Molecular & Biochemical Parasitology xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Molecular & Biochemical Parasitology

Gene family phylogeny and the evolution of parasite cell surfaces

Andrew P. Jackson

Department of Infection Biology, Institute of Infection and Global Health, University of Liverpool, Liverpool Science Park Ic2, 146 Brownlow Hill, Liverpool L3 5RF, UK

ARTICLE INFO

Article history:
Received 25 January 2016
Received in revised form 18 March 2016
Accepted 19 March 2016
Available online xxx

Keywords:
Cell-surface
Gene family
Phylogeny
Orthology
Paralogy
Protolog

ABSTRACT

Parasite genomes typically contain unique contingency gene families encoding multi-copy effector proteins that are often expressed abundantly on the parasite cell surface and beyond. The functions of these gene families are incompletely understood but it is clear that they perform fundamental roles at the host-parasite interface. Over evolutionary timescales, the evolution of these gene families is likely to have decisive effects on the pathology and virulence of parasitic infections. In this review, I will compare the evolutionary dynamics of multiple examples from trypanosomatids and apicomplexan parasites to demonstrate how their inherent mutability makes their phylogeny very different to 'normal' gene families. I will argue that phylogenetic analyses could help to understand the functions of these enigmatic genes.

1. Introduction

Comparison of anything reveals similarities and differences. Comparison of genomes invariably shows us features that change rapidly in form and quantity, features that hardly change at all over immense time, and indeed, everything in between.

The most changeable features of unicellular parasite genomes consistently pertain to cell surfaces, or to the secretory realm beyond. Hence, while trypanosomatid parasites (*Trypanosoma*, *Leishmania*, *Leptomonas* and others) display a common physiology and ultrastructure, their cell surfaces molecules are lineage-specific and mutually exclusive [1–3]. Similarly, apicomplexan parasites (*Plasmodium*, *Babesia*, *Eimeria*, *Toxoplasma* and others) share an underlying ultrastructure and developmental regimen that is reflected in their genome content [4], but their cell surface architectures are so distinct that a common ancestral structure cannot currently be imagined [5,6]. The genomes of other unicellular parasites such as the plant-pathogenic *Phytophthora* spp. [7,8] and *Entamoeba* spp. [9] continue this trend.

Rapid evolutionary change of these genes is intuitive. The cell surface and its immediate environs comprise the host-parasite interface, and cell surface-expressed gene families have strong associations with pathology and virulence. No other compartment is subject to such powerful co-evolutionary pressures. This is reflected in the structural diversity of cell surface gene families, often called *contingency gene families* [10] because precise regulation of the expression of structural isoforms allows pathogens to respond flexibly to diverse environmental pressures. To this we might also add the idea of *contingency regions* of genomes, e.g. sub-telomeres. These are typically outside of regular chromosomal cores, and house contingency gene loci in conditions that facilitate their specialized (often irregular) expression [4,11].

The true scale of parasite contingency gene families has only become apparent in the genome sequencing era. *Plasmodium* genomes can contain hundreds of *pir* genes [12,13]; African trypanosomes can have in excess of 2500 *VSG* genes in their genome [14–17]. In parasites with complex life cycles, multi-copy families are often developmentally regulated. However, it generally remains the case that we do not understand all (or any) of the functions they perform.

In this review, I will present a series of phylogenies for parasite cell-surface gene families of trypanosomatids and apicomplexan parasites that reflect the diversity in evolutionary dynamics that exist. I will show how phylogenetic analysis of these genes can help to understand their enigmatic origins and the processes that regulate their conspicuous diversity.

I will also advance the view that phylogenetics provides an experimental rationale for determining gene functions. Deciding which paralog(s) to manipulate in an experiment is a difficult but crucial question when dealing with large and diverse families. Phylogenetic analysis identifies differences in evolutionary dynamic among paralogous loci, the simplest being whether they are conserved or species-specific. It can identify the locus most likely to

