Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Microbiology

Lateral gene transfers and the origins of the eukaryote proteome: a view from microbial parasites

Robert P Hirt[1], Cecilia Alsmark[2,3] and T Martin Embley[1]

Our knowledge of the extent and functional impact of lateral gene transfer (LGT) from prokaryotes to eukaryotes, outside of endosymbiosis, is still rather limited. Here we review the recent literature, focusing mainly on microbial parasites, indicating that LGT from diverse prokaryotes has played a significant role in the evolution of a number of lineages, and by extension throughout eukaryotic evolution. As might be expected, taxonomic biases for donor prokaryotes indicate that shared habitat is a major factor driving transfers. The LGTs identified predominantly affect enzymes from metabolic pathways, but over a third of LGT are genes for putative proteins of unknown function. Finally, we discuss the difficulties in analysing LGT among eukaryotes and suggest that high-throughput methodologies integrating different approaches are needed to achieve a more global understanding of the importance of LGT in eukaryotic evolution.

Addresses
[1] Institute for Cell and Molecular Biosciences, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, NE2 4HH, UK
[2] Division of Pharmacognosy, Department of Medicinal Chemistry, Uppsala University, Biomedical Center, S-751 23 Uppsala, Sweden
[3] Department of Virology, Immunobiology and Parasitology, National Veterinary Institute, Uppsala, Sweden

Corresponding author: Hirt, Robert P (Robert.Hirt@ncl.ac.uk)

Current Opinion in Microbiology 2015, **23**:155–162

This review comes from a themed issue on **Genomics**

Edited by **Neil Hall** and **Jay CD Hinton**

For a complete overview see the Issue and the Editorial

Available online 5th December 2014

http://dx.doi.org/10.1016/j.mib.2014.11.018

Introduction

Novel genes derived from a number of processes; including gene duplications, *de novo* gene formation, and LGT; contribute to genomic and phenotypic plasticity and can drive adaptive evolution [1]. LGT in prokaryotes is recognised to play a major role in providing novel protein coding genes and contributing adaptive traits, including the archetypical resistance to antibiotics [2]. The frequency and origins of LGT among eukaryotes and its impact on their biology is still relatively poorly understood [3] but is also increasingly recognised as a significant source of novel genes [4,5]. Compared to prokaryotes identifying LGT in eukaryotes is more difficult due to the confounding effect of their (i) complex origins involving at least two prokaryotic lineages, (ii) more complex genome architecture and protein coding capacities, (iii) sparse and biased taxonomic sampling of genome sequence data and (iv) lack of phylogenetic resolution for the major eukaryotic lineages [6]. These factors, along with the intrinsic difficulties of inferring single gene phylogenies, render annotations and evolutionary inferences of eukaryotic protein coding genes often less reliable and more sensitive to sequence database taxa sampling and to different parameters of evolutionary models in bioinformatic tools [6].

Protein coding genes in eukaryote nuclear genomes are currently thought to have originated from DNA from at least two distinct prokaryotic lineages, an archaeal source, thought to represent the original host that evolved into a nucleated cell and an alpha-proteobacterial endosymbiont that eventually evolved into mitochondria [6,7]. Additional nuclear genes of bacterial origin can be identified among eukaryotes possessing plastids, derived from a cyanobacterial primary endosymbiont or from secondary/tertiary endosymbioses involving eukaryotic endosymbionts with primary/secondary plastids [7,8]. Eukaryotic nuclear genes derived from endosymbionts are defined as endosymbiotic gene transfers (EGT) [7], which for convenience we differentiate here from LGT from other sources. Mobile genetic elements, including viruses and transposable elements, can also be integrated into nuclear genomes [1,9,10]. We shall focus here on eukaryotic genes of prokaryotic origins in microbial parasites and discuss how these data are pertinent to the question of the relative contribution of prokaryotic LGT during eukaryote diversification more generally. Notably, in a given eukaryotic genome the number of genes of bacterial origin are typically more numerous (~2/1 ratio across 14 genomes analysed in [11]) and significantly more variable than those that can be traced to an archaeal origin, highlighting the higher evolutionary plasticity of the former [11]. The growing list of LGT identified from various prokaryotic donor lineages in different eukaryotic lineages suggests that LGT has played a significant role in shaping eukaryote protein coding capacity throughout eukaryote diversification [12•].

