life

MDPI

Article

Functional Annotations of Paralogs: A Blessing and a Curse

Rémi Zallot, Katherine J. Harrison, Bryan Kolaczkowski and Valérie de Crécy-Lagard *

Department of Microbiology and Cell Science, Institute of Food and Agricultural Sciences, University of Florida, Gainesville, FL 32611, USA; remizallot@ufl.edu (R.Z.); katherinejh@ufl.edu (K.J.H.); bryank@ufl.edu (B.K.)
* Correspondence: vcrecy@ufl.edu; Tel.: +1-352-392-9416

Academic Editor: Niles Lehman
Received: 1 July 2016; Accepted: 2 September 2016; Published: 8 September 2016

Abstract: Gene duplication followed by mutation is a classic mechanism of neofunctionalization, producing gene families with functional diversity. In some cases, a single point mutation is sufficient to change the substrate specificity and/or the chemistry performed by an enzyme, making it difficult to accurately separate enzymes with identical functions from homologs with different functions. Because sequence similarity is often used as a basis for assigning functional annotations to genes, non-isofunctional gene families pose a great challenge for genome annotation pipelines. Here we describe how integrating evolutionary and functional information such as genome context, phylogeny, metabolic reconstruction and signature motifs may be required to correctly annotate multifunctional families. These integrative analyses can also lead to the discovery of novel gene functions, as hints from specific subgroups can guide the functional characterization of other members of the family. We demonstrate how careful manual curation processes using comparative genomics can disambiguate subgroups within large multifunctional families and discover their functions. We present the COG0720 protein family as a case study. We also discuss strategies to automate this process to improve the accuracy of genome functional annotation pipelines.

Keywords: paralog; ortholog; homolog; annotation; genome context; phylogeny; signature motif; genome; multigene families

1. Introduction

The advent of low-cost, high-throughput sequencing technologies has created the global capacity to generate complete genome sequences at an unprecedented rate, with tens of thousands of organisms currently at various stages in genome sequencing pipelines [1–3]. This rate is expected to continue increasing, at least for the foreseeable future, as technologies continue to improve. This genome sequencing "revolution" has profound implications. Classically, analysis methodologies focused heavily on rigorous small-scale experimentation and statistical approaches for boosting power to detect differences in small data sets [4–6]. The deluge of genome sequence data makes small-scale experimental approaches impractical, while the coincident increase in computational power has made high-throughput bioinformatics analyses an attractive alternative [7–9]. Out of necessity, biologists are relying increasingly on bioinformatics-based predictions to generate information [10].

Although the development of efficient algorithms for identifying protein-coding sequences (CDSs) is still an active area of bioinformatics research, "gene finding" has improved greatly in recent years, particularly for prokaryotes [11–15] (even if the calling of start sites is still often problematic [16]). However, the functional annotation of CDSs is particularly difficult to automate. Current state-of-the-art functional annotation pipelines integrate multiple types of evidence [17–19], but unfortunately the quality of functional annotations remains generally poor [20–23] and is highly dependent on resource-intensive manual curation [24,25].

Early pioneers in function annotation were quick to identify potential problems with large-scale annotation efforts [26–28], and misannotation is a growing concern among the general research community, as misannotated genes can have a "ripple effect" impacting diverse areas of biological inquiry [29–31]. By different measures, 10%–25% of functional calls are wrong, even in very small bacterial genomes [32]. High-throughput functional calls can be incorrect due to a variety of factors [33,34], but the most common errors (>85%) are over-annotations, in which a gene is given a specific but incorrect function [21,32,35]. Once made, functional annotation errors can be difficult to correct in large sequence databases and can often "propagate" themselves to newly sequenced genomes through "(mis)annotation transfer" [36–39].

Multigene families—defined as having more than one member of the family in a given genome—can be notoriously difficult to correctly annotate, primarily due to over-annotation problems [21]. Large multigene families are also a potential goldmine for discovering new functions, because many of the enzymes within a family are likely to perform chemically similar reactions on similar substrates, reducing the functional space to explore in comparison with totally unknown enzymes (see Table 1 for examples). Overcoming the challenges of annotating multigene families and exploiting their potential for discovering new gene functions will likely require integrating large-scale and small-scale annotation methods. To date, the majority of genome-scale functional annotation pipelines have been developed in isolation from small-scale annotation efforts or experimental validation. While large-scale approaches are inherently limited due to the need for computational efficiency and generalizability, many smaller-scale efforts are combining advanced bioinformatics tools with experimental validation to disambiguate multigene families and are producing high quality functional annotations and discovering novel or missing functions in the process [40–43]. In our view, "closing the loop" between these different approaches has a tremendous potential to improve the quality of functional annotations stored in major online databases. A recent call for functional microbiologists to be more involved in the genome annotation process stresses this point [44].

Table 1. Examples of paralogs with different functions elucidated by comparative genomic approaches.

COG [1]	Paralog Subfamilies			Ref.
0780 and 0302	QueF (Queuosine synthesis)	FolE (Tetrahydrofolate synthesis)	QueF-Like (Aracheosine synthesis)	[45,46]
1539	FolX (Tetrahydromonapterin synthesis)	FolB (Tetrahydrofolate synthesis)		[47]
1028 and 0262	FolM (Tetrahydromonapterin synthesis)	FadG (fatty acid synthesis)	FolA (Tetrahydrofolate synthesis)	[47]
0009	YciO (Unknown function)	YrdC/TsaC (t^6A synthesis)		[48]
1509	YjeK (Protein modification)	LamB (Lysine degradation)		[49]
2269 and 1190	YjeA (Protein modification)	LysRS (Protein synthesis)		[49]
0354 and 0404	YgfZ (Iron-sulfur cluster repair)	GcvT (One carbon metabolism)		[50]
2102	Dph6 (Diphtamide synthesis)	DUF71-B12 group (function unknown, B12 salvage)		[51]
5424	PqqC (PQQ synthesis)	CT610 (Para-aminobenzoate synthesis)		[52]
1478	CofE (F420 synthesis)	CT611 (Tetrahydrofolate synthesis)		[53]
1901	TrmY Archaeal m1Psi54 methylase	Bacterial unknown methylase		[54]
0720	PTPS-I (QueD, Queuosine synthesis)	PTPS-II (Biopterin synthesis)	PTPS-III (Folate synthesis)	[55]
0523	15 subfamilies identified			[56]
0212	5-formyltetrahydrofolate cycloligase (5-FCL)	Thiamin metabolism		[57]

[1] Data obtained from the Clusters of Orthologous Groups of proteins (COGs) database, 2003 COGs, 2014 update [58]. Please note that if some paralog pairs are currently mapping to more than one COG, it was not the case at the time of the reported studies.

In this work, we argue that focusing on individual genomes that have readily identifiable paralogs and examining similarities and differences in the corresponding gene neighborhoods is an efficient way to identify and separate similar genes with different functions. We propose an annotation workflow that "bootstraps" information gained by identifying specific genomes harboring functionally differentiated paralogs to potentially improve annotations of related genes in other genomes (Figure 1). We expect this type of annotation workflow could both guide the experimentalist and improve automated annotation of large multigene families.

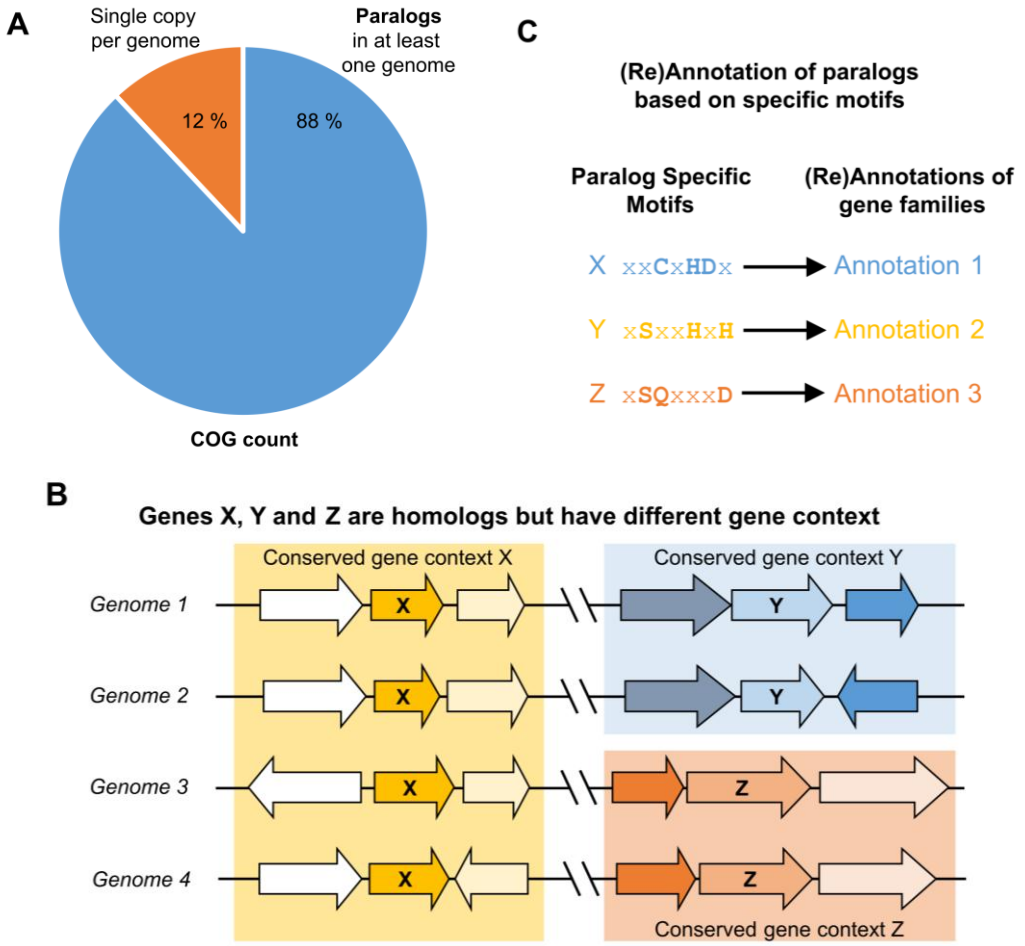

Figure 1. Differential paralog clustering can be used to identify non-isofunctional genes in multigene families. (**A**) Although single-copy genes can be accurately annotated by current high-throughput pipelines, they comprise a small percentage of the genome's total gene count (12%, according to the National Center for Biotechnology Information (NCBI) Clusters of Orthologous Groups of proteins (COGs) database (2014 update) [58]). Genomes for which there is more than one homolog among a multigene family suggest the existence of paralogs among this family, and transfer of annotation based on homology is problematic among paralogs that can be non-isofunctional; (**B**) Paralogs with different flanking genes (genomic contexts) conserved across multiple species can be inferred as having different functions, even if their functions are not known; (**C**) Once paralogs are differentiated into provisional isofunctional groups by conserved genomic context, patterns of sequence conservation within and among isofunctional groups can be used to derive paralog-specific motifs, which can be used to annotate additional members of the isofunctional group for which there is no physical clustering conservation.

2. Protein Function and Evolution

Protein functional annotation can vary widely in the degree of precision and can be defined in multiple ways, including phenotype, cellular localization, ligand interactions or interaction partners [59,60]. Efforts to unify and standardize functional annotations using controlled hierarchical vocabularies—although controversial and incomplete—have provided an objective framework for functional annotation that can be computed automatically [61,62]. However, these approaches are only beginning to be incorporated into genome annotation pipelines, requiring older genomes to be re-annotated [63,64]. For the purpose of this study, we will use a very strict definition of function in which both the molecular function and its functional context have been elucidated (sometimes called a two-dimensional annotation [65]). For an enzyme, this would be the Enzyme Commission number (EC number) and the biological pathway the enzyme participates in.

As it is impossible to experimentally verify the function of every newly sequenced gene, researchers have sought to elucidate the "evolutionary rules" governing when it is—and when it is not—appropriate to "transfer" specific functional information from a gene of known function to a related gene whose function is not known. We would expect genes with very similar sequences to have similar functions, while genes with more dissimilar sequences are more likely to have different functions. Indeed, the earliest functional annotation algorithms simply used sequence similarity searches to predict protein function by "annotation transfer" [66,67], but these generally led to unacceptably high error rates [32,33]. The fact that gene duplication and speciation events followed by mutation can lead to functional changes means that proteins with high sequence similarity might not have the same function [68].

