THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Xenolog classification

Citation for published version:
Darby, CA, Stolzer, M, Ropp, PJ, Barker, D & Durand, D 2017, 'Xenolog classification', *Bioinformatics*, vol. 33, no. 5, pp. 640-649. https://doi.org/10.1093/bioinformatics/btw686

Digital Object Identifier (DOI):
10.1093/bioinformatics/btw686

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Bioinformatics

Publisher Rights Statement:
© The Author 2016. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 23. Jan. 2025

Bioinformatics, 33(5), 2017, 640–649
doi: 10.1093/bioinformatics/btw686
Advance Access Publication Date: 29 December 2016
Original Paper

OXFORD

Phylogenetics

Xenolog classification

Charlotte A. Darby[1,†,§], Maureen Stolzer[1,§], Patrick J. Ropp[1], Daniel Barker[2,‡] and Dannie Durand[1,3,*]

[1]Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA, [2]School of Biology, University of St. Andrews, St. Andrews, Fife KY16 9TH, UK and [3]Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

*To whom correspondence should be addressed.
[†]Present address: Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA
[‡]Present address: Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3FL, UK
[§]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Cedric Chauve

Received on August 5, 2016; revised on October 11, 2016; editorial decision on October 24, 2016; accepted on October 26, 2016

Abstract

Motivation: Orthology analysis is a fundamental tool in comparative genomics. Sophisticated methods have been developed to distinguish between orthologs and paralogs and to classify paralogs into subtypes depending on the duplication mechanism and timing, relative to speciation. However, no comparable framework exists for xenologs: gene pairs whose history, since their divergence, includes a horizontal transfer. Further, the diversity of gene pairs that meet this broad definition calls for classification of xenologs with similar properties into subtypes.
Results: We present a xenolog classification that uses phylogenetic reconciliation to assign each pair of genes to a class based on the event responsible for their divergence and the historical association between genes and species. Our classes distinguish between genes related through transfer alone and genes related through duplication and transfer. Further, they separate closely-related genes in distantly-related species from distantly-related genes in closely-related species. We present formal rules that assign gene pairs to specific xenolog classes, given a reconciled gene tree with an arbitrary number of duplications and transfers. These xenology classification rules have been implemented in software and tested on a collection of ∼13 000 prokaryotic gene families. In addition, we present a case study demonstrating the connection between xenolog classification and gene function prediction.
Availability and Implementation: The xenolog classification rules have been implemented in Notung 2.9, a freely available phylogenetic reconciliation software package. http://www.cs.cmu.edu/~durand/Notung. Gene trees are available at http://dx.doi.org/10.7488/ds/1503.
Contact: durand@cmu.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Homology analysis, classifying gene pairs according to the evolutionary process by which they diverged, is a fundamental tool of comparative genomics. Identifying orthologs is integral to the functional annotation of novel genes (Wu *et al.*, 2003) and prediction of gene function by various methods, including phylogenetic profiling (Pellegrini *et al.*, 1999) and gene fusion (Enright *et al.*, 1999; Marcotte *et al.*, 1999). Phylostratigraphic investigations linking the age of a gene to its functions, disease associations, or ecological distribution exploit the fact that orthologs from the same pair of species diverged at roughly the same time (Capra *et al.*, 2013). Orthologs are used as markers for homologous chromosomal

© The Author 2016. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/33/5/640/2725487 by Edinburgh University user on 08 February 2019

regions for comparative mapping (Nadeau and Sankoff, 1998; O'Brien *et al.*, 1997), phylogenetic footprinting (Dickmeis and Muller, 2005; Duret and Bucher, 1997) and operon prediction (Chen *et al.*, 2004; Ermolaeva *et al.*, 2001; Price *et al.*, 2005; Westover *et al.*, 2005).

Identification of paralogs is a prerequisite for studying processes of gene duplication, the major source of genetic novelty in eukaryotes. Comparison of paralogous pairs with a pre-duplication ortholog reveals patterns and rates of diversification following duplication (Lynch, 2007, and work cited therein), as well as the functional fates of duplicated genes (Lynch, 2007). Spatial patterns of orthologs and paralogs are used to infer the specific duplication process that gave rise to a given set of paralogs (Durand and Hoberman, 2006; Simillion *et al.*, 2004; Van de Peer, 2004).

Homology identification is a highly active research area, comprising methodological approaches ranging from sequence comparison to phylogenetic reconciliation. More recent innovations include the exploitation of shared synteny (Shi *et al.*, 2011) and specialized methods for identifying multidomain homologs (Ali *et al.*, 2016; Song *et al.*, 2007, 2008).

Most work on homology analysis to date has not considered genes related through horizontal transfer. Studies of horizontal transfer commonly use approaches that seek to identify genes of foreign origin in a given genome, rather than homologous gene pairs that are related through horizontal transfer (reviewed in Azad and Lawrence, 2012). A few methods, such as gene tree–species tree reconciliation, do infer gene pairs that correspond to the donor and recipient of a transfer. Reconciliation algorithms that account for transfer events are relatively new (reviewed by Huson and Scornavacca, 2011; Nakhleh, 2010, 2013), computationally more complex, and are only recently coming into use for genomic analyses (e.g. David and Alm, 2011; Richards *et al.*, 2014).

Appropriate terminology for describing gene pairs related through horizontal transfer is a fundamental requirement for extending the homology analysis framework to include this evolutionary process. The term 'xenolog', proposed by Gray and Fitch (1983) to describe horizontally transferred genes, is in use, but not widely, and there is no consensus on a precise definition. Further, there has been little discussion of differentiating between xenologs to convey distinctions between horizontally transferred genes with different properties (see Koonin *et al.*, 2001, for a notable exception). Such xenolog classes would be analogous to paralog subtypes proposed to convey the relative timing of duplications and speciations (e.g. in-paralogs versus out-paralogs, Sonnhammer and Koonin, 2002) or distinguish between different mechanisms of duplication (e.g. ohnologs and tandem duplications, reviewed by Durand and Hoberman, 2006; Ramos and Ferrier, 2012).

Background: Fitch (1970) introduced the terms orthology ('homology [that] is the result of speciation') and paralogy ('homology [that] is the result of gene duplication') and proposed that 'foreign genes . . . since they are neither orthologous nor paralogous but are clearly homologous . . . should be called xenologous' (Gray and Fitch, 1983). These definitions, which are framed in terms of the event that caused the divergence, have been widely adopted. In 2000, Fitch proposed more precise definitions of orthology and xenology: Orthology includes the requirement that the 'common ancestor lies in the cenancestor of the taxa from which the two sequences were obtained,' where a *cenancestor* is the 'most recent common ancestor of the [species] taxa under consideration,' and xenology is the 'relationship of any two homologous characters whose history, since their common ancestor, involves an interspecies (horizontal) transfer of the genetic material for at least one of those characters.' In other words, a pair of genes, g_1 and g_2, are xenologs, if there is a transfer on the path connecting g_1 and g_2 in the gene tree.

In this updated definition, orthology is defined not just in terms of a speciation event, but in terms of the association of nodes in the gene and species trees. Under a duplication-loss event model, the event-based definition of orthology and this definition are equivalent. However, when transfers are included in the event model, the sets of orthologs predicted using the two definitions are not identical. Moreover, the event-based definition leads to predicted orthologs that have properties that are not usually associated with orthologs.

For example, nodes g_X and $\hat{g}$ in Figure 1 are orthologs according to the event-based definition, because the event at their most recent common ancestor (g_4) is a speciation. Yet g_X and $\hat{g}$ are genes in the same present-day species X, violating the assumption that genes in the same species cannot be orthologous. Gene pairs in species X and Z also exhibit surprising behavior according to the event-based definition. The most recent common ancestor (MRCA) of g_X and g_Z is a speciation node, as is the most recent common ancestor of $\hat{g}$ and g_Z, implying that both pairs are orthologs in the species X and Z. However, these pairs arose at very different times in the species tree, violating the assumption that orthologs drawn from the same pair of species are associated with the same species divergence and are roughly the same age (Capra *et al.*, 2013; Goodman *et al.*, 1979). Neither of these problems arises when the cenancestor-based definition is used, because neither g_X nor g_Z are orthologs of $\hat{g}$ according to that definition. In both cases, the MRCA of the genes does not lie in their cenancestor.

