Widespread of horizontal gene transfer events in eukaryotes

Kun Li[1$], Fazhe Yan[1$], Zhongqu Duan[1], David L. Adelson[2], Chaochun Wei[1, 3*]

[1] School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China

[2] School of Biological Sciences, The University of Adelaide, SA 5005, Australia

[3] Joint International Research Laboratory of Metabolic and Developmental Sciences, Shanghai Jiao Tong University, Shanghai 200240, China

Contact information

Chaochun Wei

Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China

Tel: (+86)21-34204083

E-mail: ccwei@sjtu.edu.cn

[$]: these authors contributed equally to this work.

Summary

Horizontal gene transfer (HGT) is the transfer of genetic material between distantly related organisms. While most genes in prokaryotes can be horizontally transferred, HGT events in eukaryotes are considered as rare, particularly in mammals. Here we report the identification of HGT regions (HGTs) in 13 model eukaryotes by comparing their genomes with 824 eukaryotic genomes. Between 4 and 358 non-redundant HGTs per species were found in the genomes of 13 model organisms, and most of these HGTs were previously unknown. The majority of the 824 eukaryotes with full length genome sequences also contain HGTs. These HGTs have transformed their host genomes with thousands of copies and have impacted hundreds, even thousands of genes. We extended this analysis to ~128,000 prokaryote and virus genomes and revealed a few potential routes of horizontal gene transfer involving blood sucking parasites, intracellular pathogens, and bacteria. Our findings revealed that HGTs are widespread in eukaryotic genomes, and HGT is a ubiquitous driver of genome evolution for eukaryotes.

Keyword: horizontal gene transfer, eukaryotes, genome comparisons, sequence composition bias

Main

Background

Horizontal gene transfer (HGT) is the transfer of genetic material between organisms that is not from parent to offspring, and it is a major driver of genome evolution in bacteria and archaea[1, 2]. On average, 81% of genes in prokaryotes were involved in HGT[3]. Recent evidence has shown that HGT events also exist in eukaryotes. For example, HGT events have been reported from soil bacteria to the common ancestor of *Zygnematophyceae* and *embryophyte*s, which increased its resistance to biotic and abiotic stresses during terrestrial adaptation[4]. Besides, HGT of a plant detoxification gene BtPMaT1, made whiteflies gain the ability to malonylate a common group of plant defense compounds[5]. Another remarkable example of HGT is a ~1.5Mb fragment of *Wolbachia spp*. DNA integrated into the pill bug *Armadillidium vulgare* genome, resulting in the creation of a new W sex chromosome[6]. HGT regions (HGTs) have been observed in genomes of five parasitic plants in the *Orobanchaceae* family[7], several unicellular pathogens[8] and blood-sucking parasites[9-11]. Although it has been proposed that only unicellular and early developmental stages of eukaryotes are vulnerable to HGT[12], some argue that HGT events in eukaryotes may be limited to those derived from endosymbiotic organelles[13, 14]. Mechanisms for the transfer of DNA into eukaryotic genomes have been described for viral infection, transposons, conjugation between bacteria and eukaryotes, or from endosymbionts (not only plastids and mitochondria)[15]. Some behaviors, such as predation, and life-styles, such as parasitism, have been reported to promote DNA transfer in eukaryotes[16]. Recently, more eukaryote HGTs were reported. For example, a plant detoxification gene BtPMaT1 was found transferred to whitefly *Bemisia tabaci* which greatly expanded the insect’s food spectrum[17]. Therefore, while the prevalence of HGT may be rare in eukaryotes, compared to bacteria and archaea, it does occur. However, the scale and impact of HGTs in eukaryotes are unknown.

We present here a fast identification method for HGTs in eukaryotes using both sequence composition bias and genome comparisons (Fig S1; see Methods)，and we evaluated the method using a simulated dataset. We applied this to 13 model organisms with high quality genomes, and then expanded it to 824 eukaryotes with available full length genome sequences. Many bacteria and virus genomes were also compared.

Results

A Fast HGT identification method and evaluation of the method

We created a fast identification pipeline for HGTs in eukaryotes by combining sequence composition filtering and genome sequence comparison (Fig S1; see Methods). In brief, we first identified genomic regions most different from the rest of the genome based on their k-mer frequencies and then we aligned these selected genomics regions to other eukaryotic genomes. The genomic regions were considered as candidate HGT regions if their sequence conservation levels were discordant to the phylogenetic tree of relevant species. Specifically, genomic fragments were determined as HGT sequence if they had high identity percentage with species within a distantly related group (DRG) but were missing in most species in its closely related group (CRG). Finally, each putative HGT region was used to search for homologous sequences in bacteria and viruses which may be the medium vectors of HGT events.

The pipeline was evaluated with HGT sequences previously reported as HGT regions in the human genome[18]. We tried different kmer sizes (1~6), and k=4 was selected because the highest portion of candidate HGT sequences previously reported in the human genome[18] were kept (Figure S2). A very high portion (>75%) of these human HGTs reported previously were kept in the result HGT sequences even if we only input top 5% of the fragments with highest differences to the human genome (Figure S2).

We further evaluated the pipeline with a genome containing simulated HGT regions. Since our HGT identification pipeline has two main steps, sequence composition-based filtering step and genome comparison step. The evaluation was done for the two steps (Figure S3, Table S1). While top 1% fragments were input to the pipeline, 20.6% correct results would be identified after sequence composition-based filtering and 14.3% correct results identified after genome comparison. When the percentage of fragments input was up to 50%, 83.4% and 77.7% correct results were identified after two steps respectively. It can be seen that the precision of prediction was higher than 60% for all cases. This indicated that we may have underestimated the number of HGTs (low recall rate) but majority of the identified HGTs were highly reliable.

Widespread of HGTs among eukaryotes

We applied our HGT identification method to identify HGTs in Eukaryotes. We identified between 4 and 358 non-redundant eukaryotic HGTs for 13 model organisms with high quality genomes including 1 primate, 4 mammals, 4 non-mammalian vertebrates, 2 invertebrates, 1 plant and 1 fungus (Table 1). The number of HGT regions found in lizard, a non-mammalian vertebrate, was the smallest, while the number for elephant was the largest. There are more HGT regions found in 5 mammals especially in elephant, human and cow, than that in other 8 organisms. The number of HGT regions in a species was around 20, except elephant, human cow and lizard. For 824 eukaryotes with full length genome sequences currently available, almost all (98.7%) of them contained HGT regions. A number of those HGT regions were also found to have bacteria or viruses as medium vectors (Table 1, Table S2).

For the identified HGTs in the 13 model organisms (Table 1), most of them were previously unknown compared with reported HGT regions[10, 18-25]. For each candidate HGT region, a phylogenetic tree was constructed from the homologous sequences of that HGT region in all eukaryotes (see Methods). To determine the frequency of HGT in eukaryotes, we calculated an HGT-appearance number N_{AB} for a model organism A and another eukaryotic organism B, which was defined by the frequency with which organism B appeared in the phylogenetic trees of non-redundant HGTs of model organism A. For instance, among the 313 non-redundant HGT trees for *Homo sapiens*, *Pan troglodytes* was found in 312 of them, therefore the HGT-appearance number N_{HP} between *Homo sapiens* and *Pan troglodytes* was 312. The distribution of HGT-appearance numbers between the 13 model organisms and 824 eukaryotes is shown in Figure 1A and Table S3. If model organism A and organism B were from different kingdoms, N_{AB} are shown as a line in Figure 1B and Table S3. The greater the value for N_{AB}, the thicker the line. By using this metric, we determined that 98.7% of eukaryotes (813 of 824) hosted HGTs, revealing widespread of HGTs across eukaryotes. In addition, we categorized the HGTs into cross-kingdom, cross-phylum, cross-class or unknown categories based on the taxonomy relationships of the two involved organisms. We found that 1081 pairs of cross-kingdom species contained at least one HGT, and about half of them contained multiple HGTs (Figure 1B, Table S4). The number of cross-phylum and cross-class species pairs containing HGT were 1,890 and 2,909 respectively (Figure S4, Table S4).