E-mail address: a.p.jackson@liv.ac.uk

http://dx.doi.org/10.1016/j.molbiopara.2016.03.007
0166-6851/

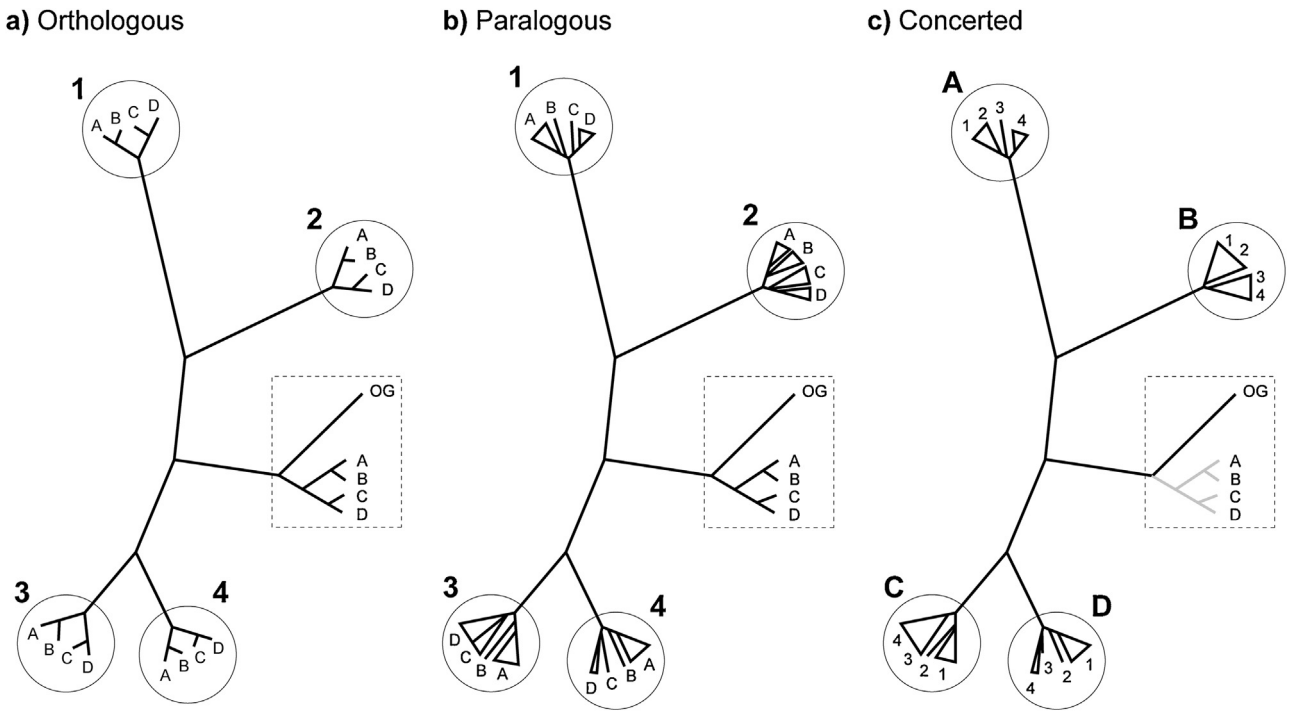

Fig. 1. Homology and the definition of a protolog. Each phylogeny considers four species (A–D) that share a gene family with four sub-families and an ancestral locus (the 'protolog'; boxed) that may or may not be present in an outgroup ('OG'). (a) In the absence of gene duplication or recombination between paralogs, clades have widespread distributions, consisting of orthologs of each sub-family from each species. (b) With gene duplication within species but no recombination, clades are still widespread, but consist of several conspecific copies of each sub-family, themselves forming clades that are co-orthologous. (c) Under conditions of concerted evolution, (i.e. rapid turnover due to gene loss or conversion), there is loss of orthology and clades consist of paralogs from a single species. A single-copy protolog may be present.

represent the origin of a family, as well as distinct sub-families that evolving under different selective environments. My contention is that such discontinuities are caused by functional differences and that, in the absence of a full understanding of functions a priori, the cladistic structure can guide our decisions as to which genes should be knocked out to expose the functional consequences of a particular evolutionary event, and which other genes we should employ in rescuing gene function to test hypotheses of redundancy or functional differentiation.

2. Cell-surface gene family phylogenetics

Given that parasite cell-surface gene families are known for their mutability and variation, phylogenetic analysis within and between species is crucial to understanding their biology. Interspecific comparisons seek to distinguish orthologs and paralogs. Orthologs are homologous gene copies present in different species and descended from a common ancestor. They generally retain the same genomic position in related species. By contrast, paralogs are descendants of a gene duplication event either in the same genome or an ancestor, often associated with the creation of a new locus in a different genomic context. Orthologs are thought more likely to maintain the same function in different species [18]; therefore, by segregating gene families in to orthologous clades we begin the task of understanding functional evolution.

All else being equal, we would expect orthologs to cluster together in a phylogeny. This is shown in Fig. 1(a) for a gene family in four species (A–D) consisting of four loci (plus a related but divergent locus that does not belong to the family, i.e. an 'outgroup'). Each gene is present in each species, leading to clades of orthologs in the phylogeny. If recent gene duplications have occurred, we may instead see each species represented by a clade of paralogs (Fig. 1(b)), but these clades remain most closely related to paralogs in a different species (i.e. they are co-orthologs).

In the phylogenies of many cell-surface gene families concerted evolution complicates this simple distinction between orthologs and paralogs. Concerted evolution describes how, for a set of species each possessing a multi-copy gene family inherited from their ancestor, the copies in each species are more closely related to each other than they are to homologs in other species. We may think of concerted evolution occurring as ancestral sequence types (which retain the signature of orthology) are gradually replaced by recently derived sequences [19] (e.g. Fig. 1(c)). This may happen due to high gene turnover (i.e. rapid and random duplication and loss of gene copies) or gene conversion, whereby gene sequences are 'overwritten' by homologous donor sequences during the repair of DNA strand breaks [20,21].

Under these circumstances we observe *loss of orthology*. A literal interpretation of Fig. 1(c) would be that all gene copies had emerged after speciation. However, as we typically have good biological reasons for thinking that the ancestral state was similar to the derived states, we do not interpret this literally, but as evidence for rapid gene turnover.

Multi-copy gene families and concerted evolution are not unique to parasites; nevertheless, they are a consistent feature of parasite genomes [7,16,22–24]. Besides the many paralogs of recent origin, these families often include rare members, which I will refer to as 'protologs' because they are the most likely to represent the ancestral state, prior to the elaboration of a gene family. Putative protologs usually have atypical structures and are located outside of the contingency region, (e.g. in a chromosome-internal locus). They are typically present as orthologs in multiple species both parasitic and free-living, and branch closest to the root in the phylogeny. Together, these various properties point to their ancient origins. Experimental approaches to the evolution of gene function clearly need to address a protolog, if there is one, in comparison with the derived functions of parasite-specific genes.

Please cite this article in press as: A.P. Jackson, Gene family phylogeny and the evolution of parasite cell surfaces, Mol Biochem Parasitol (2016), http://dx.doi.org/10.1016/j.molbiopara.2016.03.007