More sophisticated approaches attempt to identify specific evolutionary patterns that may correlate with functional conservation or divergence [69–71]. The evolutionary relationship between two genes in a protein family (i.e., "homologs", genes that have descended from a common ancestor) can be broadly classified into two types: "orthologs" are two genes from different species that derived from a single gene in the last common ancestor of the species, while "paralogs" are two genes (within the same or different species) that derived from a single gene that was duplicated within the genome of some species [72].

Current thinking and empirical evidence supports the conclusion that, in general, orthologous genes have equivalent functions more often than paralogs [73]. This has led to numerous approaches for identifying groups of orthologs (e.g., OrthoMCL, OMA and eggNOG [74–76]) and transferring functional annotations among members of an ortholog group. Some researchers have gone so far as to "redefine" the term "orthologs" to mean "genes with equivalent functions" [77]. In agreement with R.A. Jensen [78], we feel that the "functional" definitions of "ortholog" and "paralog" are unwarranted and invite confusion, and we retain the original "evolutionary" definitions in this work. There are certainly examples of orthologs that have different functions in different species [78–81], and functional differentiation between orthologs from different species (particularly those separated by large evolutionary distances and/or major changes in lifestyle or habitat) may be common enough that simply transferring functional annotations among orthologs—without considering additional information—is likely to lead to potential errors [82].

It may be comforting to assume that the converse relationship may hold more universally: perhaps paralogous genes in the same species always have different functions. However, duplicate genes with equivalent functions can be retained to supply specific gene or protein dosage, providing a scenario in which paralogous genes have equivalent functions [27,83–86]. In multicellular eukaryotes, this can occur when gene regulatory functions are partitioned between daughter genes (i.e., the same protein is expressed by different genes in different tissues, cell types, or different periods of time/development). This regulatory split also occurs in bacteria [87,88]. Hence, orthology/paralogy—in and of itself—is insufficient as a universal means for dividing related genes into functional classes, although it can provide crucial information as part of a more comprehensive strategy.

3. Identifying Orthologs and Paralogs in Practice

In theory, identifying orthologs and paralogs with phylogenetic trees should be straightforward. "Ortholog" and "paralog" are evolutionary concepts that are well-defined from a phylogenetic perspective; given a gene family tree and a species tree that are both known with absolute certainty, orthologs and paralogs can be identified without error [89,90]. However, neither gene nor species trees are ever known with absolute certainty, and current gene-species tree reconciliation methods have limited capacity to incorporate phylogenetic uncertainty or complex realistic evolutionary scenarios [49,91–94]. Even sophisticated phylogenetic reconstruction methods can be subject to methodological bias under some conditions [95–98], and strongly supported errors in either the gene tree or the species tree can produce radical errors in ortholog/paralog identification [97]. Even if the phylogenies of extant genes and species could be reconstructed without error, certain patterns of differential gene losses can make paralogous genes erroneously look like orthologs (Figure 2).

Even with these caveats, recent phylogeny-based functional annotation methods have performed very well, compared to competing approaches [99,100]. Unfortunately, rigorous phylogenetic analysis does not scale well computationally, making it difficult to apply the most accurate tree reconstruction algorithms to large data sets [71,101]. Fast tree reconstruction algorithms do exist but are more prone to errors in both topology and confidence assessment than more computationally intensive methods [97,102–104]. Even with the dramatic recent speed-up of phylogenetic reconstruction algorithms, these methods are likely to remain too computationally expensive to deploy across the growing wealth of whole-genome sequence data [71,105].

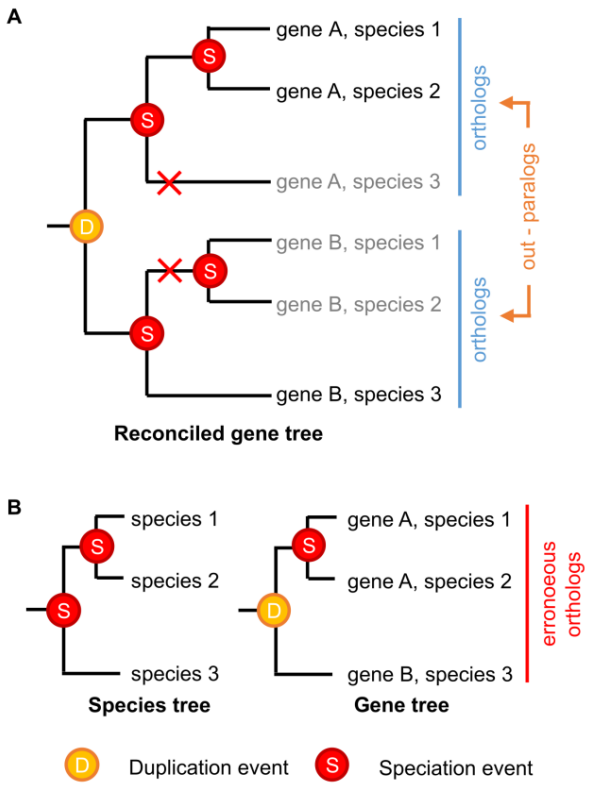

Figure 2. Differential gene loss can obscure the identification of paralogous genes. (**A**) A hypothetical example showing the historical evolutionary history of a simple multigene family across three species. Gene duplication and speciation events are indicated, with red crosses indicating gene losses. Grey names indicate unobservable "extinct" genes; (**B**) Even if the correct species tree were known with certainty, differential loss of paralogs across the species in (**A**) would result in the indicated gene tree, which is "correct," given extant sequence data, but cannot observe the "extinct" genes. Gene-species tree reconciliation results in the erroneous grouping of paralogous genes A and B as orthologs.

Numerous approaches have been devised that use computationally efficient sequence similarity comparisons, often combined with other sources of information, to approximate phylogenetically based ortholog/paralog identification while improving throughput [106,107]. At their core, most of these approaches use patterns of pairwise sequence similarity among genes to infer patterns of orthology/paralogy. Historically, Koonin and co-workers were the first to implement such an approach with their Clusters of Orthologous Groups (COGs), based on best birectional BLAST hits [108]. Many other approaches to ortholog/paralog identification use pairwise sequence similarity as a starting point, differing mainly in how pairwise similarity networks are divided into ortholog/paralog groups [74,109,110]. For example, Sequence Similarity Networks (SNNs) have been used successfully to identify paralogs and isofunctional groups in large multigene families [111]. A major advantage of SSNs over traditional phylogenetic trees is that relationships between protein sequences are easier to visualize, allowing larger numbers of sequences to be examined, and it is generally easier to automatically analyze similarity networks to divide them into orthology groups.

It is well recognized that patterns of sequence similarity only weakly approximate phylogenetic relationships [71,112–114], so these approaches may be error-prone in orthology/paralogy inference. As shown in Table 1, some known paralogs have been erroneously grouped in the same COG, and some orthologs are erroneously split into different COGs. To estimate how frequently COG families contained paralogs, we analyzed the genomic distribution of the latest COG set that lists 4695 COGs across 711 prokaryotic and archaeal genomes (2003 COGs, 2014 update) [58]. This analysis revealed that only 12% of COGs have only one member in any given genome. For 88% of COGs, a genome could be found that encodes at least two paralogous members of the same COG, reinforcing that sequence similarity alone is insufficient to reliably group genes into sets of orthologs (Figure 1). Of course, if errors in orthology/paralogy inference by sequence similarity happen to correlate with the historical patterns of functional divergence among orthologs/paralogs, sequence similarity methods may perform surprisingly well at predicting function, albeit because they may be fortuitously "biased in the right direction" [114,115].

4. Lessons from Comparative Genomics

Comparing similarities and differences among genomes can be a powerful tool for functional annotation. Not long after the release of the first whole genome sequence, visionary pioneers like Koonin [116] and Overbeek [117,118] provided a framework for using comparative genomic methods to improve gene annotations. In general, the "comparative genomic framework" combines evolutionary relationships with additional information gleaned from whole-genome comparisons to improve predictions about similarities and differences in gene function.

A gene's function is impacted by the specific metabolic network or subsystem within which it works as well as its role within that network or subsystem [65,119–121]. One of the underlying principles of the comparative genomic framework is that various types of information, such as physical clustering, gene fusions, co-regulation or regulons and phylogenetic co-distribution, can be used to reconstruct a gene's potential role within a network or subsystem [122–124]. These association networks can be used to efficiently identify and annotate paralogs that have diverged in function, particularly in the case of prokaryote genomes, where genes operating in the same subsystem are often physically clustered. As the first dozens of bacterial genomes were sequenced, it became apparent that conserved physical clustering of genes (conserved genome context) in phylogenetically distant genomes is a strong suggestion of functional association [118,125–127]. The underlying causes can vary but include maintaining co-regulation of functionally related genes [128–130] and facilitating horizontal transfer of multigene functional units [131]. As more genomes have become available, physical clustering and gene fusions—which can also be considered a form of physical clustering [125,132]—have become a robust measure of functional association [127,133,134] used in various integrated annotation platforms (e.g., STRING [135], SEED [118,136], Microbesonline [137], Syntax [138], MaGE [139]).

Examples of successful separation and annotation of paralogs within multigene families by our group are given Table 1. In all cases, physical clustering analysis was key to the functional separation of the multigene family into putative isofunctional subgroups. These examples are drawn from families with only two to three subgroups, but physical clustering can also be useful for functionally characterizing more complex gene families [140,141]. For example, a single bacterial genome can contain dozens of individual genes from the Nudix hydrolase family. Members of this family are all pyrophosphatases [142] but can vary widely in preferred substrate and are very difficult to functionally separate by sequence similarity alone. However, substrate specificity can be accurately predicted by incorporating physical clustering information [71,105,142,143].

Combining phylogenetic or sequence similarity information with additional comparative-genomics information—gene clustering, protein-protein interactions, regulon membership, presence/absence of specific functional domains, structure, chemistry and "signature" sequence motifs—has produced some of the most effective examples of reliably separating functional groups within multigene families [43,106,144]. These approaches are extremely effective when performed on a single gene family, often in combination with experimental validation [145,146]. Although these integrative approaches to functional annotation currently rely on expert understanding of the specific gene family under examination and careful manual curation, recent advances have improved prospects for future complete automation of integrative function-prediction analyses, with impressive results. For example, the Structure-Function Linkage Database (SFLD) analyzes 12 superfamilies that cover over 300 functional families [147], and the CATH-Gene3D and FunFams resources encompass more than 2500 families [148,149]. These "middle level" resources fill an important gap between rigorous small-scale experimentation and efficient high-throughput annotation pipelines; they are likely to improve the accuracy of functional annotations for the gene families they cover but are not currently deployed at a whole-genome scale. Future expansion of these and similar resources that carefully integrate multiple data sources using sophisticated approaches that mimic the manual curation process are expected to dramatically improve the overall quality of gene functional annotation.

5. The COG0720 Case Study

The COG0720 family is a good example of how comparative genomics can differentiate paralogs, even in complicated cases with several functional divergence events. The founding member of COG0720 is 6-pyruvoyl-tetrahydropterin synthase (PTPS). This enzyme catalyzes the second step of tetrahydrobiopterin (BH4) biosynthesis. BH4 is a cofactor used by aromatic acid hydroxylases in animals and certain bacteria [150]. It is synthesized from guanosine triphosphate (GTP) in three enzymatic steps (Figure 3A). The first step is shared with the folate biosynthesis pathway and catalyzed by GTP cyclohydrolase I (FolE). PTPS (EC 4.6.1.10) then produces 6-pyruvoyl-tetrahydropterin (PTP) from dihydroneopterin-triphosphate (DHN-P3). The last step is catalyzed by sepiapterin reductase (SR) (Figure 3A). The first cloned and sequenced PTPS-encoding gene was from rat [151]. Some bacteria known to have a BH4 biosynthesis pathway, like *Synechococcus* sp. *PCC 7942*, have two PTPS homologs. PTPS-I (BLAST *e*-value 6×10^{-31} to the rat homolog) was shown not to be involved in BH4 synthesis, as deletion of the corresponding gene did not affect the level of synthesis of the cofactor [152], while PTPS-II (BLAST *e*-value 5×10^{-20} to the rat homolog) was shown to have canonical PTPS activity in vitro [152] (Figure 3A) (default BLAST parameters were used). This is a clear case of bacterial paralogs with different functions and also demonstrates that the best BLAST hit can be the wrong paralog.

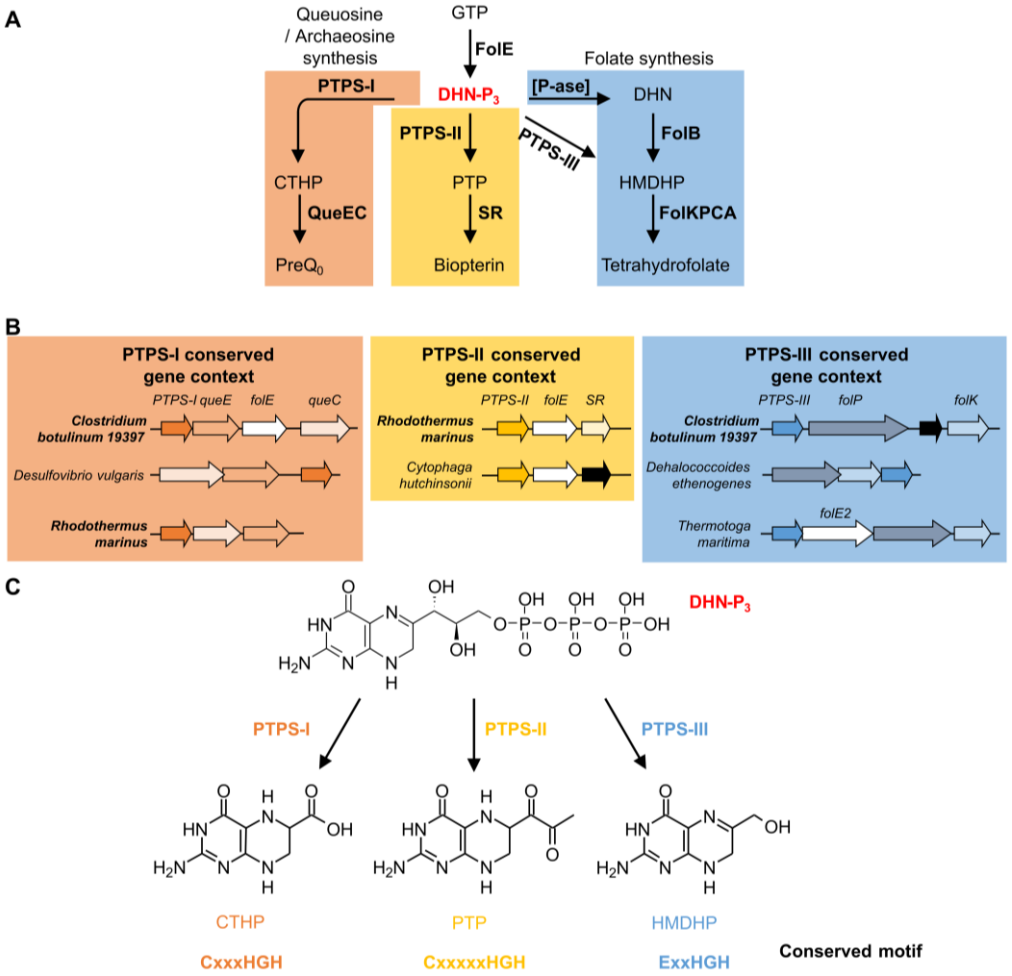

Figure 3. Functional roles of different PTPS subfamilies. Biosynthesis pathways in which PTPS-I, II and III are involved (**A**) and example of conserved gene context of PTPS-I, II and III linking them with these pathways (**B**); Note that *Clostridium botulinum 19397* and *Rhodothermus marinus* are among the organisms having more than one PTPS gene belonging to different conserved gene contexts. Specific reactions catalyzed by PTPS-I, II and III, and conserved motifs identified (**C**). Abbreviations: GTP: guanosine triphosphate; FolE: GTP cyclohydrolase I; FolE2: GTP cyclohydrolase II; DHN-P3: dihydroneopterin-triphosphate; CTHP: 6-Carboxytetrahydropterin; PTP: 6-pyruvoyl-tetrahydropterin; HMDHP: 6-hydroxymethyldihydropterin; PreQ0: 7-cyano-7-deazaguanosine; QueE: 7-carboxy-7-deazaguanine synthase; QueC: 7-cyano-7-deazaguanine synthase; SR: sepiapterin reductase; DHN: dihydroneopterin; [P-ase]: phosphatase; FolB: Dihydroneopterin aldolase; FolK: 6-Hydroxymethyl-7,8-dihydropterin pyrophosphokinase; FolP: Dihydropteroate synthase; FolC: Dihydrofolate:folylpolyglutamate synthase; FolA: Dihydrofolate reductase.

Although COG0720 proteins are found in most bacteria, only ~50 are predicted to take part in the BH4 biosynthesis pathway based on literature review and co-occurrence with the SR protein (see "PTPS paralogs" subsystem http://pubseed.theseed.org//SubsysEditor.cgi?page=ShowSubsystem&subsystem=PTPS_paralogs) [119]. Hence, metabolic reconstruction could have flagged the 6-pyruvoyl-tetrahydropterin synthase annotation for most members of the COG0720 family as questionable. However, this correction did not occur, and most COG0720 genes were incorrectly annotated. In the initial COG database [116], COG0720 contained 55 proteins in 43 archaeal or bacterial genomes. Even though eight genomes encoded multiple COG0720 copies, all COG0720 members were annotated as 6-pyruvoyl-tetrahydropterin synthases. Our more recent analysis of the Joint Genome Institute (JGI) Integrated Microbial Genomes (IMG) database (Table 2) found that ~15 percent of

bacterial genomes contain more than one COG0720 member. The COG0720 family is a clear case of over-annotation caused by transfer of annotations between functionally divergent paralogs.

Table 2. COG0720 genome count from diverse genomes [1].

Domain	Genome Count with COG0720	COG0720 Gene Count	Genomes with COG0720 Paralogs	Total Genomes
Archaea	561	698	134	771
Bacteria	636	821	164	1056
Eukaryota	60	65	5	220

[1] Data extracted from IMG using the search function as of 30 November 2015. Genome list is provided as supplementary information in Table S1.

Using different types of comparative genomics evidence, we were able to predict, and then experimentally verify, that at least two other COG0720 sub-families (PTPS-III and PTPS-I) perform different reactions in other pathways (Figure 3A) [55]. PTPS-III is part of the tetrahydrofolate pathway in *Plasmodium falciparum* and various bacteria [55,153,154]. In specific Archaea, members of the PTPS-III group are involved in tetrahydromethanopterin biosynthesis, but in these cases the substrate might be the monophosphate and not the triphosphate form [51]. PTPS-I is involved in the synthesis of archaeosine and queuosine in tRNA [55,155,156]. In this case, physical clustering evidence (Figure 3B) was sufficient to link the three PTPS gene families to their respective biochemical pathways.

Multiple alignments and structure comparisons allowed us to identify diagnostic signature motifs for the subfamilies (Figure 3C) [55]. Enzymes belonging to the PTPS-II subfamily harbor the motif CxxxxxHGH. The motif found in PTPS-III enzymes ExxHGH is not mutually exclusive with the one found in PTPS-I enzymes QueD (CxxxHGH), i.e., an enzyme with a CExxHGH can exist. Some organisms indeed have a PTPS-I/III hybrid version of the enzyme that can function in both folate and Q pathways [55]. In addition to the examples presented here, other subfamilies have been identified: PTPS-IV is involved in the synthesis of the archaeal cofactor F420 (de Crécy-Lagard and R. White unpublished); PTPS-VI has been validated as a folate enzyme [51]; and the function of PTPS-V is still unclear. Each of these subfamilies harbors specific signature motifs [55]. In summary, the functional complexity of the COG0720 family—which appears only separable using a combination of comparative techniques—makes it a good example to test automated annotation pipelines.

6. Current Methods for Paralog Annotation in Genome Annotation Pipelines

The above discussion suggests that careful methods for separating multigene families into functional subgroups have become highly effective. However, these robust approaches are not regularly incorporated into automated genome annotation pipelines, so the "best" methods are not being used to annotate most genes. Indeed, in most cases, after a given CDS has been predicted, it is aligned using sequence similarity to either a reference genome or to protein sequences available in large databases such as Uniprot. If a high scoring homolog is identified, the annotation is typically transferred, often without any significant additional analysis [67]. This process is far from perfect, as mentioned above and discussed in detail by Richardson et al. [67]. Issues include the use of references with outdated or misspelled annotations, inconsistent annotations among close orthologs, protein fusions and, of course, paralogs.

A recent improvement in annotation pipelines has been to include family relationships (e.g., OrthoMCL, CDD, COG, EggNOG, Pfam, TigerFam, FigFam) as part of the annotation [67]. This is very useful for identifying fusion proteins that are a well-known misannotation problem [67,132]. However, unless domain families have been specifically designed to separate paralogs (such as FigFam, used by the RAST pipeline), domain annotations do not disambiguate paralogs; on the contrary, domain-family annotations typically group paralogs together. We evaluated if and how paralogs were separated in the most popular genome annotation pipelines, including: IMG [17,157,158],

RAST [13,159,160], NCBI prokaryotic annotation [161], EnsemblBacteria [162] and MaGE [139]. We specifically examined two COG0720 proteins from *Pyrococcus furiosis* DSM 3638: PF0219 and PF1278. The first is a (preQ0) synthesis enzyme (PTPS-I) and the second a folate synthesis enzyme (PTPS-III).

The NCBI Prokaryotic Genome Annotation Pipeline relies on protein cluster membership to transfer functional annotations [161,163]. This method is not designed to separate paralogs without additional curation. We found that PF0219 was correctly annotated, but PF1278 was not, and it remains hypothetical (Table 3). IMG uses a combination of methods to assign IMG terms as a controlled vocabulary for describing functional roles [164,165], but IMG relies on BLAST or family relationships that do not explicitly separate paralogs. Interestingly, IMG has developed a series of integrated tools to identify identical functional annotations within a genome and flag paralogs [166]. However, these tools are not directly used in the annotation pipeline and are not easily accessible from a given gene page through the IMG web browser interface. Many genomes have been re-annotated in IMG [157], but if PF0219 was correctly annotated as a QueD enzyme in the first annotation, the re-annotation introduced a functional ambiguity that obscured the PF0219 annotation. For PF1278, the function was not called in the first IMG annotation, and it was miscalled in the re-annotation (Table 3).

RAST uses a curated set of protein families (FigFams) as a source of annotations for submitted sequences. FigFams are defined as sets of isofunctional homologs based on sequence similarity and conserved genomic context across closely related genomes [159,167]. This family definition should allow separation of paralogs, and the two test proteins were indeed correctly called (Table 3). We should note that, because of our long association with the SEED database that is the basis for many RAST/PATRIC annotations, these genes were actually manually curated by our group. If the initial curation of a few members of the FigFam is not correctly done, the paralog families will not be separated.

MaGe (Magnifiying Genomes) annotates proteins based on sequence similarity searches against non-redundant protein sequences, protein family identification and using HAMAP (High quality Automated and Manual Annotation Proteins) profiles [168]. Gene contexts are said to be taken into account in annotation calls [139], so in theory this should allow paralog separation. However, it does not appear that this information is integrated at the level of functional annotation, as neither of the two test proteins was correctly annotated in MaGe (Table 3). Finally, Ensembl Genomes uses the InterProScan 5 pipeline for protein annotation [169,170], which primarily uses HMM (Hidden Markov Model) signatures for identifying protein families [170]. If paralogs have not been separated by the HMM signatures beforehand, they will not be separated in any new annotation. In addition, the Ensembl platform provides a gene-oriented phylogenetic resource that should allow paralog identification (see [171]). However, it is not available for every gene entry and does not appear to be used for annotation purposes, as neither of the two test proteins was correctly annotated in Ensembl (Table 3).

In summary, we found a major discrepancy between the tools available in large integrated databases, which should in theory be capable of disambiguating or at least flagging paralogs, and the actual annotation pipelines, which rarely appear to take advantage of these tools to transfer functional annotations. In practice, large-scale functional annotation is still mainly based on similarity scores alone. The identification of paralogs remains primarily a low-throughput manual process.

Table 3. Annotations and information available in diverse databases for the genes belonging to the classes PTPS-I and PTPS-III of *Pyrococcus furiosus* DSM 3638 [1].

Annotation	Database Identifier	Annotation	Database Identifier	Annotation
NCBI annotation	WP_011011332.1	**6-carboxy-5,6,7,8-tetrahydropterin synthase (NCBI Reference Sequence)**	WP_011012422.1	6-pyruvoyltetrahydropterin synthase (NCBI Reference Sequence)
Ensemble bacteria	AAL80343	putative 6-pyruvoyl tetrahydrobiopterin synthase (PF0219)	AAL81402	hypothetical protein (PF1278)
PATRIC (uses the RAST annotation pipeline)	fig186497.12.peg.227	**Queuosine biosynthesis QueD, PTPS-I**	fig186497.12.peg.1340	**Folate biosynthesis protein PTPS-III, catalyzes a reaction that bypasses dihydroneopterin aldolase (FolB)**
MaGe	PF0219	putative 6-pyruvoyl tetrahydrobiopterin synthase putative 6-pyruvoyl tetrahydrobiopterin synthase automatic/finished	PF1278	hypothetical protein hypothetical protein automatic/finished
IMG	638172701 [2]	NCBI RefSeq Annotation: putative 6-pyruvoyl tetrahydrobiopterin synthase TrEMBL annotation: Putative 6-pyruvoyl tetrahydrobiopterin synthase preQ(0) biosynthesis protein QueD	638173858 [2]	NCBI RefSeq Annotation: hypothetical protein TrEMBL annotation: Dihydroneopterin monophosphate aldolase hypothetical protein
	2625830234 [3]	6-pyruvoyltetrahydropterin/ **6-carboxytetrahydropterin synthase**	2625831353 [3]	6-pyruvoyltetrahydropterin/ 6-carboxytetrahydropterin synthase

[1] Annotations considered correct are highlighted in bold; [2] Initial annotation; [3] Re-annotated genome.

7. Integration of Tools for Paralog Separation in a Workflow

As discussed above, paralogs within large multigene families remain a potentially major source of errors in high-throughput gene annotation pipelines. A "stop-gap" first measure for mitigating the effects of paralogous genes on annotation quality could be to simply flag potential paralogs and integrate this flagging into annotation confidence scores [18]. If a gene has a homolog in the same genome, or even if it is part of a family that has several members in another genome, the confidence in its functional annotation should be reduced. This would introduce a level of "healthy skepticism" into predicted functional annotations that is lacking in most cases. This first step could be implemented efficiently and easily incorporated into existing annotation pipelines, as data on the distribution of gene families across genomes is already available in most integrated databases such as PATRIC, IMG or MaGE. While flagging potential paralogs is an appropriate first measure, the ultimate goal here is to examine the extent to which tools for rigorously separating multigene families into isofunctional subgroups could be integrated into high-throughput annotation pipelines.

Information from the highly curated SFLD [147] or CATH-Gene3D platforms [148,149] are considered the current "state of the art" in subfamily separation and classification. The COG0720 family is not currently covered by SFLD, but it is in CATH (CATH ID: 3.30.479.10). However, the CATH pipeline was not able to correctly separate PTPS subfamilies, even if it did capture the chemistry for two of them, PTPS-I and PTPS-II. That said, for the families covered, information from SFLD and CATH-Gene3D has the potential to improve gene functional annotations and should therefore be incorporated into annotation pipelines. Even if only some gene family annotations are expected to be improved by incorporation of manually curated database information, incorporating this information should help mitigate misannotation transfer errors and allow more generalizable—but computationally expensive—methods to be applied in the more difficult cases for which they are needed.

We also tested whether SSNs and Genome Neighborhood Networks (GNNs) developed by EFI [111] could separate the different COG0720 subgroups. We created an SSN from the COG0720 proteins extracted from the "PTPS paralogs" subsystem, a set of sequences for which the annotations have been verified. Increasing stringency allowed good separation of the different PTPS subfamilies into clusters, but the alignment score threshold required to fully separate the functional groups was so high that the same subfamily is also split into different clusters (Figure 4A). This problem persisted when we selected a more stringent set of starting sequences by examining only subsystem sequences from genomes that contained more than one COG0720 family member (Figure 4B). As this example demonstrates, in some "real-world" cases, sequence similarity may not be sufficient to identify all isofunctional groups within a multigene family without also breaking some isofunctional groups into multiple clusters. We then built the corresponding GNNs, as these might allow the regrouping of clusters based on common gene neighborhoods (Figure A1). The GNN generated did find neighbors that are relevant to the functions of the PTPS subfamilies. For example, we observed the expected physical clustering between PTPS-I and QueC, QueF and FolE. However, proteins of the same family were found in the neighborhood of proteins from different similarity clusters and more generally the noisiness of the GNN results make them difficult to interpret. If nothing were known about the family being analyzed, determining functional groups precisely using this data would be a challenge.

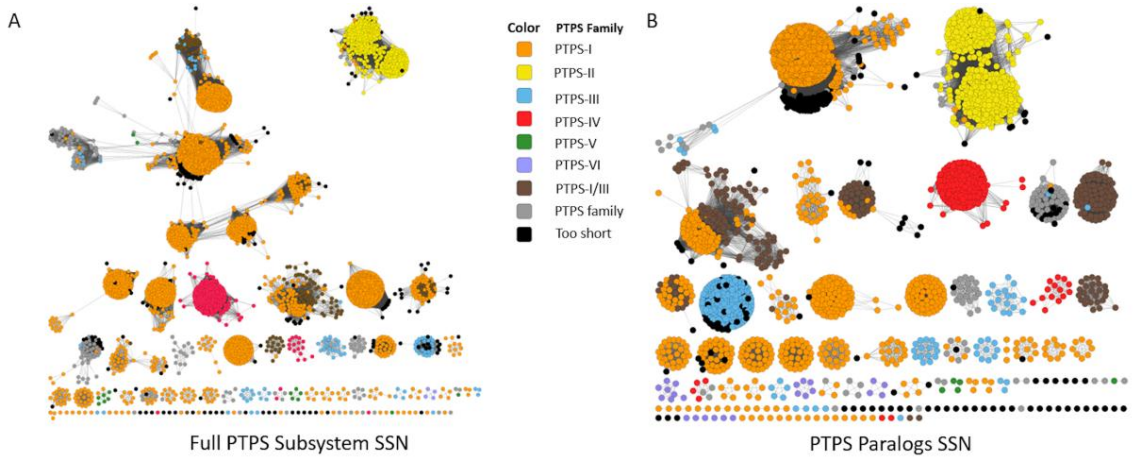

Figure 4. Sequence Similarity Networks Generated from the "PTPS paralog" Subsystem. An SSN was generated from the sequences extracted from the "PTPS paralog" subsystem, at alignment score threshold of 30 (**A**); A specific subset of sequences, limited to only to PTPS homologs in the same genomes (paralogs), were used to generate an SSN with identical parameters (**B**).

In summary, the EFI network tools were able to create networks that separate the COG0720 subfamilies. However, their analysis requires prior knowledge of the families as well as trial and error in order to provide accurate and reliable information about protein function. Noise-reduction techniques and other similar statistical tools could be developed to make these methods more useful for the average user or easier to integrate into automated annotation pipelines, but these additional analyses would need to be validated prior to deployment on a large scale.

None of the automated functional annotation pipelines were able to correctly separate COG0720 paralogs, whereas these paralogs were easily identified manually using comparative genomic tools. This suggests that an automated workflow inspired by our manual annotation strategy might be appropriate for separating challenging paralog groups (Figure 5). The key idea in this workflow is to focus initial analyses only on those genomes in which paralogs can be easily identified. Simply, any single genome containing at least two members of a protein family must encode paralogs, although the paralogs could perform the same or different functions. Once paralogs have been identified in a single genome, physical clustering by gene neighborhood can be used to group paralogs likely to have similar functions—because they physically group with the same genes across different genomes—and separate paralogs likely to have different functions—because they cluster with different groups of genes across genomes. Patterns of sequence conservation within and among paralog groups can then be used to identify "signature sequence motifs" that reliably identify the group of paralogs performing each functional role (Figure 1).

Tools to identify genomes encoding several genes of the same given family and extract the physical gene-neighborhood clusters associated with these genes are already available in the same integrative databases that already serve genome annotation pipelines such as IMG, MaGE or SEED, and we were able to use them to characterize the COG0720 family. Similarly, automated tools to identify signature motifs are now available. As shown in Figure 6, signature motifs identifying the three COG0720 families (PTPS-I, PTPS-II and PTPS-III) that we had created by manual analysis [55] could be recreated automatically using the Weblogo [172] and Web2logo platforms [173]. Once these signature motifs have been created, they can be used by tools such as HAMAP [168] to annotate all members of a multigene family, regardless of gene neighborhood (Figure 5). The value of using well-defined signature motifs as effective annotation tools has already been documented [174].

Based on our comparative genomics experience and approach, we propose the creation of a workflow, assembled from tools available, that we use and that are individually working well, to automate the identification of paralogs and their annotation. We believe that focusing on the difficult

task of paralog annotation is needed, and will lead to valuable discoveries. In addition, improving the annotation process will be beneficial for the general research community.

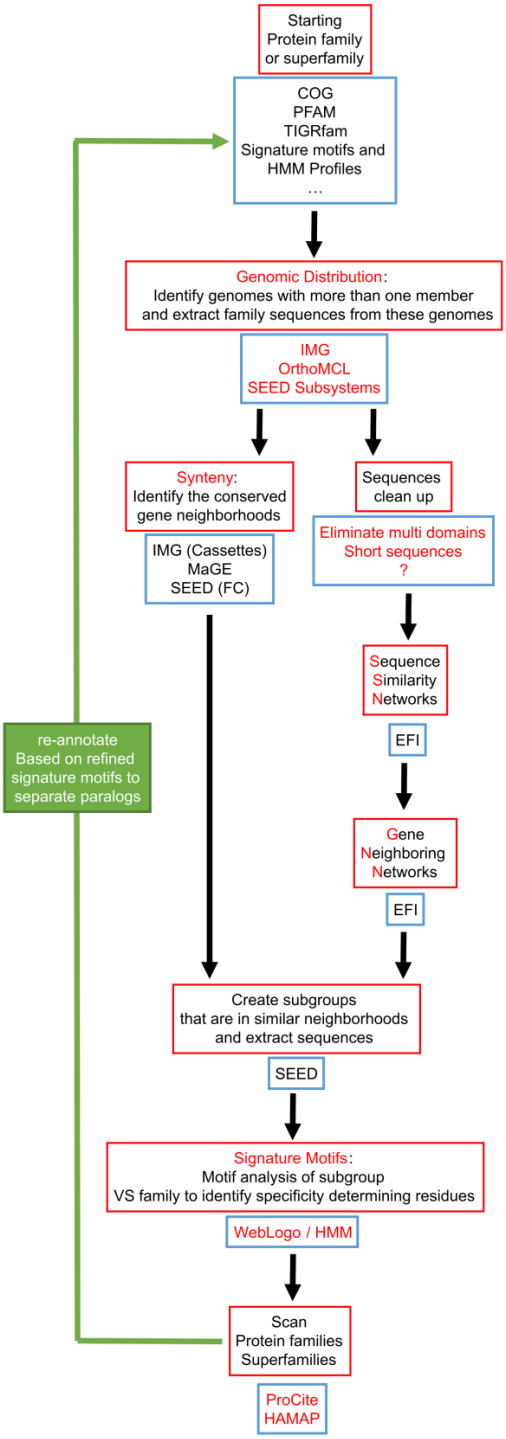

Figure 5. Proposed workflow for improving paralog annotation that integrates existing tools. The key features of this workflow are: focusing on genomes that already have paralogs, and building signature motifs only on groups of sequences that share similar physical neighborhoods. Red boxes describe the different required operations, and blue boxes list existing databases or software that provide the tools to perform these operations. The use of SSN and GNN is proposed as a strategy to identify the groups of sequences to build signatures as we feel their pipeline could become more automatable once it has matured. FC: Functionally coupled.

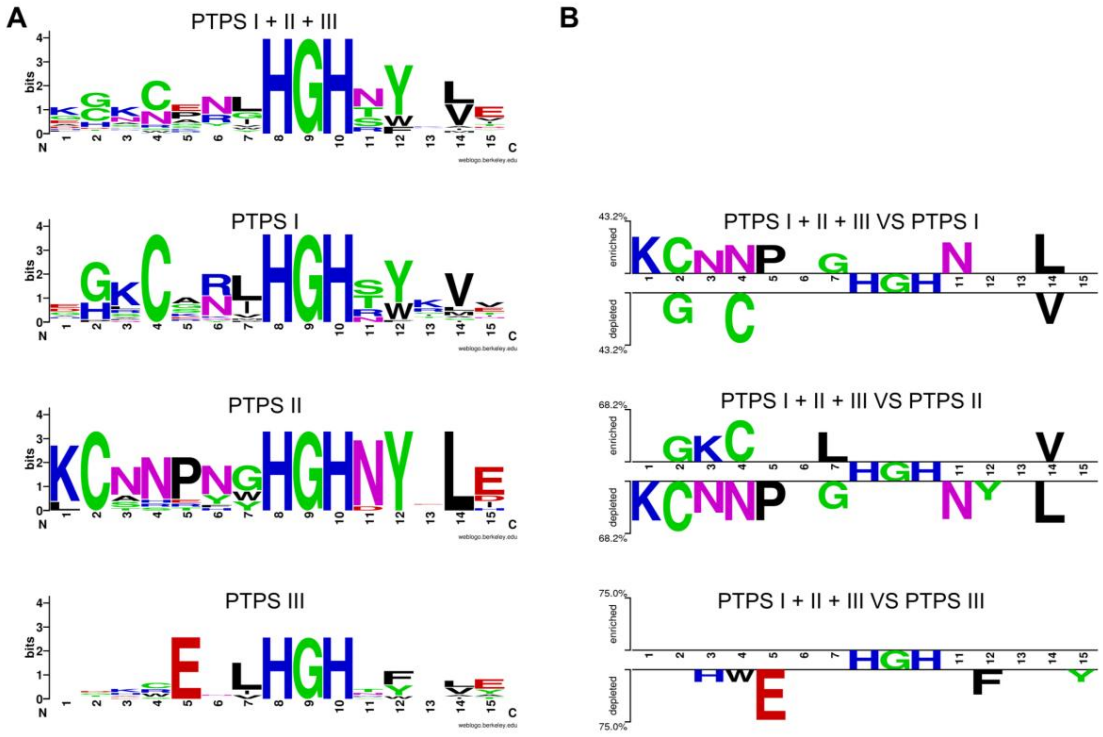

Figure 6. Motifs identified in paralogs from the same genome based on comparing with homologs with conserved physical clustering. Logos of the region containing the signature sequence (amino acids 23 to 38 in *E. coli* Uniprot P65870) from different COG0720 families (**A**); The choice of the sequences for the logos was as followed: starting from genomes that had two or more members of the family, only sequences that were in similar gene neighborhoods were chosen to generate the logos. The set of sequences used is given in Supplementary Material Document S1. The logos were generated using Weblogo [172]. Logos from the different subfamilies were compared using the Two Sample Logo platform [173] (**B**). This automated comparison tool objectively identifies key residues for each subfamily.

8. Materials and Methods

8.1. Bioinformatic Analyses

The BLAST tools and resources at NCBI were routinely used [175]. Analysis of the distribution of genes of the COG0720 on the different kingdoms was performed using the tools of the IMG database (Phylogenetic Profiler for Single Genes) [176] and by a parsing of data obtained from the COG database [58].

Analysis of the phylogenetic distribution and physical clustering was performed in the SEED database [119] on the 11,411 genomes available at the time of the analysis (November 2015). Results are available in the "PTPS paralogs" subsystem on the public SEED server (http://pubseed.theseed.org//SubsysEditor.cgi?page=ShowSubsystem&subsystem=PTPS_paralogs) [119]. A subset of the analysis is summarized Figure 3. Clustal Ω [177] was used to verify the belonging of sequences to the PTPS-I, II, III or IV subfamilies, based on the strict presence of previously identified motifs [55] and re-annotated when needed. Sequences that were too short (typically with a motif partly missing or integrally missing) were re-annotated as "PTPS family" and flagged with the comment "#short, possible wrong start call." We estimate that around 15% to 20% of the genes seem to be wrongly called. Sequences that align but for which no classification was possible based on the variability among the recognizable motif were annotated as "PTPS family." The PATRIC annotations [178] were extracted and mapped to Uniprot genes when needed.

For sequence logo analysis and comparison, selected sequences corresponding to PTPS-I, II and III were extracted from the "PTPS paralogs subsystem" from genomes for which there is more than one PTPS (and are defined as paralogs) and for which there is evidence of functional coupling with genes previously identified involved in the specific PTPS pathway. Sequences were aligned with Clustal Ω [177]. Obtained alignments were trimmed with AliView-1.17.1 [179] to focus on the conserved motif area. The reference for the comparison contains all the extracted sequences and is used as a representative for the PTPS family. From the alignment, WebLogo [172] was used to create logos of the HGH motif region. Each logo was compared with the others in Two Sample Logo using the *t*-test option [173]. The sequences extracted are available in document S1.

8.2. Sequence Similarity Network (SSN) and Neighborhood

A script allowing the extraction of homologous sequences that are present in the same genome from the PTPS Exploration subsystem on PubSEED was used as input to generate a Sequence Similarity Network (SSN) and a Genome Neighborhood Network (GNN). Networks were also generated using PTPS sequences from the entire subsystem as input. The networks were generated using the EFI Enzyme Similarity Tool and Genome Neighborhood Tool [111,180] available at http://efi.igb.illinois.edu/efi-est/. The SSN was generated with an initial alignment score threshold of 15 using the custom Fasta option (option C). Cytoscape was used to visualize and edit the networks [181]. Uniprot accessions were imported into the SSNs and all nodes that did not contain a Uniprot accession were removed. The alignment score threshold of the networks was increased by deleting edges with increasing alignment score values in order to visualize separation of clusters. A table with manually determined annotations extracted from the "PTPS paralogs" subsystem was added to the network in order to test whether the clusters were iso-functional groups. The GNN was generated with default values using the SSN with only nodes containing a Uniprot accession at an alignment score threshold of 25 as input. This score was chosen based on the complete separation of PTPS families.

9. Conclusions

The issue of gene functional annotation quality will need to be solved for systems biology approaches to reach their full potential. As discussed here, paralogs are a major source of errors in gene functional annotation, and annotation quality cannot generally improve until paralog families are better annotated. Given recent advances in computational infrastructure and recent development of robust small-scale annotation efforts, we feel the time is right for large-scale annotation pipelines to begin integrating the methods used in small-scale expert-based annotations. With genome sequencing poised to generate a super-exponential increase in total sequence data over the coming years, failure to incorporate reliable gene-annotation methods now could leave us with a challenging "re-annotation" problem in the future. Although the methods described here are mainly applicable to prokaryotic genomes, as physical clustering is rare in eukaryotes, they may provide information informing annotation of the portion of eukaryote genomes that have bacterial homologs [48]. In addition to improving annotation quality, reliable identification of paralogs within multigene families is expected to provide identification of subfamilies with specific—but unknown—chemistry, potentially driving the discovery of previously unanticipated novel gene functions as shown with the examples in Table 1.

Supplementary Materials: The following are available online at http://www.mdpi.com/2075-1729/6/3/39/s1. Table S1: Genome list is provided as supplementary information. Document S1: Sequences. The sequences from which the logos and compared logos figures were created are available in this file.

Acknowledgments: We would like to thank Arcady R. Mushegian for his book, *Foundations of Comparative Genomics*, that was a constant inspiration for this work and Svetlana Gerdes and Jennifer J. Thiaville for their critical reading of the manuscript. This work was supported by the National Institutes of Health (grant number R01 GM70641 to V.d.C.-L.) and the National Science Foundation (grant number MCB 00116984 to B.K.).

Author Contributions: V.d.C.-L. and R.Z. conceived and designed the experiments; R.Z. and K.J.H. performed the experiments; and R.Z., K.J.H. and V.d.C.-L. analyzed the data; R.Z., B.K. and V.d.C.-L wrote the paper.

Conflicts of Interest: The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

Appendix

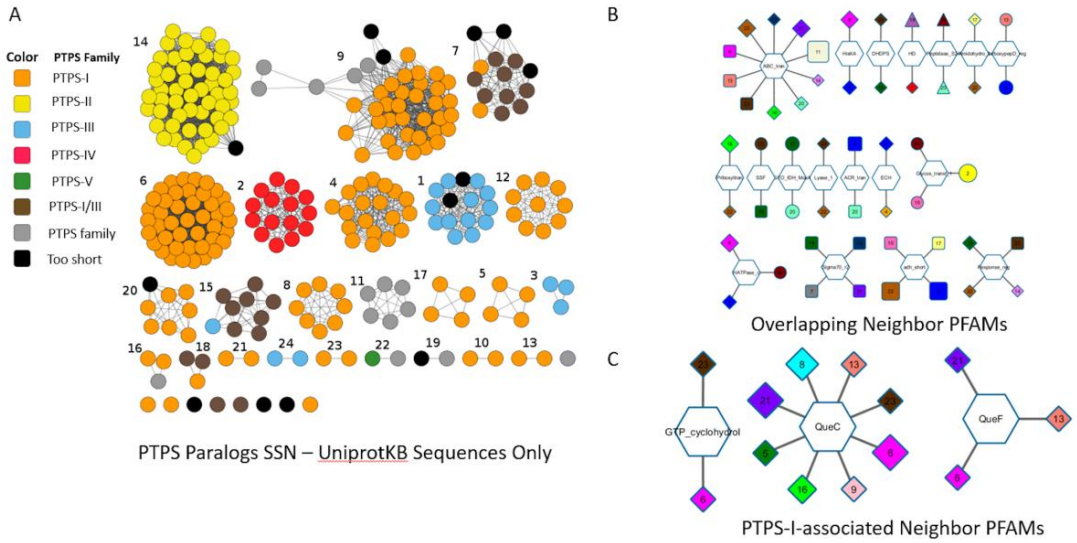

Figure A1. Genome Neighborhood Network (GNN) generated from the "PTPS paralog" subsystem. A Sequence Similarity Network was generated with the COG0720 proteins extracted from the "PTPS paralog" subsystem, but only from genomes that harbored several proteins of the family and that could be mapped to an Uniprot ID (**A**); The alignment score threshold was 25. Each SSN cluster is identified with a unique number. To create the GNN (**B**), the neighboring (10) genes of each gene in a given cluster are captured as a PFAM ID. The hexagonal inside nodes are these PFAMs and the external nodes represent the clusters that have this PFAM in their neighborhoods. Broad PFAM families such as ABC transporters (ABC-trans) or Histidine kinase (HisKA) are linked to clusters that belong to different subfamilies introducing background noise that make a blind analysis difficult. When looking at PFAMs associated with PTPS-I clusters (**C**), relevant neighbors such as QueC, QueF and FolE (PFAM name: GTP_CyclohydroI) are identified. However, not all PTPS-I clusters are present: the clusters 4, 12, 20, 17, 16, and 10 are missing.

References

1. Reddy, T.B.K.; Thomas, A.D.; Stamatis, D.; Bertsch, J.; Isbandi, M.; Jansson, J.; Mallajosyula, J.; Pagani, I.; Lobos, E.A.; Kyrpides, N.C. The Genomes OnLine Database (GOLD) v.5: A metadata management system based on a four level (meta)genome project classification. *Nucleic Acids Res.* **2015**, *43*, D1099–D1106. [CrossRef] [PubMed]
2. Lasken, R.S.; McLean, J.S. Recent advances in genomic DNA sequencing of microbial species from single cells. *Nat. Rev. Genet.* **2014**, *15*, 577–584. [CrossRef] [PubMed]
3. Shendure, J.; Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **2008**, *26*, 1135–1145. [CrossRef] [PubMed]
4. Mellis, I.A.; Raj, A. Half dozen of one, six billion of the other: What can small- and large-scale molecular systems biology learn from one another? *Genome Res.* **2015**, *25*, 1466–1472. [CrossRef] [PubMed]
5. Fisher, R.A. On the interpretation of χ^2 from contingency tables, and the calculation of P. *J. R. Stat. Soc.* **1922**, *85*, 87. [CrossRef]
6. Larntz, K. Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics. *J. Am. Stat. Assoc.* **1978**, *73*, 253–263. [CrossRef]
7. Klimke, W.; O'Donovan, C.; White, O.; Brister, J.R.; Clark, K.; Fedorov, B.; Mizrachi, I.; Pruitt, K.D.; Tatusova, T. Solving the Problem: Genome Annotation Standards before the Data Deluge. *Stand. Genom. Sci.* **2011**, *5*, 168–193. [CrossRef] [PubMed]

8. Brent, R. Genomic biology. *Cell* **2000**, *100*, 169–183. [CrossRef]
9. Davidson, D.; Baldock, R. Bioinformatics beyond sequence: Mapping gene function in the embryo. *Nat. Rev. Genet.* **2001**, *2*, 409–417. [CrossRef] [PubMed]
10. Murali, T.M. Computationally Driven Experimental Biology. *Computer* **2012**, *45*, 22–23.
11. Tritt, A.; Eisen, J.A.; Facciotti, M.T.; Darling, A.E. An integrated pipeline for de novo assembly of microbial genomes. *PLoS ONE* **2012**, *7*, e42304. [CrossRef] [PubMed]
12. Dunitz, M.I.; Lang, J.M.; Jospin, G.; Darling, A.E.; Eisen, J.A.; Coil, D.A. Swabs to genomes: A comprehensive workflow. *PeerJ* **2015**, *3*, e960. [CrossRef] [PubMed]
13. Overbeek, R.; Olson, R.; Pusch, G.D.; Olsen, G.J.; Davis, J.J.; Disz, T.; Edwards, R.A.; Gerdes, S.; Parrello, B.; Shukla, M.; et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.* **2014**, *42*, D206–D214. [CrossRef] [PubMed]
14. Crappé, J.; Ndah, E.; Koch, A.; Steyaert, S.; Gawron, D.; de Keulenaer, S.; de Meester, E.; de Meyer, T.; van Criekinge, W.; van Damme, P.; Menschaert, G. Proteoformer: Deep proteome coverage through ribosome profiling and MS integration. *Nucleic Acids Res.* **2015**, *43*, e29. [CrossRef] [PubMed]
15. Siezen, R.J.; van Hijum, S.A.F.T. Genome (re-)annotation and open-source annotation pipelines. *Microb. Biotechnol.* **2010**, *3*, 362–369. [CrossRef] [PubMed]
16. Overmars, L.; Siezen, R.J.; Francke, C. A Novel Quality Measure and Correction Procedure for the Annotation of Microbial Translation Initiation Sites. *PLoS ONE* **2015**, *10*, e0133691. [CrossRef] [PubMed]
17. Chen, I.-M.A.; Markowitz, V.M.; Chu, K.; Anderson, I.; Mavromatis, K.; Kyrpides, N.C.; Ivanova, N.N. Improving microbial genome annotations in an integrated database context. *PLoS ONE* **2013**, *8*, e54859. [CrossRef] [PubMed]
18. Bastian, F.B.; Chibucos, M.C.; Gaudet, P.; Giglio, M.; Holliday, G.L.; Huang, H.; Lewis, S.E.; Niknejad, A.; Orchard, S.; Poux, S.; et al. The Confidence Information Ontology: A step towards a standard for asserting confidence in annotations. *Database (Oxford)* **2015**, *2015*, bav043. [CrossRef] [PubMed]
19. Óhéigeartaigh, S.S.; Armisén, D.; Byrne, K.P.; Wolfe, K.H. SearchDOGS bacteria, software that provides automated identification of potentially missed genes in annotated bacterial genomes. *J. Bacteriol.* **2014**, *196*, 2030–2042. [CrossRef] [PubMed]
20. Bork, P.; Bairoch, A. Go hunting in sequence databases but watch out for the traps. *Trends Genet.* **1996**, *12*, 425–427. [CrossRef]
21. Schnoes, A.M.; Brown, S.D.; Dodevski, I.; Babbitt, P.C. Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.* **2009**, *5*, e1000605. [CrossRef] [PubMed]
22. Anton, B.P.; Kasif, S.; Roberts, R.J.; Steffen, M. Objective: Biochemical function. *Front. Genet.* **2014**, *5*, 210. [CrossRef] [PubMed]
23. Wu, Q.; Ye, Y.; Ng, M.K.; Ho, S.-S.; Shi, R. Collective prediction of protein functions from protein-protein interaction networks. *BMC Bioinform.* **2014**, *15* (Suppl. 2), S9. [CrossRef] [PubMed]
24. Pfeiffer, F.; Oesterhelt, D. A manual curation strategy to improve genome annotation: Application to a set of haloarchael genomes. *Life (Basel, Switzerland)* **2015**, *5*, 1427–1444. [CrossRef] [PubMed]
25. Poux, S.; Magrane, M.; Arighi, C.N.; Bridge, A.; O'Donovan, C.; Laiho, K. UniProt Consortium Expert curation in UniProtKB: A case study on dealing with conflicting and erroneous data. *Database (Oxford)* **2014**, *2014*, bau016. [CrossRef] [PubMed]
26. Brenner, S.E. Errors in genome annotation. *Trends Genet.* **1999**, *15*, 132–133. [CrossRef]
27. Bell, M.J.; Collison, M.; Lord, P. Can inferred provenance and its visualisation be used to detect erroneous annotation? A case study using UniProtKB. *PLoS ONE* **2013**, *8*, e75541. [CrossRef] [PubMed]
28. Poptsova, M.S.; Gogarten, J.P. Using comparative genome analysis to identify problems in annotated microbial genomes. *Microbiology* **2010**, *156*, 1909–1917. [CrossRef] [PubMed]
29. Radivojac, P.; Clark, W.T.; Oron, T.R.; Schnoes, A.M.; Wittkop, T.; Sokolov, A.; Graim, K.; Funk, C.; Verspoor, K.; Ben-Hur, A.; et al. A large-scale evaluation of computational protein function prediction. *Nat. Methods* **2013**, *10*, 221–227. [CrossRef] [PubMed]
30. Gillis, J.; Pavlidis, P. Characterizing the state of the art in the computational assignment of gene function: Lessons from the first critical assessment of functional annotation (CAFA). *BMC Bioinform.* **2013**, *14* (Suppl. 3), S15. [CrossRef]
31. Kahanda, I.; Funk, C.S.; Ullah, F.; Verspoor, K.M.; Ben-Hur, A. A close look at protein function prediction evaluation protocols. *Gigascience* **2015**, *4*, 41. [CrossRef] [PubMed]

32. Koonin, E.V.; Galperin, M.Y. *Sequence—Evolution—Function*; Springer US: Boston, MA, USA, 2003.
33. Lee, D.; Redfern, O.; Orengo, C. Predicting protein function from sequence and structure. *Nat. Rev. Mol. Cell Biol.* **2007**, *8*, 995–1005. [CrossRef] [PubMed]
34. Percudani, R.; Carnevali, D.; Puggioni, V. Ureidoglycolate hydrolase, amidohydrolase, lyase: How errors in biological databases are incorporated in scientific papers and vice versa. *Database (Oxford)* **2013**, *2013*, bat071. [CrossRef] [PubMed]
35. Mao, F.; Su, Z.; Olman, V.; Dam, P.; Liu, Z.; Xu, Y. Mapping of orthologous genes in the context of biological pathways: An application of integer programming. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 129–134. [CrossRef] [PubMed]
36. Bork, P.; Koonin, E.V. Predicting functions from protein sequences—Where are the bottlenecks? *Nat. Genet.* **1998**, *18*, 313–318. [CrossRef] [PubMed]
37. Green, M.L.; Karp, P.D. Genome annotation errors in pathway databases due to semantic ambiguity in partial EC numbers. *Nucleic Acids Res.* **2005**, *33*, 4035–4039. [CrossRef] [PubMed]
38. Devos, D.; Valencia, A. Intrinsic errors in genome annotation. *Trends Genet.* **2001**, *17*, 429–431. [CrossRef]
39. Promponas, V.J.; Iliopoulos, I.; Ouzounis, C.A. Annotation inconsistencies beyond sequence similarity-based function prediction—Phylogeny and genome structure. *Stand. Genom. Sci.* **2015**, *10*, 108. [CrossRef] [PubMed]
40. Dornfeld, C.; Weisberg, A.J.; K C, R.; Dudareva, N.; Jelesko, J.G.; Maeda, H.A. Phylobiochemical characterization of class-Ib aspartate/prephenate aminotransferases reveals evolution of the plant arogenate phenylalanine pathway. *Plant Cell* **2014**, *26*, 3101–3114. [CrossRef] [PubMed]
41. Verdel-Aranda, K.; López-Cortina, S.T.; Hodgson, D.A.; Barona-Gómez, F. Molecular annotation of ketol-acid reductoisomerases from Streptomyces reveals a novel amino acid biosynthesis interlock mediated by enzyme promiscuity. *Microb. Biotechnol.* **2015**, *8*, 239–252. [CrossRef] [PubMed]
42. Brown, S.D.; Babbitt, P.C. New insights about enzyme evolution from large scale studies of sequence and structure relationships. *J. Biol. Chem.* **2014**, *289*, 30221–30228. [CrossRef] [PubMed]
43. Das, S.; Orengo, C.A. Protein function annotation using protein domain family resources. *Methods* **2016**, *93*, 24–34. [CrossRef] [PubMed]
44. Barona-Gómez, F. Re-annotation of the sequence > annotation: Opportunities for the functional microbiologist. *Microb. Biotechnol.* **2015**, *8*, 2–4. [CrossRef] [PubMed]
45. Van Lanen, S.G.; Reader, J.S.; Swairjo, M.A.; de Crécy-Lagard, V.; Lee, B.; Iwata-Reuyl, D. From cyclohydrolase to oxidoreductase: Discovery of nitrile reductase activity in a common fold. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 4264–4269. [CrossRef] [PubMed]
46. Phillips, G.; Swairjo, M.A.; Gaston, K.W.; Bailly, M.; Limbach, P.A.; Iwata-Reuyl, D.; de Crécy-Lagard, V. Diversity of archaeosine synthesis in crenarchaeota. *ACS Chem. Biol.* **2012**, *7*, 300–305. [CrossRef] [PubMed]
47. Pribat, A.; Blaby, I.K.; Lara-Núñez, A.; Gregory, J.F.; de Crécy-Lagard, V.; Hanson, A.D. FolX and FolM are essential for tetrahydromonapterin synthesis in *Escherichia coli* and *Pseudomonas aeruginosa*. *J. Bacteriol.* **2010**, *192*, 475–482. [CrossRef] [PubMed]
48. Gerdes, S.; El Yacoubi, B.; Bailly, M.; Blaby, I.K.; Blaby-Haas, C.E.; Jeanguenin, L.; Lara-Núñez, A.; Pribat, A.; Waller, J.C.; Wilke, A.; et al. Synergistic use of plant-prokaryote comparative genomics for functional annotations. *BMC Genom.* **2011**, *12* (Suppl. 1), S2. [CrossRef] [PubMed]
49. Bailly, M.; de Crécy-Lagard, V. Predicting the pathway involved in post-translational modification of elongation factor P in a subset of bacterial species. *Biol. Direct* **2010**, *5*, 3. [CrossRef] [PubMed]
50. Waller, J.C.; Alvarez, S.; Naponelli, V.; Lara-Nuñez, A.; Blaby, I.K.; da Silva, V.; Ziemak, M.J.; Vickers, T.J.; Beverley, S.M.; Edison, A.S.; et al. A role for tetrahydrofolates in the metabolism of iron-sulfur clusters in all domains of life. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 10412–10417. [CrossRef] [PubMed]
51. De Crécy-Lagard, V.; Forouhar, F.; Brochier-Armanet, C.; Tong, L.; Hunt, J.F. Comparative genomic analysis of the DUF71/COG2102 family predicts roles in diphthamide biosynthesis and B12 salvage. *Biol. Direct* **2012**, *7*, 32. [CrossRef] [PubMed]
52. Adams, N.E.; Thiaville, J.J.; Proestos, J.; Juárez-Vázquez, A.L.; McCoy, A.J.; Barona-Gómez, F.; Iwata-Reuyl, D.; de Crécy-Lagard, V.; Maurelli, A.T. Promiscuous and adaptable enzymes fill "holes" in the tetrahydrofolate pathway in Chlamydia species. *mBio* **2014**, *5*, e01378–e013714. [CrossRef] [PubMed]

53. De Crécy-Lagard, V.; El Yacoubi, B.; de la Garza, R.D.; Noiriel, A.; Hanson, A.D. Comparative genomics of bacterial and plant folate synthesis and salvage: Predictions and validations. *BMC Genom.* **2007**, *8*, 245. [CrossRef] [PubMed]
54. Chatterjee, K.; Blaby, I.K.; Thiaville, P.C.; Majumder, M.; Grosjean, H.; Yuan, Y.A.; Gupta, R.; de Crécy-Lagard, V. The archaeal COG1901/DUF358 SPOUT-methyltransferase members, together with pseudouridine synthase Pus10, catalyze the formation of 1-methylpseudouridine at position 54 of tRNA. *RNA* **2012**, *18*, 421–433. [CrossRef] [PubMed]
55. Phillips, G.; Grochowski, L.L.; Bonnett, S.; Xu, H.; Bailly, M.; Blaby-Haas, C.; El Yacoubi, B.; Iwata-Reuyl, D.; White, R.H.; de Crécy-Lagard, V. Functional promiscuity of the COG0720 family. *ACS Chem. Biol.* **2012**, *7*, 197–209. [CrossRef] [PubMed]
56. Haas, C.E.; Rodionov, D.A.; Kropat, J.; Malasarn, D.; Merchant, S.S.; de Crécy-Lagard, V. A subset of the diverse COG0523 family of putative metal chaperones is linked to zinc homeostasis in all kingdoms of life. *BMC Genom.* **2009**, *10*, 470. [CrossRef] [PubMed]
57. Pribat, A.; Blaby, I.K.; Lara-Núñez, A.; Jeanguenin, L.; Fouquet, R.; Frelin, O.; Gregory, J.F.; Philmus, B.; Begley, T.P.; de Crécy-Lagard, V.; et al. A 5-formyltetrahydrofolate cycloligase paralog from all domains of life: Comparative genomic and experimental evidence for a cryptic role in thiamin metabolism. *Funct. Integr. Genom.* **2011**, *11*, 467–478. [CrossRef] [PubMed]
58. Galperin, M.Y.; Makarova, K.S.; Wolf, Y.I.; Koonin, E.V. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.* **2015**, *43*, D261–D269. [CrossRef] [PubMed]
59. Lan, N.; Montelione, G.T.; Gerstein, M. Ontologies for proteomics: Towards a systematic definition of structure and function that scales to the genome level. *Curr. Opin. Chem. Biol.* **2003**, *7*, 44–54. [CrossRef]
60. Lan, N.; Jansen, R.; Gerstein, M. Toward a systematic definition of protein function that scales to the genome level: Defining function in terms of interactions. *IEEE Proc.* **2002**, *90*, 1848–1858.
61. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef] [PubMed]
62. Mao, X.; Cai, T.; Olyarchuk, J.G.; Wei, L. Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics* **2005**, *21*, 3787–3793. [CrossRef] [PubMed]
63. Reference Genome Group of the Gene Ontology Consortium. The Gene Ontology's Reference Genome Project: A unified framework for functional annotation across species. *PLoS Comput. Biol.* **2009**, *5*, e1000431.
64. Conesa, A.; Götz, S.; García-Gómez, J.M.; Terol, J.; Talón, M.; Robles, M. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **2005**, *21*, 3674–3676. [CrossRef] [PubMed]
65. Reed, J.L.; Famili, I.; Thiele, I.; Palsson, B.O. Towards multidimensional genome annotation. *Nat. Rev. Genet.* **2006**, *7*, 130–141. [CrossRef] [PubMed]
66. Yandell, M.; Ence, D. A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* **2012**, *13*, 329–342. [CrossRef] [PubMed]
67. Richardson, E.J.; Watson, M. The automatic annotation of bacterial genomes. *Brief. Bioinform.* **2013**, *14*, 1–12. [CrossRef] [PubMed]
68. Jensen, L.J.; Ussery, D.W.; Brunak, S. Functionality of system components: Conservation of protein function in protein feature space. *Genome Res.* **2003**, *13*, 2444–2449. [CrossRef] [PubMed]
69. Pereira, C.; Denise, A.; Lespinet, O. A meta-approach for improving the prediction and the functional annotation of ortholog groups. *BMC Genom.* **2014**, *15* (Suppl. 6), S16. [CrossRef] [PubMed]
70. Brown, D.P.; Krishnamurthy, N.; Sjölander, K. Automated protein subfamily identification and classification. *PLoS Comput. Biol.* **2007**, *3*, e160. [CrossRef] [PubMed]
71. Engelhardt, B.E.; Jordan, M.I.; Srouji, J.R.; Brenner, S.E. Genome-scale phylogenetic function annotation of large and diverse protein families. *Genome Res.* **2011**, *21*, 1969–1980. [CrossRef] [PubMed]
72. Fitch, W.M. Distinguishing homologous from analogous proteins. *Syst. Biol.* **1970**, *19*, 99–113. [CrossRef]
73. Altenhoff, A.M.; Studer, R.A.; Robinson-Rechavi, M.; Dessimoz, C. Resolving the ortholog conjecture: Orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Comput. Biol.* **2012**, *8*, e1002514. [CrossRef] [PubMed]
74. Chen, F.; Mackey, A.J.; Stoeckert, C.J.; Roos, D.S. OrthoMCL-DB: Querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* **2006**, *34*, D363–D368. [CrossRef] [PubMed]

75. Altenhoff, A.M.; Škunca, N.; Glover, N.; Train, C.-M.; Sueki, A.; Piližota, I.; Gori, K.; Tomiczek, B.; Müller, S.; Redestig, H.; et al. The OMA orthology database in 2015: Function predictions, better plant support, synteny view and other improvements. *Nucleic Acids Res.* **2015**, *43*, D240–D249. [CrossRef] [PubMed]
76. Huerta-Cepas, J.; Szklarczyk, D.; Forslund, K.; Cook, H.; Heller, D.; Walter, M.C.; Rattei, T.; Mende, D.R.; Sunagawa, S.; Kuhn, M.; et al. eggNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **2016**, *44*, D286–D293. [CrossRef] [PubMed]
77. Gerlt, J.A.; Babbitt, P.C. Can sequence determine function? *Genome Biol.* **2000**, *1*, S0005. [CrossRef] [PubMed]
78. Jensen, R.A. Orthologs and paralogs—We need to get it right. *Genome Biol.* **2001**, *2*, S1002. [CrossRef]
79. Studer, R.A.; Robinson-Rechavi, M. How confident can we be that orthologs are similar, but paralogs differ? *Trends Genet.* **2009**, *25*, 210–216. [CrossRef] [PubMed]
80. Nehrt, N.L.; Clark, W.T.; Radivojac, P.; Hahn, M.W. Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Comput. Biol.* **2011**, *7*, e1002073. [CrossRef] [PubMed]
81. Gharib, W.H.; Robinson-Rechavi, M. When orthologs diverge between human and mouse. *Brief. Bioinform.* **2011**, *12*, 436–441. [CrossRef] [PubMed]
82. Gabaldón, T.; Koonin, E.V. Functional and evolutionary implications of gene orthology. *Nat. Rev. Genet.* **2013**, *14*, 360–366. [CrossRef] [PubMed]
83. Gout, J.-F.; Lynch, M. Maintenance and loss of duplicated genes by dosage subfunctionalization. *Mol. Biol. Evol.* **2015**, *32*, 2141–2148. [CrossRef] [PubMed]
84. Papp, B.; Pál, C.; Hurst, L.D. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **2003**, *424*, 194–197. [CrossRef] [PubMed]
85. Gout, J.-F.; Kahn, D.; Duret, L.; Paramecium Post-Genomics Consortium. The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. *PLoS Genet.* **2010**, *6*, e1000944. [CrossRef]
86. Qian, W.; Liao, B.-Y.; Chang, A.Y.-F.; Zhang, J. Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends Genet.* **2010**, *26*, 425–430. [CrossRef] [PubMed]
87. Chan, C.T.Y.; Pang, Y.L.J.; Deng, W.; Babu, I.R.; Dyavaiah, M.; Begley, T.J.; Dedon, P.C. Reprogramming of tRNA modifications controls the oxidative stress response by codon-biased translation of proteins. *Nat. Commun.* **2012**, *3*, 937. [CrossRef] [PubMed]
88. Fillinger, S.; Boschi-Muller, S.; Azza, S.; Dervyn, E.; Branlant, G.; Aymerich, S. Two glyceraldehyde-3-phosphate dehydrogenases with opposite physiological roles in a nonphotosynthetic bacterium. *J. Biol. Chem.* **2000**, *275*, 14031–14037. [CrossRef] [PubMed]
89. Rusin, L.Y.; Lyubetskaya, E.V.; Gorbunov, K.Y.; Lyubetsky, V.A. Reconciliation of gene and species trees. *BioMed Res. Int.* **2014**, *2014*, 642089. [CrossRef] [PubMed]
90. Szöllősi, G.J.; Tannier, E.; Daubin, V.; Boussau, B. The inference of gene trees with species trees. *Syst. Biol.* **2015**, *64*, e42–e62. [CrossRef] [PubMed]
91. Wu, Y.-C.; Rasmussen, M.D.; Bansal, M.S.; Kellis, M. Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees. *Genome Res.* **2014**, *24*, 475–486. [CrossRef] [PubMed]
92. Doyon, J.-P.; Ranwez, V.; Daubin, V.; Berry, V. Models, algorithms and programs for phylogeny reconciliation. *Brief. Bioinform.* **2011**, *12*, 392–400. [CrossRef] [PubMed]
93. Arvestad, L.; Berglund, A.-C.; Lagergren, J.; Sennblad, B. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics* **2003**, *19* (Suppl. 1), i7–i15. [CrossRef] [PubMed]
94. Vernot, B.; Stolzer, M.; Goldman, A.; Durand, D. Reconciliation with non-binary species trees. *J. Comput. Biol.* **2008**, *15*, 981–1006. [CrossRef] [PubMed]
95. Kolaczkowski, B.; Thornton, J.W. Long-branch attraction bias and inconsistency in Bayesian phylogenetics. *PLoS ONE* **2009**, *4*, e7891. [CrossRef] [PubMed]
96. Kolaczkowski, B.; Thornton, J.W. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* **2004**, *431*, 980–984. [CrossRef] [PubMed]
97. Hahn, M.W. Bias in phylogenetic tree reconciliation methods: Implications for vertebrate genome evolution. *Genome Biol.* **2007**, *8*, R141. [CrossRef] [PubMed]
98. Jeffroy, O.; Brinkmann, H.; Delsuc, F.; Philippe, H. Phylogenomics: The beginning of incongruence? *Trends Genet.* **2006**, *22*, 225–231. [CrossRef] [PubMed]

99. Engelhardt, B.E.; Jordan, M.I.; Muratore, K.E.; Brenner, S.E. Protein molecular function prediction by Bayesian phylogenomics. *PLoS Comput. Biol.* **2005**, *1*, e45. [CrossRef] [PubMed]
100. Sahraeian, S.M.; Luo, K.R.; Brenner, S.E. SIFTER search: A web server for accurate phylogeny-based protein function prediction. *Nucleic Acids Res.* **2015**, *43*, W141–W147. [CrossRef] [PubMed]
101. Giribet, G. Efficient tree searches with available algorithms. *Evol. Bioinform. Online* **2007**, *3*, 341–356. [PubMed]
102. Price, M.N.; Dehal, P.S.; Arkin, A.P. FastTree 2—Approximately maximum-likelihood trees for large alignments. *PLoS ONE* **2010**, *5*, e9490. [CrossRef] [PubMed]
103. Howe, K.; Bateman, A.; Durbin, R. QuickTree: Building huge Neighbour-Joining trees of protein sequences. *Bioinformatics* **2002**, *18*, 1546–1547. [CrossRef] [PubMed]
104. Hillis, D.M. Approaches for assessing phylogenetic accuracy. *Syst. Biol.* **1995**, *44*, 3–16. [CrossRef]
105. Cotton, J.A. Analytical methods for detecting paralogy in molecular datasets. *Methods Enzymol.* **2005**, *395*, 700–724. [PubMed]
106. Lechner, M.; Hernandez-Rosales, M.; Doerr, D.; Wieseke, N.; Thévenin, A.; Stoye, J.; Hartmann, R.K.; Prohaska, S.J.; Stadler, P.F. Orthology detection combining clustering and synteny for very large datasets. *PLoS ONE* **2014**, *9*, e105015. [CrossRef] [PubMed]
107. Kristensen, D.M.; Wolf, Y.I.; Mushegian, A.R.; Koonin, E.V. Computational methods for Gene Orthology inference. *Brief. Bioinform.* **2011**, *12*, 379–391. [CrossRef] [PubMed]
108. Tatusov, R.L.; Galperin, M.Y.; Natale, D.A.; Koonin, E.V. The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **2000**, *28*, 33–36. [CrossRef] [PubMed]
109. Kuzniar, A.; van Ham, R.C.H.J.; Pongor, S.; Leunissen, J.A.M. The quest for orthologs: Finding the corresponding gene across genomes. *Trends Genet.* **2008**, *24*, 539–551. [CrossRef] [PubMed]
110. Anderson, C.N.K.; Liu, L.; Pearl, D.; Edwards, S.V. Tangled trees: The challenge of inferring species trees from coalescent and noncoalescent genes. *Methods Mol. Biol.* **2012**, *856*, 3–28. [PubMed]
111. Gerlt, J.A.; Bouvier, J.T.; Davidson, D.B.; Imker, H.J.; Sadkhin, B.; Slater, D.R.; Whalen, K.L. Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks. *Biochim. Biophys. Acta* **2015**, *1854*, 1019–1037. [CrossRef] [PubMed]
112. Cantarel, B.L.; Morrison, H.G.; Pearson, W. Exploring the relationship between sequence similarity and accurate phylogenetic trees. *Mol. Biol. Evol.* **2006**, *23*, 2090–2100. [CrossRef] [PubMed]
113. Kelly, S.; Maini, P.K. DendroBLAST: Approximate phylogenetic trees in the absence of multiple sequence alignments. *PLoS ONE* **2013**, *8*, e58537. [CrossRef] [PubMed]
114. Trachana, K.; Forslund, K.; Larsson, T.; Powell, S.; Doerks, T.; von Mering, C.; Bork, P. A phylogeny-based benchmarking test for orthology inference reveals the limitations of function-based validation. *PLoS ONE* **2014**, *9*, e111122. [CrossRef] [PubMed]
115. Swofford, D.L.; Waddell, P.J.; Huelsenbeck, J.P.; Foster, P.G.; Lewis, P.O.; Rogers, J.S. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Syst. Biol.* **2001**, *50*, 525–539. [CrossRef] [PubMed]
116. Tatusov, R.L.; Koonin, E.V.; Lipman, D.J. A genomic perspective on protein families. *Science* **1997**, *278*, 631–637. [CrossRef] [PubMed]
117. Overbeek, R.; Larsen, N.; Pusch, G.D.; D'Souza, M.; Selkov, E.; Kyrpides, N.; Fonstein, M.; Maltsev, N.; Selkov, E. WIT: Integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.* **2000**, *28*, 123–125. [CrossRef] [PubMed]
118. Overbeek, R.; Fonstein, M.; D'Souza, M.; Pusch, G.D.; Maltsev, N. Use of contiguity on the chromosome to predict functional coupling. *In Silico Biol.* **1999**, *1*, 93–108. [PubMed]
119. Overbeek, R.; Begley, T.; Butler, R.M.; Choudhuri, J.V.; Chuang, H.-Y.; Cohoon, M.; de Crécy-Lagard, V.; Diaz, N.; Disz, T.; Edwards, R.; et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* **2005**, *33*, 5691–5702. [CrossRef] [PubMed]
120. Ye, Y.; Osterman, A.; Overbeek, R.; Godzik, A. Automatic detection of subsystem/pathway variants in genome analysis. *Bioinformatics* **2005**, *21* (Suppl. 1), i478–i486. [CrossRef] [PubMed]
121. Liberal, R.; Pinney, J.W. Simple topological properties predict functional misannotations in a metabolic network. *Bioinformatics* **2013**, *29*, i154–i161. [CrossRef] [PubMed]
122. Osterman, A.; Overbeek, R. Missing genes in metabolic pathways: A comparative genomics approach. *Curr. Opin. Chem. Biol.* **2003**, *7*, 238–251. [CrossRef]

123. Earnshaw, W.C. Deducing protein function by forensic integrative cell biology. *PLoS Biol.* **2013**, *11*, e1001742. [CrossRef] [PubMed]
124. Hanson, A.D.; Pribat, A.; Waller, J.C.; de Crécy-Lagard, V. "Unknown" proteins and "orphan" enzymes: The missing half of the engineering parts list—And how to find it. *Biochem. J.* **2010**, *425*, 1–11. [CrossRef] [PubMed]
125. Pellegrini, M.; Thompson, M.; Fierro, J.; Bowers, P. Computational method to assign microbial genes to pathways. *J. Cell. Biochem. Suppl.* **2001**, *84* (Suppl. 37), 106–109. [CrossRef] [PubMed]
126. Dandekar, T.; Snel, B.; Huynen, M.; Bork, P. Conservation of gene order: A fingerprint of proteins that physically interact. *Trends Biochem. Sci.* **1998**, *23*, 324–328. [CrossRef]
127. Yanai, I.; Mellor, J.C.; de Lisi, C. Identifying functional links between genes using conserved chromosomal proximity. *Trends Genet.* **2002**, *18*, 176–179. [CrossRef]
128. Price, M.N.; Huang, K.H.; Arkin, A.P.; Alm, E.J. Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Res.* **2005**, *15*, 809–819. [CrossRef] [PubMed]
129. Ream, D.C.; Bankapur, A.R.; Friedberg, I. An event-driven approach for studying gene block evolution in bacteria. *Bioinformatics* **2015**, *31*, 2075–2083. [CrossRef] [PubMed]
130. Junier, I.; Rivoire, O. Conserved units of co-expression in bacterial genomes: An evolutionary insight into transcriptional regulation. *PLoS ONE* **2016**, *11*, e0155740. [CrossRef] [PubMed]
131. Lawrence, J.G.; Roth, J.R. Selfish operons: Horizontal transfer may drive the evolution of gene clusters. *Genetics* **1996**, *143*, 1843–18460. [PubMed]
132. Henry, C.S.; Lerma-Ortiz, C.; Gerdes, S.Y.; Mullen, J.D.; Colasanti, R.; Zhukov, A.; Frelin, O.; Thiaville, J.J.; Zallot, R.; Niehaus, T.D.; et al. Systematic identification and analysis of frequent gene fusion events in metabolic pathways. *BMC Genom.* **2016**, *17*, 473. [CrossRef] [PubMed]
133. Green, M.L.; Karp, P.D. Using genome-context data to identify specific types of functional associations in pathway/genome databases. *Bioinformatics* **2007**, *23*, i205–i211. [CrossRef] [PubMed]
134. Moreno-Hagelsieb, G. The power of operon rearrangements for predicting functional associations. *Comput. Struct. Biotechnol. J.* **2015**, *13*, 402–406. [CrossRef] [PubMed]
135. Szklarczyk, D.; Franceschini, A.; Wyder, S.; Forslund, K.; Heller, D.; Huerta-Cepas, J.; Simonovic, M.; Roth, A.; Santos, A.; Tsafou, K.P.; et al. STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **2015**, *43*, D447–D452. [CrossRef] [PubMed]
136. Overbeek, R.; Fonstein, M.; D'Souza, M.; Pusch, G.D.; Maltsev, N. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 2896–2901. [CrossRef] [PubMed]
137. Dehal, P.S.; Joachimiak, M.P.; Price, M.N.; Bates, J.T.; Baumohl, J.K.; Chivian, D.; Friedland, G.D.; Huang, K.H.; Keller, K.; Novichkov, P.S.; et al. MicrobesOnline: An integrated portal for comparative and functional genomics. *Nucleic Acids Res.* **2010**, *38*, D396–D400. [CrossRef] [PubMed]
138. Oberto, J. SyntTax: A web server linking synteny to prokaryotic taxonomy. *BMC Bioinform.* **2013**, *14*, 4. [CrossRef] [PubMed]
139. Vallenet, D.; Labarre, L.; Rouy, Z.; Barbe, V.; Bocs, S.; Cruveiller, S.; Lajus, A.; Pascal, G.; Scarpelli, C.; Médigue, C. MaGe: A microbial genome annotation system supported by synteny results. *Nucleic Acids Res.* **2006**, *34*, 53–65. [CrossRef] [PubMed]
140. Goyer, A.; Hasnain, G.; Frelin, O.; Ralat, M.A.; Gregory, J.F.; Hanson, A.D. A cross-kingdom Nudix enzyme that pre-empts damage in thiamin metabolism. *Biochem. J.* **2013**, *454*, 533–542. [CrossRef] [PubMed]
141. Klaus, S.M.J.; Wegkamp, A.; Sybesma, W.; Hugenholtz, J.; Gregory, J.F.; Hanson, A.D. A nudix enzyme removes pyrophosphate from dihydroneopterin triphosphate in the folate synthesis pathway of bacteria and plants. *J. Biol. Chem.* **2005**, *280*, 5274–5280. [CrossRef] [PubMed]
142. McLennan, A.G. The Nudix hydrolase superfamily. *Cell. Mol. Life Sci.* **2006**, *63*, 123–143. [CrossRef] [PubMed]
143. Gunawardana, D.; Likic, V.A.; Gayler, K.R. A comprehensive bioinformatics analysis of the Nudix superfamily in *Arabidopsis thaliana*. *Comp. Funct. Genom.* **2009**, *2009*, 820381. [CrossRef] [PubMed]
144. Piovesan, D.; Giollo, M.; Leonardi, E.; Ferrari, C.; Tosatto, S.C.E. INGA: Protein function prediction combining interaction networks, domain assignments and sequence similarity. *Nucleic Acids Res.* **2015**, *43*, W134–W140. [CrossRef] [PubMed]

145. Bastard, K.; Smith, A.A.T.; Vergne-Vaxelaire, C.; Perret, A.; Zaparucha, A.; de Melo-Minardi, R.; Mariage, A.; Boutard, M.; Debard, A.; Lechaplais, C.; et al. Revealing the hidden functional diversity of an enzyme family. *Nat. Chem. Biol.* **2014**, *10*, 42–49. [CrossRef] [PubMed]
146. Huang, H.; Pandya, C.; Liu, C.; Al-Obaidi, N.F.; Wang, M.; Zheng, L.; Toews Keating, S.; Aono, M.; Love, J.D.; Evans, B.; et al. Panoramic view of a superfamily of phosphatases through substrate profiling. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, E1974–E1983. [CrossRef] [PubMed]
147. Akiva, E.; Brown, S.; Almonacid, D.E.; Barber, A.E.; Custer, A.F.; Hicks, M.A.; Huang, C.C.; Lauck, F.; Mashiyama, S.T.; Meng, E.C.; et al. The Structure-Function Linkage Database. *Nucleic Acids Res.* **2014**, *42*, D521–D530. [CrossRef] [PubMed]
148. Furnham, N.; Sillitoe, I.; Holliday, G.L.; Cuff, A.L.; Rahman, S.A.; Laskowski, R.A.; Orengo, C.A.; Thornton, J.M. FunTree: A resource for exploring the functional evolution of structurally defined enzyme superfamilies. *Nucleic Acids Res.* **2012**, *40*, D776–D782. [CrossRef] [PubMed]
149. Furnham, N.; Dawson, N.L.; Rahman, S.A.; Thornton, J.M.; Orengo, C.A. Large-scale analysis exploring evolution of catalytic machineries and mechanisms in enzyme superfamilies. *J. Mol. Biol.* **2016**, *428*, 253–267. [CrossRef] [PubMed]
150. Thöny, B.; Auerbach, G.; Blau, N. Tetrahydrobiopterin biosynthesis, regeneration and functions. *Biochem. J.* **2000**, *347*, 1–16. [CrossRef] [PubMed]
151. Inoue, Y.; Kawasaki, Y.; Harada, T.; Hatakeyama, K.; Kagamiyama, H. Purification and cDNA cloning of rat 6-pyruvoyl-tetrahydropterin synthase. *J. Biol. Chem.* **1991**, *266*, 20791–20796. [PubMed]
152. