest that while fruit flies and nematodes have continued to acquire foreign genes throughout their evolution, humans and other primates have gained relatively few since their common ancestor. We also resolve the controversy surrounding previous evidence of HGT in humans and provide at least 33 new examples of horizontally acquired genes.

Conclusions

We argue that HGT has occurred, and continues to occur, on a previously unsuspected scale in metazoans and is likely to have contributed to biochemical diversification during animal evolution.

Background

The acquisition of genes from an organism other than a direct ancestor (that is, horizontal gene transfer (HGT) also called lateral gene transfer) is well known in bacteria and unicellular eukaryotes, where it plays an important role in evolution [1], with recent estimates suggesting that on average 81% of prokaryotic genes have been involved in HGT at some point [2]. However, relatively few cases have been documented in multicellular organisms [3]-[7]. Reports of HGT in animals are usually limited to the description of the transfer of only one or a few genes, making the extent of horizontal gene transfer in animals unclear. Examples include the transfer of fungal genes for carotenoid biosynthesis to the pea aphid, which results in a red pigmentation and is thought to be beneficial to the aphid [8] and the transfer of a cysteine synthase from a bacterium into the arthropod lineage (likely two independent transfers into a phytophagous mite ancestor and a lepidopteran ancestor), which allows the detoxification of cyanide produced by host plants [9]. This activity is also found in nematodes, where it may have been acquired by HGT from plants [9]. Other examples of putatively adaptive HGT have been characterised in plant-parasitic nematodes, which produce cell-wall degrading enzymes from a number of horizontally transferred genes [3], and the coffee berry borer beetle, where a mannanase has been transferred from bacteria allowing the hydrolysation of coffee berry galactomannan [10].

In exceptional cases, high levels of HGT in animals have been reported, but this has been attributed to the lifestyles of the recipient organisms. For example, in bdelloid rotifers, which are desiccation-tolerant asexuals, up to approximately 10% of transcripts derive from horizontally acquired genes [11]-[13]. Desiccation results in both DNA breakage [14],[15] and loss of membrane integrity (reviewed in [16]), both of which may potentiate HGT. Another unusual example is the transfer of the entire genome (>1 Mb) of the bacterium *Wolbachia* into the fruit fly*Drosophila ananassae*, although relatively few *Wolbachia* genes are transcribed in this case [17]. Genes from *Wolbachia* are frequently transferred to invertebrates [17],[18], probably because the long-term association (either parasitic or mutualistic) between the bacterium and its hosts maintains their genomes in close proximity. Furthermore, as *Wolbachia* frequently infects the testes and ovaries of its hosts, it has access to their germlines, a prerequisite for the transmission of the acquired genes to the next generation. These studies have led to the perception that HGT occurs very infrequently in most animals, especially in vertebrates [5],[6]. Furthermore, there are concerns over the validity of the examples of HGT reported in humans [19]-[22]. The original report on the human genome sequence [19] described prokaryote-to-vertebrate HGT discovered by aligning human sequences to those of a small number of species (not many genomes were available at the time), including only two metazoans, *D. melanogaster* and *Caenorhabditis elegans*. Any proteins aligning to bacteria but not to these two metazoans, or to the other two eukaryotic proteomes used (*Arabidopsis thaliana* and *Saccharomyces cerevisiae*), were considered to be a result of prokaryote-to-vertebrate HGT. However, these four eukaryotic species do not contain orthologs of all ‘native’ human genes (that is, those not horizontally acquired), leading to incorrect identification of HGT (false positives) and the subsequent rejection of many cases by phylogenetic analyses[20]-[22]. The problem (the availability of a limited number of eukaryotic genomes for comparison in studies of HGT) has lessened in the intervening decade; thousands of proteomes (including several primates) are now available in UniProt, allowing prediction of HGT using alignment to hundreds of species and subsequent phylogenetic validation, as shown in recent work in invertebrates (for example, [12],[23],[24]). In the human, however, there have been no follow-up studies since the original genome paper, and the true scale of HGT in humans, and metazoans generally, remains unclear.

To remedy this, we initially identified non-metazoan to metazoan HGT in multiple *Drosophila*,*Caenohabditis* and primate (including human) species. Due to the controversy surrounding the human studies [19]-[22], we then took our analysis a step further by comparing multiple closely related species and combining information on horizontally transferred (‘foreign’) genes found in more than one species in the group, thereby reducing mis-identification of HGT caused by spurious alignments. In this way, we identified up to hundreds of active foreign genes in animals, including humans, suggesting that HGT provides important contributions to metazoan evolution.

Results

Drosophila species, *Caenorhabditis* species and primates have up to hundreds of active foreign genes

To determine the scale of HGT across well-characterised taxonomic groups, we examined 12*Drosophila* species, four *Caenorhabditis* species and 10 primates (Figure 1) for which high quality genomes and transcriptomes are available. For each transcribed gene, we calculated the HGT index, *h* (the difference between the bitscores of the best non-metazoan and the best metazoan matches), which gives a relative quantitative measure of how well a given gene aligns to non-metazoan versus metazoan sequences, with positive numbers indicating a better alignment to non-metazoan sequences [12]. For example, the *C. elegans* gene gut-obstructed 1 (*gob-1*), which encodes a trehalose-6-phosphate phosphatase, has a best non-metazoan match with a bitscore of 135 and a best metazoan match with a bitscore of 39.3 resulting in an HGT index of 95.7. As we were interested in more than just very recent HGT, we excluded members of the test species' phylum from the metazoan matches. This allowed us to identify HGT over evolutionary periods encompassing hundreds of millions of years, as opposed to only identifying HGT that occurred since the test species' divergence from its most closely related species (likely up to tens of millions of years). Hereafter, when we refer to matches to metazoan sequences, we mean these subsets.

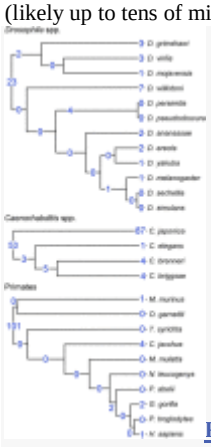

Figure 1. Phylogenetic relationships of the main taxonomic groups studied. The blue numbers indicate the ortholog groups mapping to each branch (HGT events). Events may have occurred anywhere along the branch, not just where the number is indicated. Events found at the base of the tree have occurred anywhere between the origin of the phylum and the base of the tree. Trees are not drawn to scale with each other.

We first identified a base level of HGT (called class C) by using conservative thresholds of $h \geq 30$ (as in [12]) (meaning that the gene aligns much better, and is therefore much more similar, to non-metazoan genes) and bitscore of best non-metazoan match ≥100 (thereby excluding bad alignments to non-metazoans). The example given above (*gob-1*) passes these thresholds and is therefore at least class C HGT. This per-species information was then combined for each taxon (*Drosophila*, *Caenorhabditis* and primates) to construct ortholog groups. For each ortholog group we calculated the average *h* value of all members (h_{orth}) and defined the genes with $h_{orth} \geq 30$ as class B, a subset of class C. These genes are, on average, predicted as HGT in all tested species they are found in. The gene *gob-1* has homologs in *C. brenneri*, *C. briggsae* and *C. japonica*, with values of $h = 102$, $h = 97.1$ and $h = 86.4$ respectively, giving an average *h* (h_{orth}) of 95.3 and as such *gob-1* (ands its homologs) are also class B HGT. Finally, we applied a still more stringent filter to define class A foreign genes (a subset of class B), which had only very poor alignments to metazoan sequences and whose orthologs, as used to define class B, also had similarly poor alignments to metazoan sequences. To do this, we identified those sequences whose best match to a metazoan had a bitscore <100 and whose ortholog groups contain no genes with metazoan matches of bitscore ≥100 (Figure 2A). The gene *gob-1* has no metazoan matches with bitscore ≥100 (best metazoan match = 39.3) and the same is true for its homologs (best matches of 37, 38.9 and 36.6, respectively), as such it is also class A HGT.

Figure 2. HGT genes by class. (A) The left panel shows a schematic representation of the HGT classes: class B and C genes have *h* index ≥ 30 and bitscore of the best non-metazoan blastx hit ≥ 100 (they are distinguished by h_{orth}, which is not shown on this figure), while class A genes must additionally have bitscore <100 for the best metazoan blastx hit. The right panel shows the scores for all genes in *H. sapiens*, colour-coded according to their classification (class A: red, class B: orange, class C: blue, native genes: grey). **(B)** Box-plot of the number of genes in each class, for the three main taxa analysed (*Drosophila* spp. *Caenorhabditis* spp., primates species), colour-coded according to the same scheme (class A: red, class B: orange, class C: blue).

We then performed phylogenetic analyses for all genes of each of the above classes and found that an average of 55% of all class C genes, 65% of all class B genes and 88% of all class A genes were phylogenetically validated as foreign. This validation and further manual analysis (Additional files 1 and 2) suggested that, while false positives are minimised as C → B → A, some true positives are also lost. Therefore, class A genes represent a minimum estimate of the level of HGT for a given species.

We found that *Caenorhabditis* species have, on average, 173, 127 and 68 genes in HGT classes C, B and A, respectively. In contrast, *Drosophila* species have fewer active foreign genes with, on average, 40 genes in class C, 25 in class B, and only four in class A. Primate HGT levels fall between those of the invertebrate taxa, with an average of 109, 79 and 32 genes per species in classes C, B and A, respectively (Figure 2B, Additional files 2 and 3).

Identified foreign genes are unlikely to be explained by alternative hypotheses

To verify that the foreign genes we identified do indeed belong to the species under study and are not contamination (this is a problem in a number of animal genome sequences; see 'Phylogenetic validation' in Additional file 1), we tested whether they were found on the same genomic scaffolds as (that is, were linked to) genes of metazoan origin (native genes). Across all species we found an average of only nine class C genes per species (6.6% of foreign genes) that were not linked to native genes (Additional file 2), with correspondingly low proportions for class B and A genes. Demonstration of such high levels of linkage was only pos