The additional cenancestor requirement results in a restricted set of orthologs that excludes these problematic cases. However, a consequence of defining orthologs narrowly is that xenologs are defined broadly: the set of gene pairs whose history, since their divergence, includes a transfer is substantially larger than the set of genes that diverged through a transfer event at their MRCA. Xenologs, when broadly defined, exhibit diverse properties. First, not all xenologs have the same event at their MRCA in the gene tree. We observe xenologs where this divergence arose via transfer (e.g. $\hat{g}$ and g_Y), speciation (e.g. $\hat{g}$ and g_Z) and duplication (e.g. $\hat{g}$ and h_Z). Second, xenologs can occur in the same species (e.g. $\hat{g}$ and g_X). Third, xenologs may vary greatly in how closely they are related, and the divergence of a pair of xenologs may pre- or post-date the divergence of their associated species. For example, genes $\hat{g}$ and g_Z diverged more recently than species X and species Z, whereas genes $\hat{g}$ and g_W diverged before species X and species W.

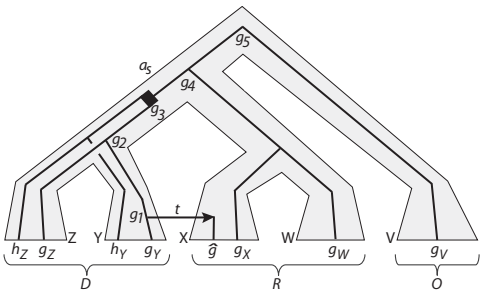

Fig. 1. Gene tree (thin black lines), with a duplication and a transfer from species *Y* to species *X*, embedded in a species tree (shown in gray). The cenancestor of the transfer is designated a_s. Species sets *D*, *R* and *O* are labeled below the leaves

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/33/5/640/2725487 by Edinburgh University user on 08 February 2019

Our contributions: This broad definition of xenologs does not convey important distinctions between the diverse and complex xenologous relationships that arise due to horizontal gene transfer. To address this, we propose xenolog classes that reflect the events associated with the divergence of a xenologous gene pair, and the relative timing of transfer and speciation events. We present formal definitions of these classes in the context of a reconciled gene tree and rules to assign xenologous gene pairs to classes. Further, we show that these classes form a hierarchy, connecting the relationship of xenologs to their placement in the gene and species trees.

An algorithm implementing these rules has been integrated into the NOTUNG 2.9 software package. An analysis of ~13 000 prokaryotic gene families demonstrates that all of the proposed classes arise in real gene tree data. We further present a case study that illustrates the potential functional implications of xenolog classification. Finally, we discuss how this framework could be used in future research to explore the evolutionary and functional fates of transferred genes.

Notation: Before stating formal definitions of the xenolog subtypes, we introduce the following notation. For a binary, rooted tree $T_i = (V_i, E_i)$ with node set V_i and edge set E_i, $L(T_i)$ designates the leaf set of T_i. $V \backslash U$ denotes vertices in set V that are not in set U, where $U \subset V$. $p(v)$ refers to the parent of node v. If v is an ancestor (resp., descendant) of u in T_i, we write $v >_i u$ (resp., $v <_i u$). The set $\Delta(u)$ represents the improper descendants of node u, i.e. u and all nodes in the subtree rooted at u. If $v \notin \Delta(u)$ and $u \notin \Delta(v)$, then we say that u and v are incomparable (denoted $u \not\lessgtr v$). The most recent common ancestor of u and v is denoted $\mathrm{MRCA}(u, v)$. Given $v_1, v_2, v_3 \in V_i$, we say that v_1 is more closely related to v_2 than to v_3, if $\mathrm{MRCA}(v_1, v_2) <_i \mathrm{MRCA}(v_1, v_3)$.

2 Methods

Our classification takes as input a gene tree, $T_G = (V_G, E_G)$, that has been reconciled with a species tree, $T_S = (V_S, E_S)$, using a duplication-transfer model. The model may also include losses; losses have no impact on xenolog classification and we do not discuss them further. Reconciliation infers a mapping, $\mathcal{M}(\cdot)$, between genes and species, where $\mathcal{M}(g) = s$ indicates that gene $g \in V_G$ was present in the genome of species $s \in V_S$. Each internal node, g, is annotated with $\mathcal{E}(g)$, the event that caused the divergence at g, where $\mathcal{E}(g)$ can be a duplication (δ), a transfer (τ), or a speciation (σ). Transfer edges are denoted by $t = (g_d, g_r)$, where g_r is the recipient gene node, $g_d = p(g_r)$ is the donor gene node, and $\mathcal{E}(g_d) = \tau$. We say transfer t is on the path from g_i to g_j, if the path from g_i to g_j passes through both g_d and g_r.

The output of our classification scheme is a homology table $H[g_i, g_j], \forall g_i, g_j \in L(V_G)$. In this classification, which is based on the definitions introduced by Fitch (2000), genes g_i and g_j are

orthologs iff $\mathcal{E}(\mathrm{MRCA}(g_i, g_j)) = \sigma$ and there is no transfer on the path from g_i to g_j;

paralogs iff $\mathcal{E}(\mathrm{MRCA}(g_i, g_j)) = \delta$ and there is no transfer on the path from g_i to g_j;

xenologs iff there is at least one transfer on the path from g_i to g_j.

Note that by explicitly defining orthologs to be gene pairs that are not connected by a transfer, this definition of ortholog ensures that the ancestor of orthologous genes lie in their cenancestor; i.e. $\mathcal{M}(\mathrm{MRCA}(g_i, g_j)) = \mathrm{MRCA}(\mathcal{M}(g_i), \mathcal{M}(g_j))$.

If g_i and g_j are orthologs, then $H[g_i, g_j] = H[g_j, g_i] = \mathrm{O}$. If they are paralogs, $H[g_i, g_j] = H[g_j, g_i] = \mathrm{P}$. If g_i and g_j are xenologs, then $H[g_i, g_j] = \mathcal{X}(g_i, g_j)$, where $\mathcal{X}(g_i, g_j)$ is the xenolog class of genes g_i and g_j. In contrast to orthology and paralogy, xenology is not symmetric, due to the directional nature of horizontal transfer.

In the remainder of this section, we define new xenolog classes and give formal rules for determining the xenolog class, $\mathcal{X}(g_i, g_j)$. In Section 2.1, we consider the case where there is a single transfer on the path from g_i to g_j and they did not diverge by duplication (i.e. $\mathcal{E}(\mathrm{MRCA}(g_i, g_j) \neq \delta)$. In Section 2.2, we provide xenolog classification rules for the case where the MRCA of g_i and g_j is a duplication and introduce a subclass of xenologs, called paraxenologs, for designating genes that are related through both duplication and transfer. Finally, in Section 2.3, we extend these definitions to allow an arbitrary number of transfers on the path from g_i to g_j.

2.1 Xenolog classification with a single transfer

Consider a gene tree with a single transfer $t = (g_d, g_r)$ from donor species $s_d = \mathcal{M}(g_d)$ to recipient species $s_r = \mathcal{M}(g_r)$. Let $a_s = \mathrm{MRCA}(s_d, s_r)$ be the cenancestor of t and let A be the set of nodes in the subtree of T_S rooted at a_s. Transfer t defines three, non-overlapping sets of species tree nodes:

$D = \{s \in V_S | \mathrm{MRCA}(s, s_d) <_s a_s\}$, i.e. the species that are more closely related to the donor than the recipient;

$R = \{s \in V_S | \mathrm{MRCA}(s, s_r) <_s a_s\}$, i.e. the species that are more closely related to the recipient than the donor;

$O = V_S \backslash A$, i.e. the nodes in the species tree equally related to the donor and recipient.