Duplications of HGTs and their impact on their host genomes

Horizontal transferred active transposable elements may continue to transpose in the new host. Therefore, we compared the non-redundant eukaryotic HGT sequences we identified with their host genomes. Overall, about 22.2% of HGTs (242 of 1,090) have multiple copies in their host genomes, and 47 HGTs have more than 1,000 copies (Table S2). In particular, BovB related HGT region "chr8:96500648-96500854" in the cow genome has 56,890 copies (total length 11.3 Mbp), which is consistent with a previous study that BovB are present as many copies[9]. In newly identified HGTs, elephant HGT region "scaffold_90:4162401-4162729" has 13,484 copies and occupies 0.15% of the elephant genome (4.4Mbp/3.2Gb). Frog HGT region "chr1:8559133-8559400" has 7,027 copies and occupies 1.7Mb.

These HGT copies have affected many genes as well. There are 51 HGT regions, each of which impacts more than 100 protein coding genes in their host genome (Table S2). For example, the frog

HGT region mentioned above and its copies overlap (with at least 1bp) more than 10% of all protein-coding genes (2,149 of 19,983), which is a dramatic impact on the frog genome functions. HGTs with similar (but different degree of) impact on genome functions can also be found for most of the 13 model organisms. Especially cow, human, frog, elephant, zebrafish, and lizard, each of them has more than 100 genes impacted by HGTs. More information can be found in Table S2.

Repetitive sequence composition of HGTs

We compared the non-redundant HGTs detected in 13 model organisms with the repetitive sequences annotated in their reference genomes. Between 0~100% of their HGTs overlapped with interspersed repeats (excluding simple repeats) (Table 1), revealing significant species and repeat-specificity (Fig S5A). The types of repeats overlapping with HGTs showed significant correlation with overall genomic repeat composition (Fig S5B). Retrotransposons (SINEs and LINEs) were common in HGTs detected in mammals, consistent with their frequencies in their host genomes. In a frog genome (*Xenopus tropicalis*), DNA transposons, the main repeats for that genome, were frequently found in HGTs. In comparison, in the rat genome (Rattus norvegicus), the distribution of DNA transposons in HGTs was not consistent with their distribution in the host genomes. In the rat genome, DNA transposons appeared in as many as 6 non-redundant HGTs (46%), while that repeat only accounted for 3.1% of repeats in the genome.

BovB and L1 retrotransposons are the two most abundant transposable elements (TEs) in ruminant and afrotherian genomes and replicate via an RNA intermediate[26]. The horizontal transfer of BovB is known to be widespread in animals[10] and horizontal transfer of L1 has been shown in plants, animals and several fungi[19]. In total, 44 of our non-redundant HGTs overlapped with BovB retrotransposons in *Bos taurus* (Ruminantia) and *Loxodonta africana* (Afrotheria) (Table 1), supporting previous results for horizontal transfer of BovB[10, 19]. Furthermore, 461 L1 horizontal transfer events were identified in five mammals (Cow, Human, Elephant, mouse, and rat), providing more evidence that L1 elements are horizontally transferred[19]. Surprisingly, 95.2% (20 of 21) (Table S5) HGTs that overlapped with BovB retrotransposons in *Bos taurus*, were associated with the possible intermediary species, the blood-sucking parasite *Cimex lectularius* (bed bug), which has been reported by Ivancevic et al.[11]. *Cimex lectularius* is known to feed on animal blood and can host over 40 zoonotic pathogens[27], thus transmitting many infectious diseases[28].

Figure 2A showed the tree from bovine HGT region “chr25:1343971-1344200” and its homologs. In addition to the candidate vector species, this HGT tree also included 12 mammals (9 Ruminantia, 2 Metatheria and *Macaca mulatta*) and 5 non-mammalian vertebrates (2 fishes, 3 reptiles), which were clearly clustered in distinct branches (Table S6). In addition, we identified these mobile DNA sequences in several bacteria, including *Enterococcus faecium*, *Mycolicibacterium malmesburyense*, *Escherichia coli* and *Anaplasma phagocytophilum* (Table S7). Using WGS data, we confirmed high similarity homologs (sequence coverage>80%, sequence identity>90%) of this HGT region from *Cimex lectularius* (NW_014465023.1|11681076-11681736) in 10 samples (collected from PRJNA259363, PRJNA167477 and PRJNA432971, sequenced in different centers) (Table S8). Like in other bugs, it appears that *Cimex lectularius* transferred DNA between the hosts it feeds on.

Apicomplexan intracellular pathogens often participate in HGTs

A considerable number of genes of intracellular pathogens have been acquired through HGT, including Apicomplexa[8, 29]. In particular, *Toxoplasma gondii* is an obligate intracellular, apicomplexan parasite that causes the disease toxoplasmosis in a wide range of warm-blooded animals including humans[30, 31], where it has been reported to infect up to one third of the world’s population[32]. About 0.21% of *Toxoplasma gondii* protein-coding genes were acquired through HGT[33]. Our analysis identified 401 HGTs from 11 model organisms and 25 apicomplexans (Table 1). *Toxoplasma gondii ME49*, *Plasmodium vivax* and *Plasmodium knowlesi strain H* appeared more frequently in HGTs (Fig 3A).

Toxoplasma gondii ME49 participated in 218 human HGTs correlated with Apicomplexan intracellular pathogens, making these cross-kingdom HGTs (Fig 3B). For instance, the HGT tree of HGT region “chr11:24184801-24185043” shown in Figure 2B includes 1 Apicomplexan pathogen, 2 invertebrates and 2 non-mammalian vertebrates (Table S6). This HGT tree is inconsistent with the phylogenic tree of these organisms, and this HGT was also found in 36 bacterial strains, indicating that these same DNA sequences were able to jump into bacteria as well as eukaryotes. The apicomplexan pathogen (*Toxoplasma gondii ME49*) and a blood-sucking parasite (*Ixodes scapularis*) are good candidate sources/vectors for DNA transfer into the human genome. Several primates including *Homo sapiens*, *Gorilla gorilla gorilla*, *Pan troglodytes*, *Pongo abelii* and *Pan* *paniscus* were clustered into a branch in the HGT tree, indicating that this DNA transfer event happened in their common ancestor. Using WGS data, we successfully confirmed homologous sequences in *Toxoplasma gondii ME49*, which further supported this DNA transfer event.

Discussion

HGTs are widespread in eukaryotes (in the 13 model organisms we examined in this study and 98.7% of other eukaryotes with whole genome sequences). Compared to HGTs in prokaryotes, the number of non-redundant eukaryotic HGTs (4~358 regions) detected in these model organisms was very small. In addition, we found many HGT regions by comparing a small part of the genome sequences that were significantly different from their reference genomes. It is conceivable that the number of HGT regions is much larger than this.

As shown in Figure 1A, the HGT-appearance numbers decreased when the phylogenetic distance between two organisms increased. For example, primates appeared in most HGT trees for human, followed by mammals. Most primates appeared in almost all HGT trees of *Homo sapiens*, indicating that most these HGT sequences were inserted into the genome of their common ancestor. We observed a similar distribution of HGT-appearance numbers in other model organisms, indicating that most HGT regions identified by our pipeline were transferred before the divergence of model organisms and their sibling lineages. This also implied that these HGTs may have important functions as they have persisted[1].

For mammals (human, mouse, rat, cow, and elephant), we investigated the geographic distribution of the two organisms involved for each HGT event. Most of the organism pairs were from the same continent. For the 259 species related to HGTs that occurred to the common ancestor of mammals, 213 (82.2%) species were located in the same continent as the corresponding model organism, 43 (16.6%) species were not, and 3 (1.2%) species were undetermined (Table S9). The continents began to separate about 200 Mya, around the same time that the oldest mammals emerged[34, 35]. For 371 species related to HGTs that occurred into the common ancestors of the orders of the model organisms, 357(96.2%) of them were found in the same continent with the corresponding model organisms and 14 (3.8%) species were not, which were all related to HGTs of the elephant (Table S9). Proboscidea, the order to which elephants belong, originated 55 Mya[36], significantly later than the time that the continents separated.