Parasites as model systems to study LGT in eukaryotes

Parasitic microbial eukaryotes have dramatic impact on the health of humans, farmed animals and plants, in addition to wildlife [13,14•]. They also represent important model

systems to study the evolution of eukaryotic cells and genomes as they are dispersed across eukaryote diversity [15]. The number of genome sequences from eukaryotes is increasing rapidly although sampling is still rather biased towards animals, fungi, plants and their parasites [16]. At a finer evolutionary scale sampling of genomes from different strains of a given species and closely related species represent an important source of data to investigate patterns of LGT acquisitions and losses and to study their potential link with phenotypic diversity and adaptions [2,3].

We have recently investigated the genomes of 12 microbial parasites infecting humans and animals [12•] (Table 1 lists some examples), which include members of four of the currently recognised five eukaryotic super-groups [15]. For comparison we also included the free-living soil amoeba *Dictyostelium discoideum* [12] and list recently published data for additional free-living species in supplementary Table S1. Our analyses represent one of the broadest and most detailed investigations of relatively recent LGT, explicitly excluding EGT [12•]. This is pertinent, as numerous publications have reported eukaryotic LGT for small sets of genes or individual genomes using a range of different methodologies and selection criteria to identify candidate LGTs. This makes meaningful comparison of data between publications rather difficult. Indeed very different counts of LGT have been published for a given genome depending on the methodology and database used (Table 1 and supplementary Table S1) [12•].

Animal hosts as a bazaar for LGT and dynamics of transfer

Animal microbial parasites have specialised for infecting different tissues in a given host including extracellular and intracellular niches [13]. Some are restricted to mucosal surfaces (e.g. *Trichomonas*), others are dependent on arthropod vectors (e.g. *Trypanosoma*) and enter their vertebrate hosts through a bite to initiate infections in the skin and/or in internal tissues. Mucosal and skin surfaces of humans and other vertebrates are hosts of a diverse and abundant microbiota comprising Bacteria, Archaea, microbial eukaryotes and viruses that are increasingly recognised as playing myriad roles in host biology [17••]. LGT among the bacterial microbiota of the gut mucosa was shown to be quantitatively more important (~25× times) than among prokaryotes from other environments [18], hence the gut microbiota has been dubbed a bazaar for gene exchange [19]. Mucosal parasites interact with the highly abundant and dense

Table 1

Variation of reported cases of LGT between species in a given study or between different studies for a given species for a selection of microbial parasites.

Species name	Higher rank taxonomy[a]	Total LGT count (%Proteome)[b]	P → E LGT[c]	E → E LGT[d]	Other LGT[e]	Methodology[f]	Reference
Entamoeba histolytica	Amoebozoa (Archamoebae)	199 (2.1% – 9090?)	197	NR	2 (virus)	Blast & Phylogeny	[47]
Entamoeba histolytica[*]	Amoebozoa (Archamoebae)	63 (0.68% – 9090)	51	12	NR	Blast & Phylogeny	[12•]
Entamoeba dispar	Amoebozoa (Archamoebae)	195 (1.90% – 10,262?)	194	NR	1 (virus)	Blast & Phylogeny	[47]
Trichomonas vaginalis	Excavata (Metamonada)	149 (0.24% – 59,681)	134	15	NR	Blast & Phylogeny	[12•]
Giardia lamblia	Excavata (Metamonada)	21 (0.36% – 6394)	15	6	NR	Blast & Phylogeny	[12•]
Leishmania major	Excavata (Discoba)	68 (0.96% – 7111)	63	5	NR	Blast & Phylogeny	[12•]
Trypanosoma brucei	Excavata (Discoba)	46 (0.47% – 9750)	45	1	NR	Blast & Phylogeny	[12•]
Plasmodium falciparum	SAR (Alveolata)	19 (0.36% – 5258)	18	1	NR	Blast & Phylogeny	[12•]
Encephalitozoon cuniculi	Opisthokonta (Nucletmycea)	3 (0.16 – 1918)	1	2	NR	Blast & Phylogeny	[12•]

Additional reference for Table 1: [48•].

[a] According to [15]. The two highest taxonomic ranks are indicated. SAR stands for the Stramenopiles, Alveolata and Rhizaria group.
[b] Values in brackets represent the fraction of LGT in % of the number of annotated protein coding genes, total is indicated after the dash. A question mark indicates the ambiguity about the exact dataset analysed as different annotations exist for a given genome.
[c] Candidate prokaryote to eukaryote LGTs. The great majority of candidates LGTs are from Bacteria.
[d] Candidate Eukaryote to Eukaryote LGTs.
[e] Additional sources of LGT investigated.
[f] Different criteria (BlastP and phylogenies) were used to select candidate LGT.
[*] Same dataset analysed in different publications — only two recent publications for one species were considered here. See [12•] for additional examples.
NR: none reported.