We define four mutually exclusive xenolog classes based on these sets. Xenolog classes are defined with respect to a reference gene $\hat{g} \in L(T_G)$ that is a descendant of the recipient of the transfer; i.e. $\hat{g} \in \Delta(g_r)$. For every $g \in \{L(V_G) \backslash \Delta(g_r)\}$, t is on the path from $\hat{g}$ to g and g is a

Primary xenolog iff $g \in \Delta(g_d)$;	$\mathcal{X}(\hat{g}, g) = \mathrm{PX}$
Sibling Donor xenolog iff $\mathcal{M}(g) \in D$ and $g \notin \Delta(g_d)$;	$\mathcal{X}(\hat{g}, g) = \mathrm{SDX}$
Sibling Recipient xenolog iff $\mathcal{M}(g) \in R$;	$\mathcal{X}(\hat{g}, g) = \mathrm{SRX}$
Outgroup xenolog iff $\mathcal{M}(g) \in O$.	$\mathcal{X}(\hat{g}, g) = \mathrm{OX}$

Xenologs are classified relative to a reference gene; therefore, xenolog class assignments are not symmetric. In the homology table, when $H[\hat{g}, g] = \mathcal{X}(\hat{g}, g)$, $H[g, \hat{g}] = *$ is used to indicate that g is the xenolog of the reference gene, $\hat{g}$, and that its class is given by $H[\hat{g}, g]$.

In Figure 1, all genes are xenologous to $\hat{g}$. Both g_Y and g_Z are in set D; g_Y is a Primary xenolog ($\mathcal{X}(\hat{g}, g_Y) = \mathrm{PX}$) and g_Z is a Sibling Donor xenolog ($\mathcal{X}(\hat{g}, g_Z) = \mathrm{SDX}$), because g_Y is a descendant of the donor (i.e. $g_Y \in \Delta(g_1)$) and g_Z is not. Genes g_X and g_W are in set R and are Sibling Recipient xenologs ($\mathcal{X}(\hat{g}, g_W) = \mathrm{SRX}$). Gene g_V is an Outgroup xenolog ($\mathcal{X}(\hat{g}, g_V) = \mathrm{OX}$) because g_V is in set O. Genes h_Y and h_Z are paraxenologs and will be discussed in Section 2.2.

A xenologous gene pair can be further annotated to indicate cases where the genes are found in the same species: g is an **autoxenolog** of $\hat{g}$, iff $\mathcal{M}(g) = \mathcal{M}(\hat{g})$. We designate this $\mathcal{X}(\hat{g}, g) = \mathrm{X}'$. Autoxenologs will also be assigned to a subclass. In Figure 1, g_X and $\hat{g}$ are both in species X; g_X is a Sibling Recipient autoxenolog ($\mathcal{X}(\hat{g}, g_X) = \mathrm{SRX}'$).

Xenolog class hierarchies: The xenolog classes form a hierarchy that can elucidate how xenologs are related in both the gene and species trees. Primary xenologs are closest in the xenolog hierarchy

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/33/5/640/2725487 by Edinburgh University user on 08 February 2019

and Outgroup xenologs are most distant. We denote this hierarchy by

$$\mathrm{PX} <_X \mathrm{SDX} <_X \mathrm{SRX} <_X \mathrm{OX},$$

where $\mathcal{X}(\hat{g}, g_1) <_X \mathcal{X}(\hat{g}, g_2)$, if $\hat{g}$ and g_1 are closer in the hierarchy than $\hat{g}$ and g_2.

Genes that are more closely related in the hierarchy are also more closely related in the gene tree. Let genes g_1 and g_2 in $V_G \backslash \Delta(g_r)$ be xenologs of $\hat{g}$ such that there is no transfer ancestral to either g_1 or g_2. Then, $\mathrm{MRCA}(\hat{g}, g_1) <_G \mathrm{MRCA}(\hat{g}, g_2)$, if $\mathcal{X}(\hat{g}, g_1) <_X \mathcal{X}(\hat{g}, g_2)$. This hierarchy, which is illustrated in Figure 2, is stated formally as follows:

THEOREM 2.1. (Xenolog class hierarchy in the gene tree) *Given* $\hat{g} \in \Delta(g_r)$, *for any Primary xenolog,* g_{P}, *Sibling Donor xenolog,* g_{SD}, *Sibling Recipient xenolog,* g_{SR}, *and Outgroup xenolog,* g_{O}, *of* $\hat{g}$

$$\mathrm{MRCA}(\hat{g}, g_{\mathrm{P}}) <_G \mathrm{MRCA}(\hat{g}, g_{\mathrm{SD}}) <_G \mathrm{MRCA}(\hat{g}, g_{\mathrm{SR}}) <_G \mathrm{MRCA}(\hat{g}, g_{\mathrm{O}}).$$

PROOF. See Supplementary Section S.1. □

We sketch the basis of this theorem informally, here. For every xenolog $g \in V_G \backslash \Delta(g_r)$ of $\hat{g}$, the common ancestor of g and $\hat{g}$ is a node on the path from g_d to the root of T_G; i.e. there exists $g_i \in V_G$, such that $g_i = \mathrm{MRCA}(\hat{g}, g)$ and $g_i \geq_G g_d$. If $g_i = g_d$, then $g \in \Delta(g_d) \backslash \Delta(g_r)$ and is therefore a Primary xenolog.

For $g_i > g_d$, the descendants of c_i, the child of g_i that is incomparable to the transfer, must satisfy two requirements. First, since all xenologs in $\Delta(c_i)$ are equally related to $\hat{g}$, all xenologs in $\Delta(c_i)$ must be assigned to the same xenolog class. This will be true if all descendants of c_i are in the same species set, D, R or O. Second, for any $g_j >_G g_i$, the xenologs in $\Delta(c_j)$ are more distantly related to $\hat{g}$ than the xenologs in $\Delta(c_i)$; therefore, consistency requires that the class of xenologs in $\Delta(c_j)$ not be closer in the hierarchy than the class of xenologs in $\Delta(c_i)$. Both of these conditions are satisfied when there is no transfer that is ancestral to either g_1 or g_2. This is always true in a reconciled tree with a single transfer and no duplications. We will reexamine the hierarchical properties of xenolog classes in trees with more complex event histories in the following sections.

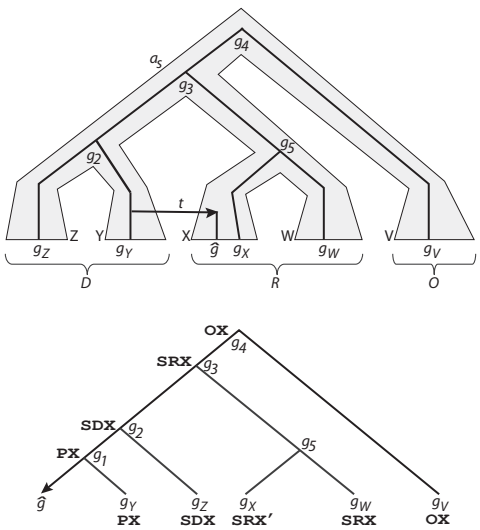

Fig. 2. Xenolog class hierarchy: (top) Gene tree with one transfer, shown in the context of the species tree. (bottom) The reconciled gene tree. Each leaf is annotated with its xenolog class. Nodes g_1, g_2, g_3 and g_4 are the common ancestors of $\hat{g}$ and, respectively, the Primary, Sibling Donor, Sibling Recipient and Outgroup xenologs in the tree, as indicated by the labels on internal nodes. The labels on the path from $\hat{g}$ to the root satisfy the hierarchy, $\mathrm{PX} <_X \mathrm{SDX} <_X \mathrm{SRX} <_X \mathrm{OX}$, consistent with Theorem 2.1

The proposed xenolog classes also convey information about the relationship of a xenolog pair in the gene tree relative to their relationship in the species tree. For xenologs, the cenancestor of $\hat{g}$ and g can predate or postdate the species containing $\mathrm{MRCA}(\hat{g}, g)$. Our xenolog classes distinguish between these three cases and are summarized in Supplementary Table S1. Primary and Sibling Donor xenologs are more closely related in the gene tree than in the species tree, whereas Sibling Recipient xenologs are more closely related in the species tree than in the gene tree. Outgroup xenologs are equally related in both trees.

2.2 Xenolog classification with transfers and duplications

We next consider the classification of genes g_i and g_j when there is a single transfer on the path from g_i to g_j and they diverged by duplication (i.e. $\mathcal{E}(\mathrm{MRCA}(g_i, g_j) = \delta)$. Such gene pairs satisfy both the paralog and the xenolog criteria proposed by Fitch (2000), leading to potential terminological confusion. To avoid this confusion, we introduce the explicit designation, **paraxenolog**, for xenologs that diverged via a duplication at their common ancestor. Note that Patterson (1988) used 'paraxenolog' to refer to a different phenomenon.

Formally, let $g_{\mathrm{DUP}} \in V_G$ be a duplication node in the gene tree with a transfer, $t = (g_d, g_r)$, in one of its two subtrees, and let $\hat{g} \in \Delta(g_r)$ be a descendant of that transfer. Then, every gene in the other subtree of g_{DUP} is a paraxenolog of $\hat{g}$, to be denoted X^P. For example, in the gene tree in Figure 1, $g_{\mathrm{DUP}} = g_3$ is a duplication node with two subtrees; the g subtree contains a transfer with reference gene $\hat{g}$. All genes in the other subtree (that is, h_Y and h_Z) are paraxenologs of $\hat{g}$.

Paraxenologs are also assigned to a specific xenolog class when it is both possible to do so and preserve the xenolog class hierarchy, as specified in Theorem 2.1. This depends on when the duplication occurred relative to a_s, the cenancestor of the transfer. If the species in which the duplication occurred is a descendant of a_s, then all descendants of g_{DUP} are more closely related to the donor than to the recipient; i.e. all paraxenologs are in species in D and must be Sibling Donor xenologs. They cannot be Primary xenologs, as, by definition, Primary xenologs are the descendants of a transfer. In this case, paraxenologs satisfy the requirements of Theorem 2.1, because all paraxenologs of $\hat{g}$ are equally related to $\hat{g}$ and are assigned to the same xenolog class; the hierarchy is preserved.

When the duplication predates or coincides with the cenancestor of the transfer, then the descendants of both children of g_{DUP} will be inherited by species in D, R and potentially O. These paraxenologs are equally related in the gene tree, but would be assigned different classes based on their location, thus violating the requirements of Theorem 2.1. To avoid violating the hierarchy, for every paraxenolog, g, of $\hat{g}$, we assign $\mathcal{X}(\hat{g}, g)$ to X^P, i.e. $\hat{g}$ and g are untyped paraxenologs. A scenario where this occurs is shown in Supplementary Figure S1.

Xenolog hierarchy with paraxenologs: The xenolog hierarchy in Theorem 2.1 holds for paraxenologs if we ignore the distinction between xenologs and paraxenologs of the same class and consider X^P to be on a par with the OX class in the hierarchy. If g_{SD} and g_{SD^P} are a Sibling Donor xenolog and a Sibling Donor paraxenolog, respectively, of $\hat{g}$, then $\mathrm{MRCA}(\hat{g}, g_{\mathrm{SD}^P})$ may be either ancestral to or a descendant of $\mathrm{MRCA}(\hat{g}, g_{\mathrm{SD}})$ (Fig. 3). Similarly, $\mathrm{MRCA}(\hat{g}, g_{\mathrm{X}^P})$ may be

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/33/5/640/2725487 by Edinburgh University user on 08 February 2019

an ancestor or a descendant of MRCA$(\widehat{g}, g_O)$, where g_O is an Outgroup xenolog of $\widehat{g}$ and g_{X^p} is an untyped paraxenolog. These results are stated formally in Theorem S.2.

The species hierarchy in Supplementary Table S1 is also preserved, with the additional observations that Sibling Donor paraxenologs behave like Sibling Donor xenologs and $\text{MRCA}(\mathcal{M}(\widehat{g}), \mathcal{M}(g_{X^p})) \geq_S \mathcal{M}(\text{MRCA}(\widehat{g}, g_{X^p}))$.

2.3 Xenolog classification with multiple transfers

With a single transfer, xenolog classes are defined in terms of the sets of species tree nodes, D, R and O, which are determined by the positions of the donor and recipient species and their common ancestor, a_s. The key issue in extending the framework to multiple transfers is how to obtain a single D, R and O given multiple donor and recipient species. We first describe a xenolog classification procedure for a pair of genes connected by a path containing $k \geq 2$ transfer edges, when all k transfers are mutually comparable. Transfers, $t^1 = (g_d^1, g_r^1)$ and $t^2 = (g_d^2, g_r^2)$, are comparable, iff g_r^1 and g_r^2 are comparable in the gene tree. Then, we describe a procedure for the case where the gene pair is separated by incomparable transfers. The remainder of this section applies to both xenologs and paraxenologs; for simplicity, we use 'xenolog' to refer to both.

Comparable transfers: Let $t^1, t^2 \ldots t^k$ be an ordered sequence of comparable transfers on the path from $\widehat{g}$ to g. We say that t^1 is ancestral to t^2 (denoted $t^1 >_G t^2$), iff $g_r^1 >_G g_r^2$. Any set of comparable transfers, $t^1, t^2 \ldots t^k$, can be ordered such that $t^i >_G t^{i+1}, \forall\, i < k-1$. In particular, $g_d^1 >_G g_d^i, \forall\, i > 1$, and $g_r^i >_G g_r^k, \ \forall\ i < k$.

Since $t^1, t^2 \ldots t^k$ are comparable, they must be on the path between $\widehat{g} \in \Delta(g_r^k)$ and MRCA$(\widehat{g}, g)$. These transfers can be summarized by a single super-transfer, $t^* = (g_d^*, g_r^*)$, where $g_d^* = g_d^1$ and $g_r^* = g_r^k$. With one exception, discussed below, t^* behaves like a single transfer that could occur in a reconciled tree: the cenancestor of the super-transfer, $a_s^* = \text{MRCA}(s_d^*, s_r^*)$, induces sets D^*, R^* and O^* (Fig. 4). These are used to determine $\mathcal{X}(\widehat{g}, g)$, using the single-transfer procedure previously described.

The exceptional case arises when the recipient species of the super-transfer is a descendant of the donor species ($s_r^* \in \Delta(s_d^*)$). This scenario (Supplementary Fig. S2) cannot occur with a single transfer because the donor and recipient species of a transfer must be incomparable. With multiple transfers, however, s_r^k may be in $\Delta(s_d^1)$. In this case, the cenancestor of the super-transfer is also its donor ($a_s^* = s_d^*$). Since all descendants of s_d^* are also descendants of a_s^*, all xenologs in A^* are Primary xenologs. All other xenologs are in O^* and are Outgroup xenologs.

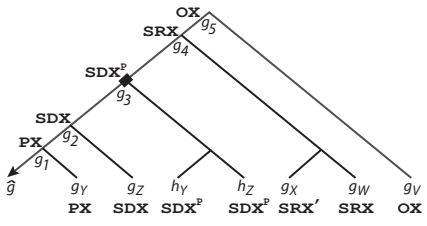

Fig. 3. Paraxenolog classification: The gene tree in Figure 1, which contains a duplication followed by a transfer. Each leaf is annotated with its xenolog class. Each internal node on the path from $\widehat{g}$ to the root is labeled with the xenolog class of all genes in its right subtree (i.e. the subtree that does not contain a transfer.) The progression of labels satisfy the xenolog hierarchy, consistent with Theorem 2.1.

A possible concern about replacing k transfers with a single super-transfer is that the intermediate species are not considered. However, these intermediate species are represented by xenologous pairs that only pass through a subset of the k transfers, namely, $g \in \Delta(g_r^1) \backslash \Delta(g_r^k)$. Information about where ancestral forms of $\widehat{g}$ spent time as $\widehat{g}$ traveled from s_d^1 to s_r^k is captured by the complete set of xenologs of $\widehat{g}$.

Incomparable transfers: We first consider the special case where $k = 2$ and the transfers are incomparable. Given a pair of genes, g_1 and g_2, connected by two incomparable transfers, t^1 and t^2 (Fig. 5), one gene is a descendant of one transfer recipient ($g_1 \in \Delta(g_r^1)$), and the other gene is a descendant of the other transfer recipient ($g_2 \in \Delta(g_r^2)$). Since g_1 and g_2 are both descendants of a transfer recipient, xenolog g_2 can be classified with respect to $\widehat{g}_1$, and vice versa.

With incomparable transfers, the xenolog classes do not satisfy the hierarchical properties of Theorem 2.1. Let $g_i = \text{MRCA}(\widehat{g}_1, g_2)$ and let c_i be the child of g_i that is ancestral to t^2 but not t^1 (i.e. $c_i \geq_G g_2$ and $c_i \not\geq_G \widehat{g}_1$). Recall that the first condition for preservation of the hierarchy is that all xenologs in $\Delta(c_i)$ must be in the same species set. Satisfaction of this condition is not guaranteed for incomparable xenologs because $\Delta(c_i)$ contains a transfer, t^2, that can move g_r^2 to a species that is not in the same set as $\mathcal{M}(g_d^2)$. Suppose, for example, the donor of t^2 is in a species in O, but its recipient is in a species in D. Since both g_d^2 and g_r^2 are in $\Delta(c_i)$, more than one species set is represented in $\Delta(c_i)$, violating the first condition. Primary xenologs are the one exception to this problem. Primary xenologs are defined in terms of g_d and not in terms of D, R and O, and are therefore unaffected by incomparable transfers.

To avoid a classification that violates the hierarchy, we do not assign xenologs separated by incomparable transfers to specific

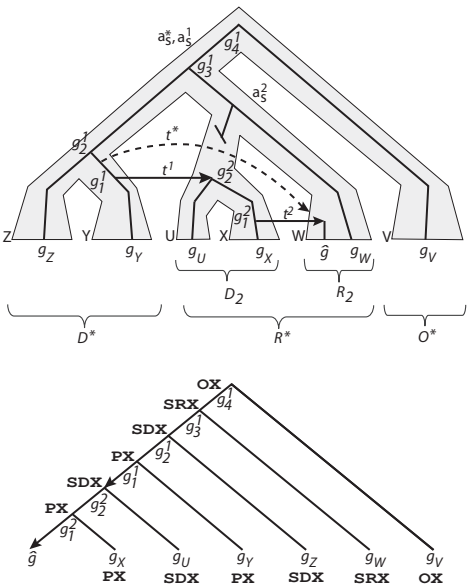

Fig. 4. Xenolog classification with multiple transfers: (top) Gene tree with two comparable transfers (solid arrows) and the associated super-transfer (dashed arrow), shown in the context of the species tree. (bottom) The reconciled gene tree. Each leaf is annotated with its xenolog class. Genes g_X and g_U are classified with respect to t^2 and obey the hierarchy: $\text{MRCA}(\widehat{g}, g_X) <_G \text{MRCA}(\widehat{g}, g_U)$ and $\mathcal{X}(\widehat{g}, g_X) = \text{PX} <_X \mathcal{X}(\widehat{g}, g_U) = \text{SDX}$. All other genes are classified with respect to the super-transfer, t^*. Their xenolog classes are consistent with the hierarchy (Theorem 2.1): $\text{MRCA}(\widehat{g}, g_Y) <_G \text{MRCA}(\widehat{g}, g_Z) <_G \text{MRCA}(\widehat{g}, g_W) <_G \text{MRCA}(\widehat{g}, g_V)$ and $\text{PX} <_X \text{SDX} <_X \text{SRX} <_X \text{OX}$

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/33/5/640/2725487 by Edinburgh University user on 08 February 2019

subclasses. Given two genes separated by incomparable transfers, t^1 and t^2, without loss of generality, let $\widehat{g}_1 \in \Delta(g_r^1)$ be the reference gene, $g_2 \in \Delta(g_r^2)$ be the xenolog under classification, and $g_m = \mathrm{MRCA}(\widehat{g}_1, g_2)$ be their common ancestor. Then g_2 is a

Primary xenolog iff $g_2 \in \Delta(g_d^1)$; $\mathcal{X}(\widehat{g}_1, g_2) = \mathrm{PX}$

Incomparable xenolog iff $g_2 \notin \Delta(g_d^1)$ and $\mathcal{E}(g_m) = \sigma$; $\mathcal{X}(\widehat{g}_1, g_2) = \mathrm{IX}$

Incomparable paraxenolog iff $g_2 \notin \Delta(g_d^1)$ and $\mathcal{E}(g_m) = \delta$. $\mathcal{X}(\widehat{g}_1, g_2) = \mathrm{IX}^P$

In the incomparable case, $H[\widehat{g}_1, g_2] = \mathcal{X}(\widehat{g}_1, g_2)$ is the classification of g_2 with respect to $\widehat{g}_1$ and $H[\widehat{g}_2, g_1] = \mathcal{X}(\widehat{g}_2, g_1)$ is the classification of g_1 with respect to $\widehat{g}_2$. Either $\mathcal{X}(g_1, g_2) = \mathrm{PX}$ and $\mathcal{X}(g_2, g_1) = \mathrm{IX}$ (or vice versa), or $\mathcal{X}(g_1, g_2) = \mathcal{X}(g_2, g_1) = \mathrm{IX}^{(\mathrm{P})}$.

We now address the case where $k > 2$ by reducing the problem to one involving two incomparable super-transfers and applying the protocol just described. Let $t^1 \cdots t^i$ be the transfers, in descending order, on the path from $\mathrm{MRCA}(g_1, g_2)$ to g_1 and $t^{i+1} \cdots t^k$ be the set of transfers on the path from $\mathrm{MRCA}(g_1, g_2)$ to g_2. Since $t^1 \cdots t^i$ must be mutually comparable, they can be replaced with super-transfer $t^{1*} = (g_d^{1*}, g_r^{1*})$, where $g_d^{1*} = g_d^1$ and $g_r^{1*} = g_r^i$. Similarly, we replace $t^{i+1} \cdots t^k$ with super-transfer $t^{2*} = (g_d^{2*}, g_r^{2*})$, where $g_d^{2*} = g_r^{i+1}$ and $g_r^{2*} = g_r^k$.

Xenolog hierarchy for multiple transfers: With multiple comparable transfers, the hierarchical properties in Theorem 2.1 hold for xenologs that share the same super-transfer from $\mathrm{MRCA}(\widehat{g}, g)$ to $\widehat{g}$. For example, in Figure 4, the xenolog class hierarchy is preserved for nodes g_X and g_U, which are xenologs of $\widehat{g}$ with respect to t^2 only. Similarly, xenologs g_Y, g_Z, g_W and g_V, which are all defined with respect to the super-transfer t^*, also obey the hierarchy.

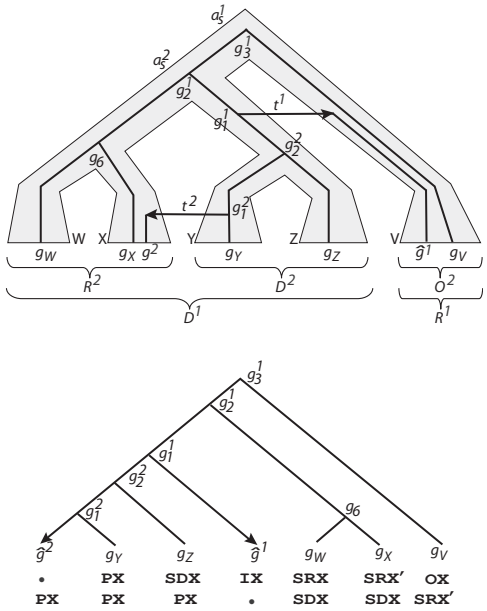

Fig. 5. Xenolog classification with incomparable transfers: (top) Gene tree with two incomparable transfers shown in the context of the species tree. Species sets associated with transfers t^1 and t^2 are shown below the leaves. (bottom) The reconciled gene tree. Each leaf is annotated with its xenolog class in reference to t^2 (top row) and t^1 (bottom row). Genes $\widehat{g}_1$ and $\widehat{g}_2$ are separated by both transfers. Since $\widehat{g}_2 \in \Delta(g_d^1)$, $\mathcal{X}(\widehat{g}_1, g_2) = \mathrm{PX}$. In contrast, $\mathcal{X}(\widehat{g}_2, g_1) = \mathrm{IX}$ since $\widehat{g}_1 \notin \Delta(g_d^2)$. Xenolog classes for other genes are consistent with their relatedness in the gene tree (Theorem 2.1): $\mathrm{MRCA}(\widehat{g}_2, g_Y) <_G \mathrm{MRCA}(\widehat{g}_2, g_Z) <_G \mathrm{MRCA}(\widehat{g}_2, g_W) = \mathrm{MRCA}(\widehat{g}_2, g_X) <_G \mathrm{MRCA}(\widehat{g}_2, g_V)$ and $\mathrm{PX} <_X \mathrm{SDX} <_X \mathrm{SRX} <_X \mathrm{OX}$

However, g_U and g_Y do not share the super-transfer and thus, do not obey the hierarchy; $\mathrm{MRCA}(\widehat{g}, g_U) <_G \mathrm{MRCA}(\widehat{g}, g_Y)$, yet $\mathcal{X}(\widehat{g}, g_U) = \mathrm{SDX} >_X \mathcal{X}(\widehat{g}, g_Y) = \mathrm{PX}$.

Primary xenologs, including those connected by incomparable transfers, are more closely related than any other class of xenologs. Incomparable xenologs that are not Primary may fall anywhere in the hierarchy; that is, a given pair of Incomparable xenologs may be more closely related, or more distantly related, than a given pair of Sibling or Outgroup xenologs. Thus, the non-specific Incomparable xenolog class provides less information about relatedness than the specific Sibling and Outgroup classes, but guarantees a classification in which relatedness in the gene tree is consistent with the hierarchy.

The species tree hierarchy for single transfers (Supplementary Table S1) also holds for multiple comparable transfers summarized by a super-transfer, with one exception. When the recipient species of the super-transfer is a descendant of the donor species (as in Supplementary Fig. S2), Primary xenologs, with respect to this super-transfer, are more or equally related in the species tree than in the gene tree.

The species tree hierarchy is not guaranteed for multiple, incomparable transfers, even when the pair are classified as Primary xenologs. The reasoning for this is that the recipient of t^2 can be in any of the sets, D^1, R^1, or O^1, defined by t^1. Therefore the cenancestor of g_1 and g_2 can be in any species in V_S. Any relationship, even an incomparable relationship, is possible between the cenancestor and the ancestor containing $\mathrm{MRCA}(g_1, g_2)$.

3 Algorithms and implementation

The classification procedure for the xenolog classes described in Section 2 is shown as pseudocode in Supplementary Section S.4. We have implemented this procedure and integrated it in Notung 2.9, a freely available software package that implements gene tree-species tree reconciliation with transfers in a parsimony framework (Stolzer *et al.*, 2012).

Upon reconciling a gene tree with a species tree, Notung 2.9 generates a homology table, H, for all pairs of leaves in the gene tree. There may be more than one minimum-cost event history that reconciles the gene and species trees. A homology table is generated for each optimal, temporally feasible reconciliation reported. Transfers imply temporal constraints because the donor and recipient of a transfer must have co-existed; a reconciliation is temporally feasible if all temporal constraints imposed by the inferred transfers are mutually compatible. In particular, temporal consistency requires that s_d and s_r be incomparable to all transfers. Notung 2.9 reports all optimal reconciliations that are temporally feasible, up to a user-specified limit (Stolzer, 2012).

Homology tables can be viewed in the graphical user interface or exported from the command line in a tab-delimited, CSV, or HTML format. Row $H[\widehat{g}_i, \cdot]$ contains the homology relationships between reference gene, $\widehat{g}_i$, and all other genes in V_G. For orthologs and paralogs, $H[g_i, g_j] = H[g_j, g_i]$. For xenologs, $H[\widehat{g}_i, g_j] = \mathcal{X}(\widehat{g}_i, g_j)$ gives the xenolog class of g_j with respect to $\widehat{g}_i$, a reference gene that is the recipient of at least one transfer on the path from $\mathrm{MRCA}(\widehat{g}_i, g_j)$ to $\widehat{g}_i$. If there is also a transfer on the path from $\mathrm{MRCA}(g_i, g_j)$ to g_j, then $H[\widehat{g}_j, g_i] = \mathcal{X}(\widehat{g}_j, g_i)$ gives the xenolog class of g_i with respect to reference $\widehat{g}_j$. Otherwise, $H[\widehat{g}_i, g_j] = *$.

The classification procedure is generally applicable to reconciled gene trees and can be implemented in any reconciliation software package that enforces temporal consistency. When temporal consistency is not enforced, reconciliations with transfers between ancestor

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/33/5/640/2725487 by Edinburgh University user on 08 February 2019

and descendant species can arise. Since this scenario is similar to super-transfers that form a loop (Supplementary Fig. S2), the classification proposed here could easily be adapted for programs that do not enforce consistency.

4 Empirical results

Genomic study: As a proof of principle, we analyzed 13 623 gene families from a dataset of 65 genomes of Proteobacteria and Cyanobacteria (Latysheva *et al.*, 2012). Phylogeny was reconstructed as described in Supplementary Section S.5. To control for spurious inference of transfers due to phylogenetic error, weakly supported branches were rearranged using a species-tree aware method as described in Supplementary Section S.5.1. The resulting rooted, rearranged trees were then reconciled with the species tree with default costs ($C_\tau = 3$, $C_\delta = 1.5$, $C_\lambda = 1$). These costs are consistent with costs used in other recent phylogenomic analyses (David and Alm, 2011; Richards *et al.*, 2014), which were selected to minimize the total net change in genome content. The time required to reconcile the 13 623 trees, including generating all optimal reconciliations and testing them for temporal feasibility, was 7.25 min on an Intel Xeon 2.3 GHz processor (128 GB RAM). The computational complexity of calculating the homology table, once the gene tree has been reconciled, is negligible.

Homology tables were computed for the 13 194 trees possessing at least one temporally feasible solution. From these, homologs of all categories were tabulated. For families with more than one optimal reconciliation, the number of pairs in each category was averaged over all reported, optimal event histories.

Orthologs, paralogs and xenologs are all represented in this dataset, and every xenolog class is also observed (Fig. 6 and Supplementary Tables S2–S6). More than a quarter of homologous gene pairs were xenologs. Of these pairs, 85.7% are xenologs with only one reference gene, where all transfers on the path from the reference to its xenolog are mutually comparable. Of these xenologs, 60.2% are either Primary or Sibling Donor (para)xenologs; thus, the majority of the inferred xenologs are closer to the donor than the recipient.

Gene pairs separated by incomparable transfers are fairly rare compared with all types of xenologs separated by any number of transfers. Such pairs have two xenologs, one for each reference gene; at most one member of each pair can be classified as a Primary xenolog (PX), otherwise they are untyped (IX). The fraction of Incomparable xenologs for which the hierarchy provides no information is quite small: 72.0% of incomparable (para)xenologs are (PX, IX) pairs; the rest are (IX, IX) or (IX^P, IX^P) pairs.

Less than 1% of all xenologous pairs are autoxenologs, which could be due to preferential transfer of novel genes or a high incidence of xenologous gene displacement (Koonin *et al.*, 2001). Paralogs constitute 2.2% of all homologs, and paraxenologs are 4.8% of all xenologs. The low level of paralogy observed is consistent with prior reports that in prokaryotes transfer is a greater source of genetic novelty than duplication (Treangen and Rocha, 2011).

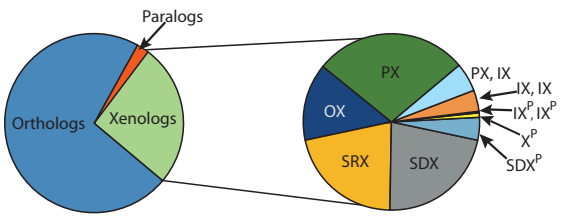

Fig. 6. *(left)* Proportions of orthologs, paralogs and xenologs (all classes) in the 13 194-tree bacterial dataset. *(right)* Proportions of xenolog classes

Interestingly, the vast majority of paraxenologs, 73.4%, are Sibling Donor paraxenologs. Recall that paraxenologs that diverged after the cenancestor of the transfer can be unambiguously classified and are always more closely related to the donor than to the recipient of the transfer. Paraxenologs that diverged before the cenancestor, i.e. closer to the root, cannot be assigned a specific class without breaking the hierarchy. As with Incomparable xenologs, the low fraction of untyped paraxenologs (X^Ps) suggests that, at least for this dataset, there are relatively few pairs for which it is impossible to extract some information from the xenolog classification.

Methodological factors may also contribute to the trends we observe. Gene families were inferred with OrthoMCL (Li *et al.*, 2003), which tends to place paralogous subfamilies in separate clusters. This could be a factor in the low level of paralogs, paraxenologs and autoxenologs in this study. It could also contribute to the preponderance of SDX^P pairs, relative to X^P pairs, as the tendency to break up paralogous subfamilies would result in relatively few inferred duplications near the root of the gene tree.

We considered to what extent the empirical parameters influenced the outcome of the analysis presented here. First, we investigated the impact of OrthoMCL on subsequent xenolog classification in a small set of curated families (Supplementary Section S.5.5). In most cases, the OrthoMCL clusters agreed with the curated family definitions. However, when OrthoMCL did split up paralogous subfamilies, the number and type of paraxenologs predicted changed dramatically.

In order to assess the impact of taxonomic breadth on our results, we also applied our classification procedure to two taxonomically-restricted subsets: families found only in the Cyanobacteria phylum (C: 49 species, 7485 trees) and families found only in the Synechococcales class (S: 30 species, 1429 trees), respectively. Orthologs, paralogs and all xenologs classes are present, and the observed trends are similar to those reported above for the full dataset (Supplementary Section S.5.4, Figs S8 and S9, and Tables S7–S16). Overall, the agreement between the full and restricted datasets suggests that our method is not highly sensitive to taxon sampling.

Finally, to probe the impact of event costs on xenolog classes observed in this study, we repeated this analysis with an increased transfer cost, $C_\tau = 4$, as described in Supplementary Section S.5.3. All xenolog classes were, again, observed. The higher transfer cost resulted in a moderate increase in the number of paralogs and paraxenologs of all classes, and a decrease in the number of non-paralogous xenologs inferred. The change in the relative frequencies of the other various classes was generally small (less than 15%) with one exception: the proportion of Outgroup xenologs decreased by more than 50%. The increase in para(xeno)logs and decrease in Outgroup xenologs, taken together, suggest that more duplications may be inferred near the roots of gene trees, when a higher transfer cost is used. Thus, in this analysis, the trade-off between duplications and transfers does not affect all xenolog classes equally.

BIO4 case study: To explore the connection between xenolog classes and protein function, we applied our approach to the *BIO4* gene family; several *BIO4* genes have been horizontally transferred and have been characterized experimentally (Hall and Dietrich, 2007). *BIO4* is part of the biotin (vitamin B7) biosynthesis pathway (Supplementary Fig. S11). Plants and some fungi possess a *BIO4* homolog that encodes a bi-functional enzyme that acts as both a 7,8-diaminopelargonic acid synthase (DAPAS) and a dethiobiotin synthetase (DTBS), steps 3 and 4 in the pathway, respectively. In bacteria,

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/33/5/640/2725487 by Edinburgh University user on 08 February 2019

the *BIO4* homolog only performs the DTBS function; the 3rd step is carried out by an unrelated protein. Unlike other fungi, however, the *BIO4* homolog in some yeasts (*Saccharomyces cerevisiae*, and its close relatives) also encodes a DTBS-only protein. Phylogenetic analysis shows that a horizontal transfer from bacteria to yeast replaced the ancestral bi-functional homolog (Hall and Dietrich, 2007). Using Notung 2.9, we reconciled the gene and species trees (Supplementary Figs S12 and S13) constructed by Hall and Dietrich (2007) and inferred xenolog classes (Fig 7 and Supplementary Fig. S14).

The hierarchical nature of the xenolog classification aids in the interpretation of the functional evolution of the family in this case study. The molecular function of yeast *BIO4* is closer to that of its Sibling Donor xenologs, which encode the DTBS-only enzyme, than its Sibling Recipient xenologs, which encode bi-functional enzymes. In contrast, the Sibling Recipient xenologs provide information about genomic context. The fact that the Sibling Recipient xenologs encode a bi-functional enzyme raises a red flag: the replacement of a bi-functional enzyme with a DTBS-only enzyme in yeast suggests loss of the DAPAS function. Either a different enzyme must be carrying out the DAPAS function or yeast no longer has a functional biotin synthesis pathway. In fact, the former is true; the DAPAS function is performed by an unrelated gene, which was also acquired horizontally (Hall and Dietrich, 2007).

In this example, a closely related gene (a DTBS-only enzyme) in a distantly related (α-proteobacterial) species is a better predictor of *BIO4* enzymatic function than a distantly related gene (the dual function homolog) in a closely related species (*Yarrowia lipolytica*). The distantly related homolog in a closely related species provides information about the genetic background; i.e. the genome could be lacking a gene encoding the DAPAS function. These insights are linked to the hierarchical structure of the xenolog classes and may represent general trends, suggesting hypotheses for future investigation. If it proves generally true, for example, that Sibling Donors are better predictors of molecular function and Sibling Recipients are better predictors of cellular context, then this system of xenolog classification could support large scale, automated analyses in comparative, evolutionary genomics.

5 Discussion

Distinguishing orthologs from paralogs, as well as the division of paralogs into subclasses based on the timing and nature of the events by which they arose, has proved to be a valuable analytical approach in molecular evolution, systematics, comparative genomics, and homology-based function prediction.

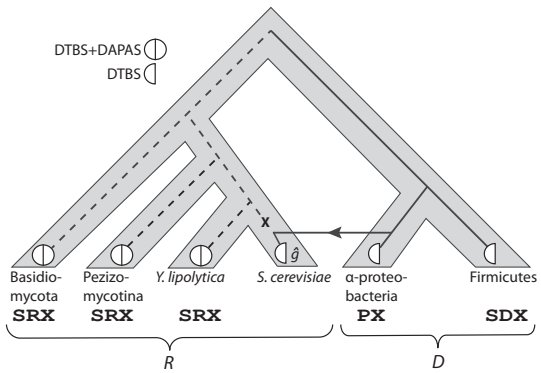

Fig. 7. Summary of the *BIO4* gene family event history. Dashed lines represent lineages with a putative dual-function DTBS + DAPAS enzyme; solid lines represent lineages with a putative DTBS-only function. With respect to the gene $\hat{g}$ in *S.cerevisiae*, all other fungal genes are SRX, α-proteobacterial genes are PX, and genes in Firmicutes are SDX

Here, we examine the challenges associated with the expansion of this framework to include horizontally transferred genes. The term 'xenolog' has been introduced to describe gene pairs related through horizontal transfer (Fitch, 2000; Gray and Fitch, 1983). However, the set of genes that share a history that includes at least one transfer encompasses a very broad set of relationships.

In this work, we propose subtypes that provide a more nuanced classification of xenologs. We provide formal rules for classification, given a reconciled gene tree with an arbitrary number of transfers and duplications. These rules have been implemented in Notung 2.9, a freely available phylogenetic reconciliation software package.

Phylogenetic reconciliation captures information about the historical association between genes and species, as well as the divergence events that characterize the xenologs in each class. A potential limitation of this approach is that it requires that species evolution be modeled as a tree. While some have argued against tree-like models, given the prevalence of horizontal gene transfer in bacteria, a tree can provide a useful heuristic, despite the reticulate nature of prokaryotic evolution (Mindell, 2013, and work cited therein).

As with most theoretical work on reconciliation, our classification assumes that the gene tree and the inferred events are correct. In practice, errors in gene tree reconstruction or incongruence due to unrecognized incomplete lineage sorting could lead to downstream errors in xenolog classification. Methods that account for phylogenetic uncertainty offer an approach to bridging this gap, and are an important direction for future work. For example, the xenolog classification proposed here could be embedded in a probabilistic reconciliation framework (e.g. Akerborg *et al.*, 2009), which would support an explicit and quantitative model of uncertainty.

Missing data is another potential source of error. If the dataset does not contain at least one descendant of the donor, a transfer will be inferred from a putative donor that is actually an ancestor of the donor species. When this occurs, some genes that are actually Sibling Donor xenologs may be incorrectly classified as Primary xenologs. The classification of Sibling Recipient, Outgroup, and all other Sibling Donor xenologs will be unaffected. Thus, classification errors due to missing taxa do not result in major changes in interpretation; these xenologs will still be correctly classified as being more closely related to the donor than to the recipient of the transfer.

Our classification is an extension of Fitch's classic framework and is based solely on information that can be extracted from gene tree–species tree reconciliation. The incorporation of other sources of information, such as synteny, sequence alignments, or structural comparison, could be used to develop richer accounts of xenology relationships. For example, Koonin *et al.* (2001) have proposed that horizontal gene transfer can result in the acquisition of a new gene family, expansion of an existing gene family, or allelic replacement without change in copy number.

Our classification provides a context for stating general hypotheses about the functional and evolutionary fates of different classes of xenologs. Since Sibling Donor xenologs are more closely related to the reference gene than Sibling Recipients, they may be more likely to share molecular functions with the reference gene. In contrast, the cellular environment of the reference gene may be more similar to that of Sibling Recipient xenologs. This could also convey information about the process of amelioration following transfer (Lawrence and Ochman, 1997). For example, the prokaryotic homologs of a fungal gene of prokaryotic origin are likely not informative with regard to the cellular compartment in which the encoded protein is

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/33/5/640/2725487 by Edinburgh University user on 08 February 2019

active. The functional fates of genes that have experienced both duplication and transfer is a largely unexplored question. Selective pressures are likely to change following both gene duplication (Lynch, 2007, and work cited therein) and horizontal gene transfer (Boto, 2010, 2016; Treangen and Rocha, 2011 and work cited therein). Little is known about the combined effect of these changes on rates of divergence and functional specialization.

Recent attempts to test the ortholog conjecture, which posits that orthologs are more functionally similar than paralogs, have demonstrated the challenges presented by confounding factors in high-throughput data, and especially in the use of ontologies (Chen and Zhang, 2012; Nehrt *et al*., 2001). Testing analogous xenolog conjectures will be even more challenging: probing all four xenolog classes would require large-scale, unbiased functional datasets for at least five species. Nevertheless, with the current pace of functional genomics, genomic-scale investigations of xenolog function are not far in the future.

Acknowledgments

We thank Minli Xu for his work preproccessing the bacterial genomic dataset, Rosie Alderson for help with biochemical information, Vivien Junker for assistance with phylogeny reconstruction, Annette McLeod for assistance with graphics, Han Lai for help coordinating the Notung implementation, and EaStCHEM for computational support via the EaStCHEM Research Computing Facility.

Funding

This material is based upon work supported by the National Science Foundation under Grant No. DBI-1262593 (to D.D.) and by Royal Society Research Grant No. RG061304 (to D.B.). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Conflict of Interest: none declared.

References

Akerborg,O. *et al*. (2009) Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 5714–5719.

Ali,R.H. *et al*. (2016) GenFamClust: an accurate, synteny-aware and reliable homology inference algorithm. *BMC Evol. Biol.*, **16**, 120.

Azad,R. and Lawrence,J. (2012) Detecting laterally transferred genes. *Methods Mol. Biol.*, **855**, 281–308.

Boto,L. (2010) Horizontal gene transfer in evolution: facts and challenges. *Proc. Biol. Sci.*, **277**, 819–827.

Boto,L. (2016) Accepting foreign genes. *J. Mol. Evol.*, **82**, 173–175.

Capra,J.A. *et al*. (2013) How old is my gene? *Trends Genet.*, **29**, 659–668.

Chen,X. and Zhang,J. (2012) The ortholog conjecture is untestable by the current gene ontology but is supported by RNA sequencing data. *PLoS Comput. Biol.*, **8**, e1002784.

Chen,X. *et al*. (2004) Operon prediction by comparative genomics: an application to the *Synechococcus* sp. WH8102 genome. *Nucleic Acids Res.*, **32**, 2147–2157.

David,L. and Alm,E. (2011) Rapid evolutionary innovation during an Archaean genetic expansion. *Nature*, **469**, 93–96.

Dickmeis,T. and Muller,F. (2005) The identification and functional characterisation of conserved regulatory elements in developmental genes. *Brief. Funct. Genomic Proteomic*, **3**, 332–350.

Durand,D. and Hoberman,R. (2006) Diagnosing duplications: can it be done? *Trends Genet.*, **22**, 156–164.

Duret,L. and Bucher,P. (1997) Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.*, **7**, 399–406.

Enright,A. *et al*. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.

Ermolaeva,M. *et al*. (2001) Prediction of operons in microbial genomes. *Nucleic Acids Res.*, **29**, 1216–1221.

Fitch,W. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–113.

Fitch,W. (2000) Homology: a personal view on some of the problems. *Trends Genet.*, **16**, 227–231.

Goodman,M. *et al*. (1979) Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst. Zool.*, **28**, 132–163.

Gray,G.S. and Fitch,W.M. (1983) Evolution of antibiotic resistance genes: The DNA sequence of a kanamycin resistance gene from *Staphylococcus aureus*. *Mol. Biol. Evol.*, **1**, 57–66.

Hall,C. and Dietrich,F.S. (2007) The reacquisition of biotin prototrophy in *Saccharomyces cerevisiae* involved horizontal gene transfer, gene duplication and gene clustering. *Genetics*, **177**, 2293–2307.

Huson,D. and Scornavacca,C. (2011) A survey of combinatorial methods for phylogenetic networks. *Genome Biol. Evol.*, **3**, 23–35.

Koonin,E.V. *et al*. (2001) Horizontal gene transfer in prokaryotes: Quantification and classification. *Annu. Rev. Microbiol.*, **55**, 709–742.

Latysheva,N. *et al*. (2012) The evolution of nitrogen fixation in cyanobacteria. *Bioinformatics*, **28**, 603–606.

Lawrence,J.G. and Ochman,H. (1997) Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.*, **44**, 383–397.

Li,L. *et al*. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178–2189.

Lynch,M. (2007) *The Origins of Genome Architecture*. Sinauer Associates Inc., Sunderland, MA.

Marcotte,E. *et al*. (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.

Mindell,D. (2013) The tree of life: Metaphor, model, and heuristic device. *Syst. Biol.*, **62**, 479–489.

Nadeau,J. and Sankoff,D. (1998) Counting on comparative maps. *Trends Genet.*, **14**, 495–501.

Nakhleh,L. (2010) Evolutionary phylogenetic networks: models and issues. In: Heath,L. and Ramakrishnan,N. (eds.) *The Problem Solving Handbook for Computational Biology and Bioinformatics*, pp. 125–158. Springer, New York, NY.

Nakhleh,L. (2013) Computational approaches to species phylogeny inference and gene tree reconciliation. *Trends Ecol. Evol.*, **28**, 719–728.

Nehrt,N.L. *et al*. (2001) Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Comput. Biol.*, **7**, e1002073.

O'Brien,S. *et al*. (1997) Comparative genomics: lessons from cats. *Trends Genet.*, **10**, 393–399.

Pellegrini,M. *et al*. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.*, **96**, 4285–4288.

Price,M. *et al*. (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res.*, **33**, 880–892.

Ramos,O.M. and Ferrier,D.E.K. (2012) Mechanisms of gene duplication and translocation and progress towards understanding their relative contributions to animal genome evolution. *Int. J. Evol. Biol.*, **2012**, 846421–31.

Richards,V.P. *et al*. (2014) Phylogenomics and the dynamic genome evolution of the genus *Streptococcus*. *Genome Biol. Evol.*, **6**, 741–753.

Shi,G. *et al*. (2011) MultiMSOAR 2.0: an accurate tool to identify ortholog groups among multiple genomes. *PLoS One*, **6**, e20892.

Simillion,C. *et al*. (2004) Recent developments in computational approaches for uncovering genomic homology. *Bioessays*, **26**, 1225–1235.

Song,N. *et al*. (2007) Domain architecture comparison for multidomain homology identification. *J. Comput. Biol.*, **14**, 496–516.

Song,N. *et al*. (2008) Sequence similarity network reveals common ancestry of multidomain proteins. *PLoS Comput. Biol.*, **4**, e1000063.

Sonnhammer,E. and Koonin,E. (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.*, **18**, 619–620.

Stolzer,M. (2012). Phylogenetic Inference for Multidomain Proteins. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/33/5/640/2725487 by Edinburgh University user on 08 February 2019

Stolzer,M. *et al.* (2012) Inferring duplications, losses, transfers, and incomplete lineage sorting with non-binary species trees. *Bioinformatics*, **28**, i409–i415.

Treangen,T.J. and Rocha,E.P.C. (2011) Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet.*, **7**, e1001284.

Van de Peer,Y. (2004) Computational approaches to unveiling ancient genome duplications. *Nat. Rev. Genet.*, **5**, 752–763.

Westover,B. *et al.* (2005) Operon prediction without a training set. *Bioinformatics*, **21**, 880–888.

Wu,C. *et al.* (2003) Protein family classification and functional annotation. *Comput. Biol. Chem.*, **27**, 37–47.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/33/5/640/2725487 by Edinburgh University user on 08 February 2019