Our study uncovered several putative routes for the exchange of genetic material between distantly related eukaryotes. We propose that blood-sucking parasites (like *Cimex lectularius* and *Ixodes scapularis*) and intracellular pathogens (like *Toxoplasma gondii ME49*) were involved in DNA transfer between mammals and other eukaryotes and these transferred DNA sequences were also found in pathogenic bacteria, suggesting exchange of genetic material between eukaryotes and bacteria (Table S10). In this fashion, bacteria might serve as the vector for DNA transfer between distantly related and eukaryotes that might not be in close contact with each other. We also found highly similar homologous sequences in viral genomes for three HGTs in human, indicating that viruses might be agents for integration of transferred DNAs in to eukaryotic genomes. Taken together, these findings revealed a putative route for DNA transfer between distantly related eukaryotes (Figure 4). Nevertheless, we observed that about 54.4% of HGTs events could be interpreted by bacteria medium (Table 1). However, the detailed routes for DNA transfer for the majority of HGT regions in this report are still unclear. With the progress of sequencing technology, especially third generation sequencing technologies, high quality whole genome sequences can be obtained for several HGT related species distributed across the tree of life, and this will provide a good opportunity to determine the route and direction of HGT.

Functional annotation for genes overlapping with HGTs (see Methods) revealed some significantly enriched Gene Ontology terms (GO terms) (Bonferroni<0.05) for protein-coding genes from mouse, fruit fly and nematode as well as non-coding genes from yeast. (Table S11). The significant GO terms for nematode were “hemidesmosome, intermediate filament”, while the significant GO term for mouse was “protein kinase A binding”. HGTs in fruit fly that overlapped with coding genes were enriched for “ATP binding, lipid particle, microtubule associated complex”, etc. HGTs in yeast overlapped with non-coding genes enriched for “retrotransposon nucleocapsid, transposition, RNA-mediated, cytosolic large ribosomal subunit”, etc.

In conclusion, comparison of 13 model eukaryote genomes against other organisms with whole available genome sequences showed that HGT is widespread in eukaryotes. We suggest that blood-sucking parasites, apicomplexan pathogens, bacteria, and viruses are nodes in the putative routes for DNA transfer between distantly related eukaryotes.

Tables

Table 1. The numbers of non-redundant HGTs in 13 model organisms. Most of these HGTs were novel. Some of these HGTs are supported by genomic evidence that they were mediated by bacteria, viruses, or apicomplexan pathogens. The numbers of HGTs overlapping with repeats, including well-known TEs, such as BovB and L1, are shown in the last two columns.

Species	Genome version	HGTs	Novel HGTs	References of known HGTs	With medium organisms			Overlapped with repeats	With TEs	
					Bacteria	Viruses	Apicomplexa		BovB	L1
Human	hg38	313	152	[18, 22]	278	159	268	313	0	117
Mouse	mm10	15	15	[21]	10	0	4	10	0	8
Rat	rn6	13	12	[21]	9	0	2	13	0	7
Cow	bosTau7	84	74	[10]	69	0	43	82	21	56
Elephant	loxAfr3	358	358	[10]	148	0	66	317	23	273
Chicken	galGal4	13	13	/	1	0	0	1	0	0
Lizard	anoCar2	4	4	[21, 24]	1	0	1	1	0	0
Frog	xenTro9	17	17	[21, 24]	0	0	0	17	0	0
Zebrafish	danRer10	25	25	[20]	2	0	1	21	4	0
Fruit fly	dm6	177	177	[22]	45	1	7	10	0	0
Nematode	ce11	22	21	[22]	1	0	1	0	0	0
Arabidopsis	tair10	22	22	[25]	16	0	3	0	0	0
Yeast	sacCer3	27	25	[23]	12	0	5	0	0	0

Figures

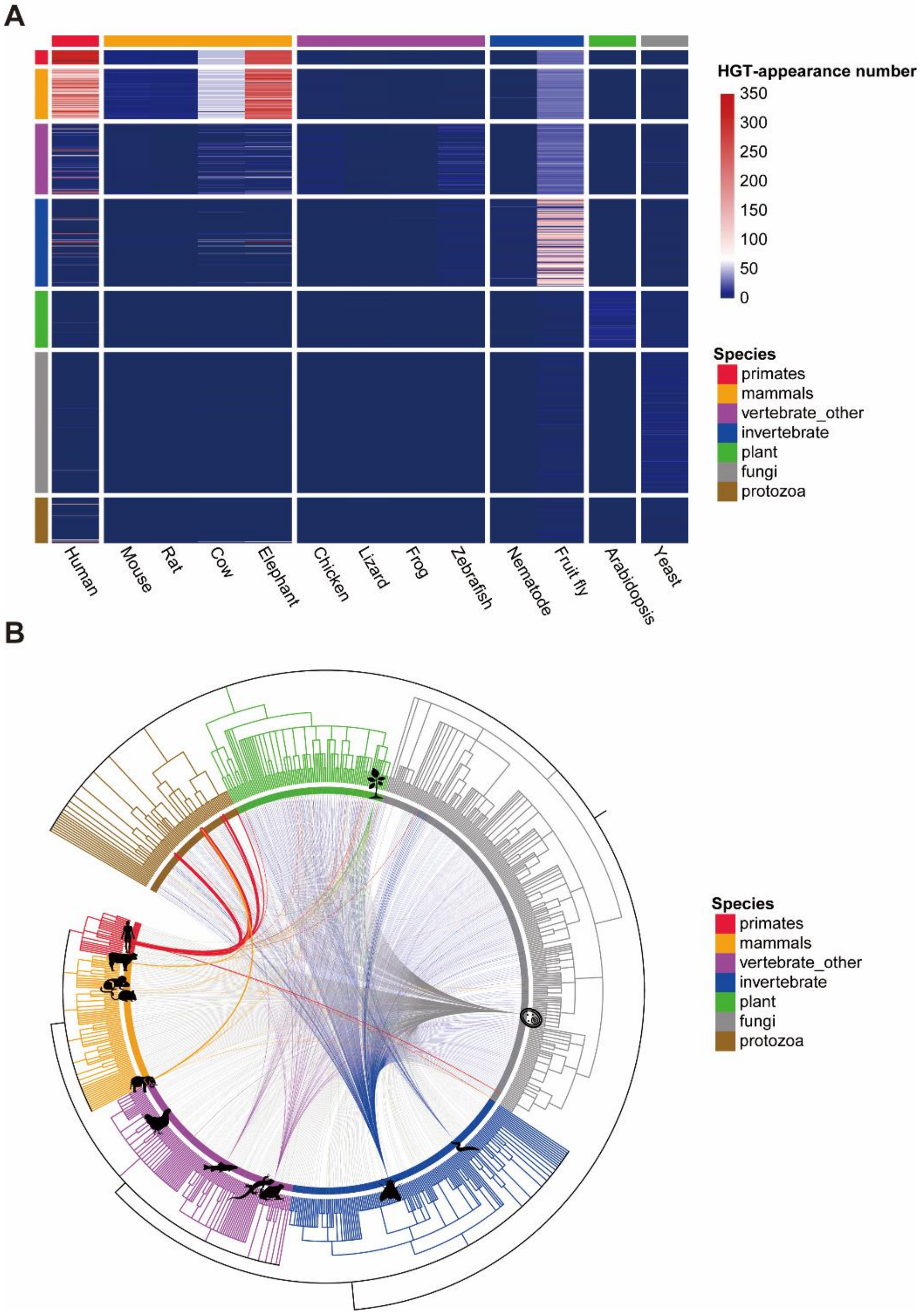

Figure 1. HGTs among eukaryotes. All 824 eukaryotes were clustered into seven sub-groups: primates, non-primate mammals, non-mammal vertebrates, invertebrates, protozoa, fungi, and plants. **(A)** The HGT-appearance numbers between 13 model organisms (X axis) and 824 eukaryotes (Y axis) are represented by the grid colors in the heatmap. (B) Cross-kingdom HGTs were shown by the lines connecting related species, and the thickness of the line represented the HGT-appearance number of the related species. Cross-phylum and cross-class HGTs are shown in Figure S2.

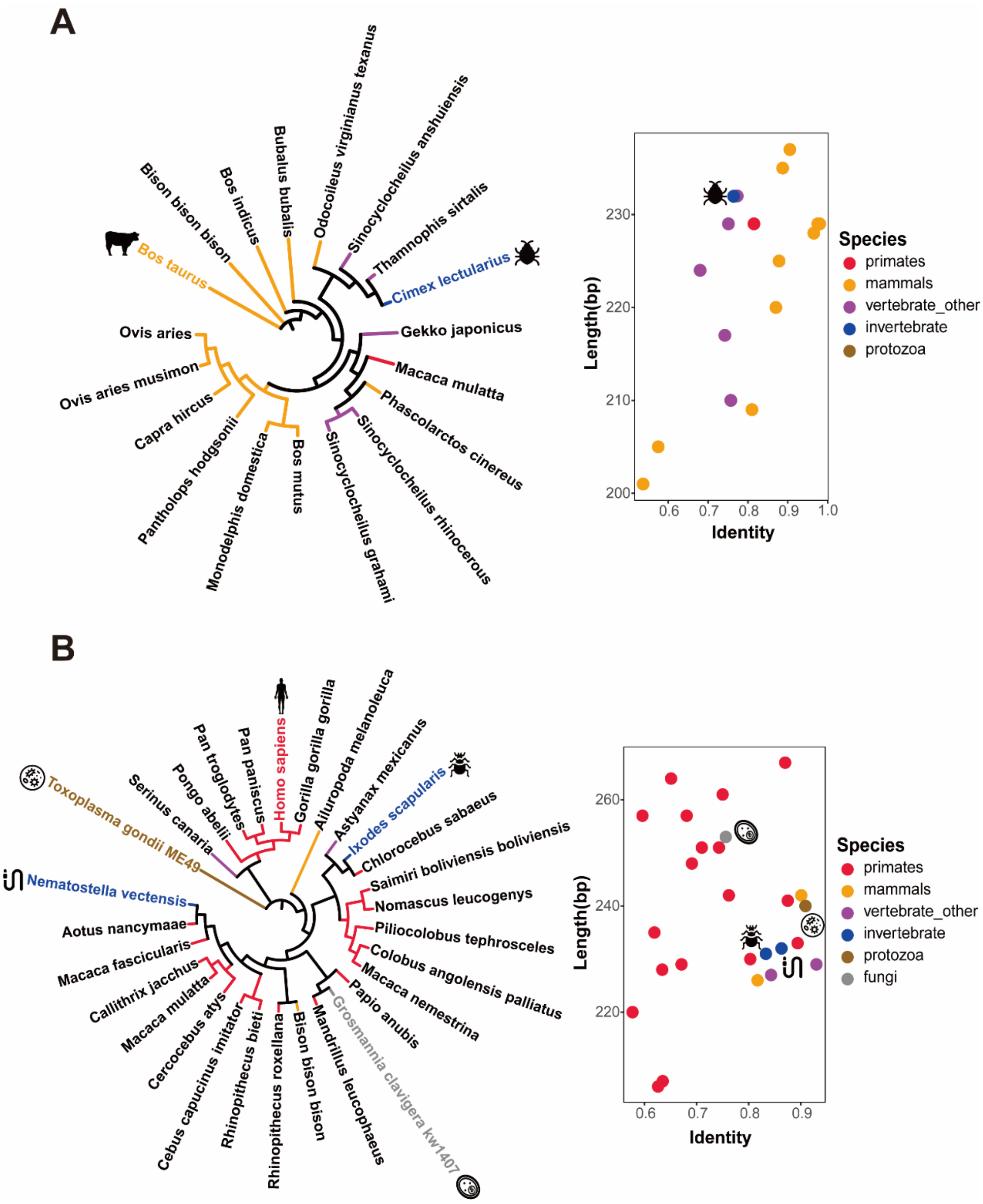

Figure 2. Phylogenic trees and length-vs-identity plots of HGT region examples. (A) *Bos taurus* HGT region “chr25:1343971-1344200”; and (B) *Homo sapiens* HGT region “chr11:24184801-24185043”. The trees on left side represent the evolutionary relationship of species linked by this HGT region, and the plots present sequence similarity between the homologous sequences from the model organism and the related species.

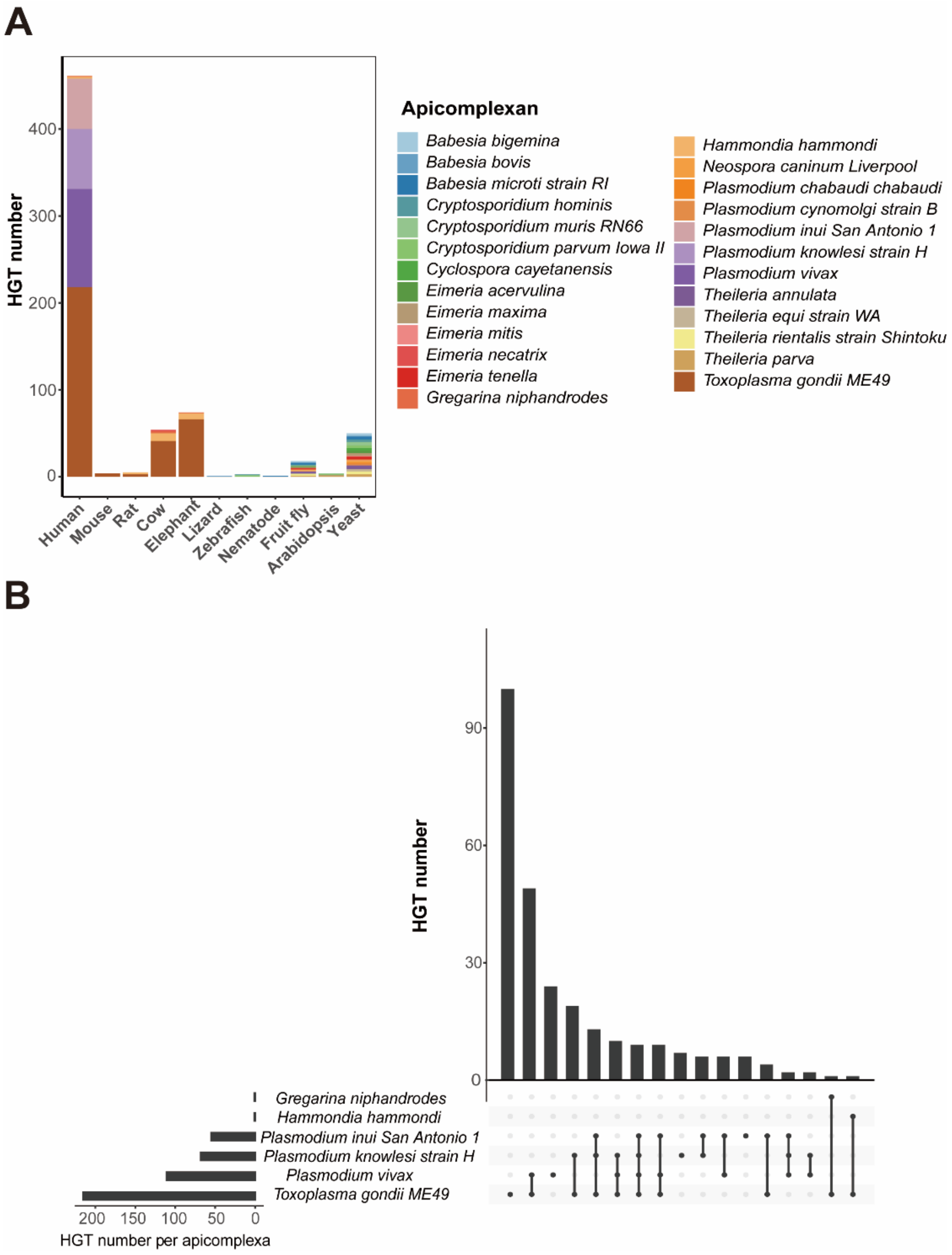

Figure 3. Apicomplexan related HGTs. **(A)** The numbers of HGTs associated with apicomplexans in different model organisms. The X-axis represented 11 different model organisms and the Y-axis represents the number of corresponding HGTs while different colors correspond to apicomplexan species. Some HGT sequences from different apicomplexan may overlap. **(B)** Detailed information about apicomplexan related HGT regions in human. The X-axis represented different combination of apicomplexans and the Y-axis represents the numbers of corresponding HGTs in the human genome.

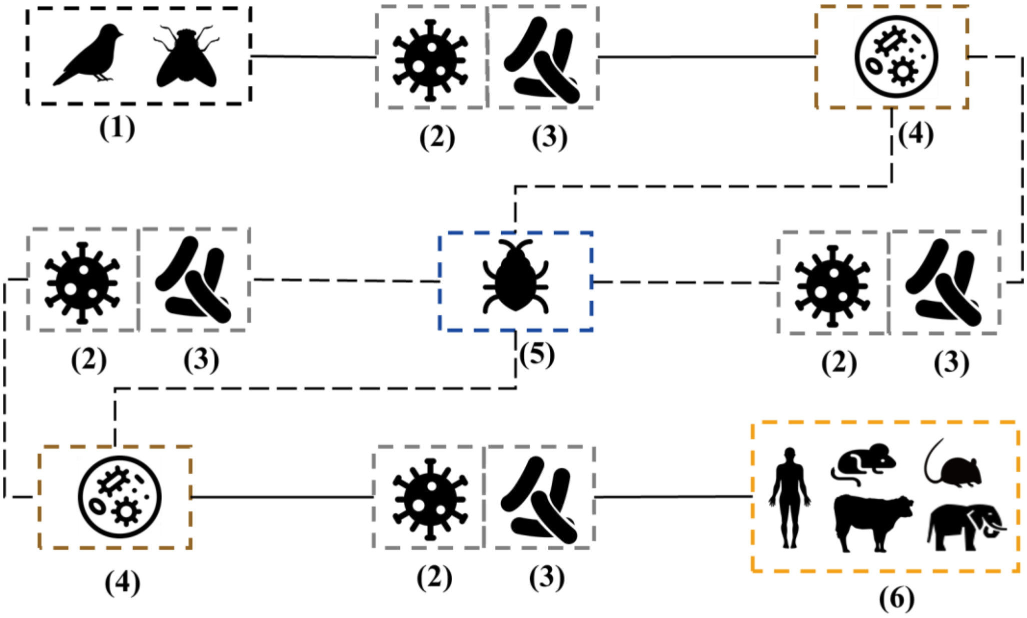

Figure 4. Putative route of horizontal gene transfer between mammals and distantly related eukaryotes. Here, boxes represent the species that participate in the DNA transfer: (1) distantly related eukaryotes, such as *Drosophila willistoni* and *Serinus canaria*; (2) viral gene pool; (3) bacterial gene pool; (4) intracellular parasites, like *Toxoplasma gondii ME49*; (5) blood-sucking parasites, like *Cimex lectularius*; and (6) mammals, including *Homo sapiens*, *Bos taurus*, *Loxodonta Africana*, *Mus musculus*, *Rattus norvegicus*. In this flowchart, solid lines stand for those well supported HGT events in this study, and dashed lines indicate untested hypothesis.

Methods

In bacterial genomes, HGT regions are also called genomic islands (GIs) and can be detected using two distinct bioinformatic approaches, based on sequence composition or comparative genomics[37]. In general, the sequence composition of GIs is significantly different from that of the recipient genome. Composition-based methods identify GIs within genome sequences by calculating the k-mer frequencies of a fragment and comparing that frequency distribution with that obtained from the whole genome. Comparative genomics approaches are based on the premise that DNA sequence based phylogenetic tree topology of GIs will be discordant with respect to known species relationships, where sequences that are absent in several closely related organisms appear in more distant species. These two methods can be adapted to the identification of HGTs in eukaryotes but not without challenges. Due to the large sizes and the high heterogeneity of eukaryotic genomes, composition-based approaches may produce a number of false-positive predictions while comparative genomic methods are computationally expensive and time-consuming when hundreds of reference genomes must be aligned. In this study, we identified HGTs between eukaryotes by combining these two approaches to reduce both the false-positive rate and computational cost.

Data collection

Three datasets were downloaded from UCSC Genome Browser[38] (http://hgdownload.soe.ucsc.edu/downloads.html) and NCBI Refseq database[39] (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq). The first dataset contained the reference genome sequences of 13 model organisms consisting of 5 mammals, 4 non-mammalian vertebrates, 2 invertebrates, 1 fungus and 1 plant. The second dataset, which was used to perform large-scale genomic comparison between model organisms with other species, contained 824 assemblies of eukaryotes including 114 mammals, 125 non-mammalian vertebrates, 155 invertebrates, 81 protozoa, 100 plants and 249 fungi. The third dataset, which contained assembled genomes of 120,838 bacteria and 7,539 viruses, was created to search the putative intermediary gene pool of DNA during transfer between eukaryotes. Detailed information about these genomes can be found in Table S12.

Pipeline to identify HGTs

Figure S1 shows the pipeline to identify HGTs. Firstly, we identified genomic regions distinguishable from the rest of the genome based on k-mer frequencies (see the next sessions for more details about this genomic fragment filtering step). The selected genomics regions were then aligned with other eukaryotic genomes using LASTZ (version 1.04.00)[40] (Supplementary Data 1). A genomic region was considered as a candidate HGT if its sequence level conservation was discordant to its species phylogenetic tree. Specifically, genomic fragments were detected with high identity percentage in species within a distantly related group (DRG) but were missing in most species in a closely related group (CRG). We then clustered the HGT sequences to obtain non-redundant HGTs. Phylogenetic trees for related species and homologous sequences were built for each HGT related species and homologous sequences of candidate HGT sequences. Finally, each putative HGT region was used to search for homologous sequences in bacteria and viruses (Table S7). Detailed information about each step is listed below.

Sequence composition-based genomic fragments filtering

Due to the large sizes of many reference genomes of model organisms, we first screened the potential genomic regions harboring HGT sequences. For each model organism species, we split the genome sequences into 1000-bp segments with 200-bp overlapped regions across all chromosomes. Sequence segments with Ns were left out. Four-bp kmer frequencies were obtained for the whole genome sequences as well as all genome segments. Euclidean distance was used to measure the difference between each segment and the whole genome sequence. All the distances were sorted in descending order. Finally, the fragments whose distances ranked in the top 10% for ce11, dm6, sacCer3 and tair10 due to their smaller genome sizes, top 20% for danRer10 or top 1% for the other model organisms were chosen for further analysis. We tried different kmer sizes (1~6), and k=4 was selected because the highest portion of candidate HGTs previously reported in the human genome[18] were kept (Figure S2).

Evaluation of the preliminary screening step

Using sequence composition to screen candidate HGT genomic regions was based on the hypothesis that different organisms have different sequence compositions. We tested this hypothesis with available genomes. First, the GC content of whole reference genome sequences showed taxon specific diversity across nine taxonomic clusters (Figure S6A). Second, principal component analysis (PCA) was performed on 824 eukaryotes using the 4-mer frequencies (Figure S6B). The resulting two-dimensional vectors were then used for binary classification to distinguish whether the organism was a mammal. This approach accurately predicted mammalian genomes 89.35% of the time, with only several non-mammalian vertebrates mis-predicted as mammals (Figure S6C).

PCA and binary-classification

For 824 eukaryotes, we conducted PCA using the "princomp" function in RStudio (version 3.5.0), to reduce the 4-mer frequencies into a lower-dimensional vector. Only the first two principal components, PC1 and PC2, were used as the features to distinguish different species. In the process of binary-classification to determine whether a given species was a mammal, two-dimensional vectors of 824 eukaryotes were randomly divided into a training dataset and a test dataset. The classifier was built from the training dataset using a logistic regression algorithm, which was implemented with the "glm" function in RStudio with default parameters, and the predicted result was evaluated by precision (exactly predicted species/all species in test dataset).

Genome comparison

The genome comparison was conducted using LASTZ for the filtered fragments of the model organisms and the whole genomes of other species with the following arguments: "--format=axt+ --ambiguous=iupac".

Re-screening the fragments and search for HGTs

Every organism belongs to its own kingdom, phylum, and class. For each classification level (kingdom, phylum, or class) for each model organism, the other species were separated into two groups: a closely related group (CRG) including all species in the same classification level as the model organism, and a distantly related group (DRG) including all species belonging to different classification levels. For example, when using class as the classification level and using human as the model organism, all mammals were regarded as part of the CRG, while the non-mammalian species formed the DRG. We further screened the filtered fragments based on alignment results from LASTZ, to identify regions with discordant evolutionary relationships. Fragments were regarded as putative HGTs when they had homologs in DRG species but not in the majority of CRG species (see below).

The aligned regions (ARs) of the input fragments were retrieved and used to identify putative HGTs. Firstly, we kept ARs that matched to DRG species that were longer than 200bp with a nucleotide identity percentage greater than 70%. For these ARs, we compared the alignment results for CRG species, for which the identity percentage threshold was set to 50%. An AR was considered to be present in a CRG species if it was aligned over 60% of its length. In addition, we counted the frequencies for each AR in CRG species (referred to as “CRG scale”). To reduce false positive results generated by incorrect alignments, we removed ARs that contained the character ‘N’ or whose GC percentages were less than 0.3 or greater than 0.6. Finally, we checked the repetitive regions overlapping with ARs. RepeatMasker tracks were downloaded from the UCSC Genome Browser, or we ran *de novo* RepeatMasker (version 4.0.7)[41] (http://www.repeatmasker.org) to label the repeats of ARs. We then removed ARs that overlapped with simple repeats or low complexity repeats. We also use TRF(version 4.09)[42] to remove ARs that overlapped with any random repeats.

We set M as the maximum number of species with sequences aligned in the CRG. For example, when using class as the classification level and using human as the target model organism, all mammals were regarded as the CRG, while the non-mammalian species formed the DRG. M can be set to the number of all primates which is the order humans belong to. The M values can be set to the numbers of vertebrates, mammals and primates in the genome dataset containing 824 eukaryotes (Table S12). At the same time, the DRG scales of ARs were limited such that they appeared in at least N of all the species in the DRG. In our analysis, we set N=1. The alignment threshold for the DRG, including identity and length coverage, were set much higher (see below) than for CRGs, and we removed ARs with high percentages of GC or repeat compositions. The remaining ARs were considered as candidate HGTs, and were used to build trees to determine discordance with known evolutionary relationships. The detailed parameter setting was shown in Table S13.

Identifying non-redundant HGTs

HGTs were clustered using the cd-hit-est program (version 4.6.6)[43] with minimum nucleotide identity set at 80%. The longest sequences from each cluster were selected to represent the non-redundant HGTs.

Counting the copy numbers of HGTs

We run BLASTN alignment for non-redundant HGT sequences against their host reference genomes, with the parameter "-e 1e-5". For each HGT, we selected aligned regions that covered at least 90% of HGT regions with nucleotide identity > 90%. We then merged those aligned regions with overlapped coordinates. The copy number of each HGT was determined from the number of merged HGT copies.

Exclusion of mitochondrial or chloroplast DNA

Complete mitochondrial genomes of 13 model organisms and Arabidopsis thaliana chloroplast DNA were obtained from NCBI, and we then searched with BLASTN against non-redundant HGTs. With the argument "-evalue 1e-5", we found no homologous DNA sequences in mitochondrial or chloroplast genomes.

Remove HGTs present in Endogenous viruses

Endogenous retroviruses (ERVs) are widespread in vertebrates, making up nearly 8% of the genome of *Homo sapiens*[44]. ERVs in human share sequence homology with other primate ERVs[45]. Therefore, in order to avoid reporting sequences as HGTs that are actually from ancestral inheritance, we removed all HGTs found in ERVs. We collected ERVs from the repeat annotation of the UCSC genome browser, except for *Saccharomyces cerevisiae S288C* and *Arabidopsis thaliana*. All HGTs that overlapped with ERV genomic coordinates or aligned to ERVs using BLASTN (identity>90% and length>100bp) were removed.

Comparison with reported HGTs in previous studies

We obtained reported HGTs for these model organisms from previous publications, including genomic coordinates and DNA sequences. HGTs in our study were considered novel if they did not match reported HGTs by genomic coordinates or sequence alignment (BLASTN[46], matched length>200bp and identity>80%).

Construction of HGT phylogenetic tree

For each HGT, we searched for homologous sequences in other species based on the LASTZ output. The nucleotide identity threshold of homologous sequences was set at 70% for DRG species and 50% for CRG species. When multiple regions in a species met the criteria, the best matched sequence, which had the maximal score weighted by the identity and multiplied by the alignment length, was picked to represent the homologous sequence. Based on the HGT sequence and the homologous sequences collected from other species, we ran multiple sequence alignment using muscle (version 3.8.31)[47] and then used FastTree (version 2.1.9) to build a maximum likelihood phylogenetic tree, which was visualized with iTOL (version 3.0)[48]. The homologous regions in other species and phylogenetic trees for non-redundant HGTs can be found in Table S14.

Homologous sequences in bacteria and viruses

HGTs were aligned to the assemblies of bacteria and viruses using NCBI BLAST(2.9.0+) with parameters: “-task blastn -evalue 1e-3”. Matched regions in assemblies were filtered to be longer than 200bp, with nucleotide identity greater than 60%.

Validation of homologous sequences in eukaryotic genomes with WGS datasets

Discordant HGT trees, constructed from discordant sequences from reference genomes, were the principal evidence for identifying HGTs from our pipeline. Thus, the power for detecting HGTs depended heavily on the quality of the reference genomes. Contaminating sequences from other species were the most likely sources of false positives. For model organisms, most candidate transferred DNA were also found in their sibling lineages, therefore the probability of sequencing contamination was negligible. However, the inaccurate reference genomes of other eukaryotes (such as parasites and protozoan pathogens) could cause false positive results due to sequencing contamination. For example, if an abnormal HGT tree consists of only one parasite and several primates, and the process of constructing the reference genome of this parasite was contaminated by human DNA, this DNA transfer would be an artifact. We checked for contamination artifacts in candidate transferred DNA by alignment with whole genome sequencing (WGS) raw data from species present in discordant cross-kingdom HGT trees. In total, we collected 59 species which were in different kingdoms with the target model organism, including 15 protozoa, 20 plants, 3 fungi, 11 invertebrates, and 10 vertebrates. For each of these species, we downloaded multiple WGS raw datasets (ranging from 3 samples to 201 samples) from the SRA database that were not used to construct the reference genome. In total, we obtained 1,190 WGS samples. Sequence alignment was done using Bowtie2 (version 2.2.4)[49] with default parameters. For each species, we calculated the length coverage percentage for homologous sequences (M sequences) of WGS samples (N samples), thus generating a coverage percentage matrix (M*N). Once a sequence had coverage of over 80% with at least 2 samples, it was classified as not an artifact. The results are shown in Table S15.

Functional annotation of genes influenced by HGTs

Genome annotation files (GFF or GTF format) were obtained for model organisms from Ensembl [50](http://asia.ensembl.org) and Tair[51] (https://www.arabidopsis.org), and they were used to identify protein-coding genes and non-coding genes likely to be affected by HGTs (overlapping with HGTs with at least 1bp). The Ensembl gene IDs were input to DAVID (version 6.8)[52] (https://david.ncifcrf.gov) for functional enrichment analysis. Significantly enriched Gene Ontology terms (GO terms) (Bonferroni<0.05) for these genes were shown in the results.

Evaluation of the pipeline using simulated datasets

We constructed a simulated genome (called genome H) with 175 HGTs from a set of distantly related genomes (called Genome set D) to the human genome. Genome set D has 4 cruciferous plant genomes, including *Arabidopsis thaliana*, *Brassica napus*, *Brassica oleracea var. oleracea* and *Brassica rapa*), while Genome set C contains 4 primate genomes, *Pan paniscus*, *Pan troglodytes*, *Pongo abelii* and *Gorilla gorilla gorilla*. The 175 HGTs are sequences that have high similarity with genomes in Genome set D (>90%) but have low similarity (<10%) with genomes in Genome set C, the closely related group of genomes.

Firstly, the genome comparison between genomes in Genome set D was conducted using LASTZ[40] and Multiz[53] to obtain sequences whose identity in all genomes of Genome set D were >90% and lengths >200bps. These sequences were compared with the genomes in Genome set C and the sequences having low similarity (identity <10%) were reserved. The obtained sequences were then clustered using the cd-hit-est program (version 4.6.6)[43] with minimum nucleotide identity set at 80%. The longest sequences from each cluster were selected as simulated HGTs, which were 175 in total. These 175 HGTs were then evenly divided into 10 groups according to their sequence lengths, and the copy numbers of which increased from 2^0 to 2^9 (Table S16). Eventually, 175 HGTs with different copy numbers were inserted into the human genome as genome H (Supplementary Data 2). Finally, we ran our pipeline with genome H as the target genome, genome set D as remote genome set, genome set C as closely related genome set and parameters M, N, L as 1, 1, 200 respectively. If the correct HGT region was covered more than 60% of its length by a predicted HGT region, the prediction was considered correct.

Declarations

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests

Authors' contributions

CCW conceived and designed the study. KL, FZY and CCW developed the pipeline and identified HGTs. KL, FZY and ZQD collected the datasets. KL and FZY conducted the visualization. KL, FZY,CCW and DLA wrote the manuscript. KL, FZY, CCW, ZQD and DLA revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by grants from the National Natural Science Foundation of China (32170643, 61472246 and J1210047), the National Basic Research Program of China (2013CB956103), the National High-Tech R&D Program (863) (2014AA021502), the SJTU JiRLMDS Joint Research Fund (MDS-JF-2019A07), and the Cross-Institute Research Fund of Shanghai Jiao Tong University (YG2017ZD01 and YG2015MS39). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We thank the High Performance Computing Center at Shanghai Jiao Tong University for the computation.

Data availability

All datasets, supplementary tables and an example of analysis pipeline application are listed in the webpage at http://cgm.sjtu.edu.cn/hgt (password: hgt2019passwd) (this webpage will become freely available after this paper is accepted).

Code availability

All scripts used in this study are available in GitHub at https://github.com/SJTU-CGM/HGT.git.

Supplementary Figures, Tables and Datasets

There are 8 Figures, 16 Tables and 2 datasets provided in multiple supplementary files. Descriptions about the figures, tables and datasets are listed below. Supplementary figures are listed in a separate file, while supplementary tables and datasets are accessible from the given URL listed in the data availability.

Supplementary Figure 1

The HGT identification system for model eukaryotes.

Supplementary Figure 2

The impact of parameter setting for the fast HGT selection step using k-mer frequency. The parameters are k-mer size and fragment percentage.

Supplementary Figure 3

Evaluation results of the HGT identification pipeline on the simulated dataset.

Supplementary Figure 4

Cross-phylum HGTs and cross-class HGTs.

Supplementary Figure 5

Repeat characteristics of HGT regions as well as reference genomes.

Supplementary Figure 6

Evaluation of the preliminary screening step.

Supplementary Figure 7

The number of HGTs associated with apicomplexan in different model organisms.

Supplementary Figure 8

Phylogenic trees of other HGT region examples.

Supplementary Table 1

Evaluation results of the HGT identification pipeline on the simulated dataset.

Supplementary Table 2

Detailed information of non-redundant HGTs, including genomic coordinates, bacterial presence, viral presence, copy numbers in the whole genomes and the number of their overlapping genes.

Supplementary Table 3

HGT-appearance numbers between the 13 model organisms and 824 eukaryotes.

Supplementary Table 4

The number of cross-kingdom HGTs, cross-phylum HGTs and cross-class HGTs.

Supplementary Table 5

The character of medium organisms of HGT regions overlapped with BovB in bosTau7.

Supplementary Table 6

Examples of HGTs in cow and human genomes.

Supplementary Table 7

BLASTN results of non-redundant HGTs against bacteria/viruses.

Supplementary Table 8

Coverage matrices of WGS data for HGT homologous sequences in selected eukaryotes.

Supplementary Table 9

The geographic information of species with HGTs in mammals.

Supplementary Table 10

Putative media of horizontal gene transfer between mammals and distantly related eukaryotes.

Supplementary Table 11

Functional annotation for genes affected by HGTs.

Supplementary Table 12

Information of 13 model organisms and assembly ID of other eukaryotes, bacteria, and viruses.

Supplementary Table 13

Parameter settings of the HGT identification pipeline.

Supplementary Table 14

Homologous regions in other species and phylogenetic trees for non-redundant HGTs.

Supplementary Table 15

Coverage matrices of WGS data for HGT homologous sequences in selected apicomplexan.

Supplementary Table 16

The detailed information of 175 simulated HGTs.

Supplementary Data 1

Raw output of LASTZ alignment between 13 model organisms with other eukaryotes (197GB)

URL: http://cgm.sjtu.edu.cn/hgt/data/Supplementary_Data_1.tar

Supplementary Data 2

The simulated genome (genome H) with 175 HGTs (2.9GB)

URL: http://cgm.sjtu.edu.cn/hgt/data/Supplementary_Data_2.fa

References

[1] Soucy, S.M., Huang, J. & Gogarten, J.P. 2015 Horizontal gene transfer: building the web of life. *Nature reviews. Genetics* **16**, 472-482. (doi:10.1038/nrg3962).

[2] Polz, M.F., Alm, E.J. & Hanage, W.P. 2013 Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends in genetics : TIG* **29**, 170-175. (doi:10.1016/j.tig.2012.12.006).

[3] Dagan, T., Artzy-Randrup, Y. & Martin, W. 2008 Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 10039-10044. (doi:10.1073/pnas.0800679105).

[4] Cheng, S.F., Xian, W.F., Fu, Y., Marin, B., Keller, J., Wu, T., Sun, W.J., Li, X.L., Xu, Y., Zhang, Y., et al. 2019 Genomes of Subaerial Zygnematophyceae Provide Insights into Land Plant Evolution. *Cell* **179**, 1057-1067. (doi:10.1016/j.cell.2019.10.019).

[5] Xia, J.X., Guo, Z.J., Yang, Z.Z., Han, H.L., Wang, S.L., Xu, H.F., Yang, X., Yang, F.S., Wu, Q.J., Xie, W., et al. 2021 Whitefly hijacks a plant detoxification gene that neutralizes plant toxins. *Cell* **184**, 1693-1705. (doi:10.1016/j.cell.2021.02.014).

[6] Leclercq, S., Theze, J., Chebbi, M.A., Giraud, I., Moumen, B., Ernenwein, L., Greve, P., Gilbert, C. & Cordaux, R. 2016 Birth of a W sex chromosome by horizontal transfer of Wolbachia bacterial symbiont genome. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 15036-15041. (doi:10.1073/pnas.1608979113).

[7] Kado, T. & Innan, H. 2018 Horizontal Gene Transfer in Five Parasite Plant Species in Orobanchaceae. *Genome Biology and Evolution* **10**, 3196-3210. (doi:10.1093/gbe/evy219).

[8] Lukes, J. & Husnik, F. 2018 Microsporidia: A Single Horizontal Gene Transfer Drives a Great Leap Forward. *Current Biology* **28**, R712-R715. (doi:10.1016/j.cub.2018.05.031).

[9] Gilbert, C., Schaack, S., Pace, J.K., Brindley, P.J. & Feschotte, C. 2010 A role for host-parasite interactions in the horizontal transfer of transposons across phyla. *Nature* **464**, 1347-U1344. (doi:10.1038/nature08939).

[10] Walsh, A.M., Kortschak, R.D., Gardner, M.G., Bertozzi, T. & Adelson, D.L. 2013 Widespread horizontal transfer of retrotransposons. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 1012-1016. (doi:10.1073/pnas.1205856110).

[11] Ivancevic, A.M., Kortschak, R.D., Bertozzi, T. & Adelson, D.L. 2018 Horizontal transfer of BovB and L1 retrotransposons in eukaryotes. *Genome Biol* **19**, 85. (doi:10.1186/s13059-018-1456-7).

[12] Huang, J.L. 2013 Horizontal gene transfer in eukaryotes: The weak-link model. *Bioessays* **35**, 868-875. (doi:10.1002/bies.201300007).

[13] Martin, W.F. 2017 Too Much Eukaryote LGT. *Bioessays* **39**. (doi:ARTN 1700115
10.1002/bies.201700115).

[14] Salzberg, S.L. 2017 Horizontal gene transfer is not a hallmark of the human genome. *Genome Biol* **18**, 85. (doi:10.1186/s13059-017-1214-2).

[15] Leger, M.M., Eme, L., Stairs, C.W. & Roger, A.J. 2018 Demystifying Eukaryote Lateral Gene Transfer. *Bioessays* **40**. (doi:ARTN 1700242
10.1002/bies.201700242).

[16] Keeling, P.J. & Palmer, J.D. 2008 Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics* **9**, 605-618. (doi:10.1038/nrg2386).

[17] Xia, J., Guo, Z., Yang, Z., Han, H., Wang, S., Xu, H., Yang, X., Yang, F., Wu, Q., Xie, W., et al. 2021 Whitefly hijacks a plant detoxification gene that neutralizes plant toxins. *Cell* **184**, 3588. (doi:10.1016/j.cell.2021.06.010).

[18] Huang, W., Tsai, L., Li, Y., Hua, N., Sun, C. & Wei, C. 2017 Widespread of horizontal gene transfer in the human genome. *BMC Genomics* **18**, 274. (doi:10.1186/s12864-017-3649-y).

[19] Ivancevic, A.M., Kortschak, R.D., Bertozzi, T. & Adelson, D.L. 2018 Horizontal transfer of BovB and L1 retrotransposons in eukaryotes. *Genome Biology* **19**. (doi:ARTN 85
10.1186/s13059-018-1456-7).

[20] Sun, B.F., Li, T., Xiao, J.H., Jia, L.Y., Liu, L., Zhang, P., Murphy, R.W., He, S.M. & Huang, D.W. 2015 Horizontal functional gene transfer from bacteria to fishes. *Scientific Reports* **5**. (doi:ARTN 18676
10.1038/srep18676).

[21] Pace, J.K., Gilbert, C., Clark, M.S. & Feschotte, C. 2008 Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 17023-17028. (doi:10.1073/pnas.0806548105).

[22] Crisp, A., Boschetti, C., Perry, M., Tunnacliffe, A. & Micklem, G. 2015 Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. *Genome Biology* **16**. (doi:ARTN 50
10.1186/s13059-015-0607-3).

[23] Carr, M., Bensasson, D. & Bergman, C.M. 2012 Evolutionary Genomics of Transposable Elements in Saccharomyces cerevisiae. *Plos One* **7**. (doi:ARTN e50978
10.1371/journal.pone.0050978).

[24] Novick, P., Smith, J., Ray, D. & Boissinot, S. 2010 Independent and parallel lateral transfer of DNA transposons in tetrapod genomes. *Gene* **449**, 85-94. (doi:10.1016/j.gene.2009.08.017).

[25] Ma, J.C., Wang, S.H., Zhu, X.J., Sun, G.L., Chang, G.X., Li, L.H., Hu, X.Y., Zhang, S.Z., Zhou, Y., Song, C.P., et al. 2022 Major episodes of horizontal gene transfer drove the evolution of land plants. *Mol Plant* **15**, 857-871. (doi:10.1016/j.molp.2022.02.001).

[26] Jurka, J., Kapitonov, V.V., Kohany, O. & Jurka, M.V. 2007 Repetitive sequences in complex genomes: Structure and evolution. *Annual Review of Genomics and Human Genetics* **8**, 241-259. (doi:10.1146/annurev.genom.8.080706.092416).

[27] Doggett, S.L., Dwyer, D.E., Penas, P.F. & Russell, R.C. 2012 Bed Bugs: Clinical Relevance and Control Options. *Clinical Microbiology Reviews* **25**, 164-+. (doi:10.1128/Cmr.05015-11).

[28] Goddard, J. & deShazo, R. 2009 Bed Bugs (Cimex lectularius) and Clinical Consequences of Their Bites. *Jama-Journal of the American Medical Association* **301**, 1358-1366. (doi:DOI 10.1001/jama.2009.405).

[29] Alexander, W.G., Wisecaver, J.H., Rokas, A. & Hittinger, C.T. 2016 Horizontally acquired genes in early-diverging pathogenic fungi enable the use of host nucleosides and nucleotides. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 4116-4121. (doi:10.1073/pnas.1517242113).

[30] Kim, K. & Weiss, L.M. 2004 Toxoplasma gondii: the model apicomplexan. *International Journal for Parasitology* **34**, 423-432. (doi:10.1016/j.ijpara.2003.12.009).

[31] van Helden, P.D., van Helden, L.S. & Hoal, E.G. 2013 One world, one health. *Embo Reports* **14**, 497-501. (doi:10.1038/embor.2013.61).

[32] Montoya, J.G. & Liesenfeld, O. 2004 Toxoplasmosis. *Lancet* **363**, 1965-1976. (doi:Doi 10.1016/S0140-6736(04)16412-X).

[33] Alsmark, C., Foster, P.G., Sicheritz-Ponten, T., Nakjang, S., Embley, T.M. & Hirt, R.P. 2013 Patterns of prokaryotic lateral gene transfers affecting parasitic microbial eukaryotes. *Genome Biology* **14**. (doi:ARTN R19
10.1186/gb-2013-14-2-r19).

[34] Seton, M., Muller, R.D., Zahirovic, S., Gaina, C., Torsvik, T.H., Shephard, G., Talsma, A., Gurnis, M., Turner, M., Maus, S., et al. 2012 Global continental and ocean basin reconstructions since 200 Ma. *Earth-Science Reviews* **113**, 212-270. (doi:10.1016/j.earscirev.2012.03.002).

[35] Goswami, A. 2012 A dating success story: genomes and fossils converge on placental mammal origins. *Evodevo* **3**. (doi:Artn 18
10.1186/2041-9139-3-18).

[36] Gheerbrant, E. 2009 Paleocene emergence of elephant relatives and the rapid radiation of African ungulates. *Proc Natl Acad Sci U S A* **106**, 10717-10721. (doi:10.1073/pnas.0900251106).

[37] Langille, M.G.I., Hsiao, W.W.L. & Brinkman, F.S.L. 2010 Detecting genomic islands using bioinformatics approaches. *Nature Reviews Microbiology* **8**, 372-382. (doi:10.1038/nrmicro2350).

[38] Casper, J., Zweig, A.S., Villarreal, C., Tyner, C., Speir, M.L., Rosenbloom, K.R., Raney, B.J., Lee, C.M., Lee, B.T., Karolchik, D., et al. 2018 The UCSC Genome Browser database: 2018 update. *Nucleic Acids Research* **46**, D762-D769. (doi:10.1093/nar/gkx1020).

[39] Pruitt, K.D., Tatusova, T. & Maglott, D.R. 2007 NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* **35**, D61-D65. (doi:10.1093/nar/gkl842).

[40] Harris, R.S. 2007 Improved pairwise alignment of genomic dna., The Pennsylvania State University.

[41] Price, A.L., Jones, N.C. & Pevzner, P.A. 2005 De novo identification of repeat families in large genomes. *Bioinformatics* **21**, I351-I358. (doi:10.1093/bioinformatics/bti1018).

[42] Benson, G. 1999 Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573-580. (doi:DOI 10.1093/nar/27.2.573).

[43] Li, W. & Godzik, A. 2006 Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659. (doi:10.1093/bioinformatics/btl158).

[44] Paces, J., Pavlicek, A. & Paces, V. 2002 HERVd: database of human endogenous retroviruses. *Nucleic Acids Research* **30**, 205-206. (doi:DOI 10.1093/nar/30.1.205).

[45] Johnson, W.E. 2015 Endogenous Retroviruses in the Genomics Era. *Annual Review of Virology, Vol 2* **2**, 135-159. (doi:10.1146/annurev-virology-100114-054945).

[46] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T.L. 2009 BLAST plus : architecture and applications. *Bmc Bioinformatics* **10**. (doi:10.1186/1471-2105-10-421).

[47] Edgar, R.C. 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792-1797. (doi:10.1093/nar/gkh340).

[48] Letunic, I. & Bork, P. 2016 Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Research* **44**, W242-W245. (doi:10.1093/nar/gkw290).

[49] Langmead, B. & Salzberg, S.L. 2012 Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357-U354. (doi:10.1038/Nmeth.1923).

[50] Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Giron, C.G., et al. 2018 Ensembl 2018. *Nucleic Acids Research* **46**, D754-D761. (doi:10.1093/nar/gkx1098).

[51] Poole, R.L. 2007 The TAIR database. *Methods Mol Biol* **406**, 179-212. (doi:10.1007/978-1-59745-535-0_8).

[52] Huang, D.W., Sherman, B.T. & Lempicki, R.A. 2009 Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4**, 44-57. (doi:10.1038/nprot.2008.211).

[53] Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., et al. 2004 Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14**, 708-715. (doi:10.1101/gr.1933104).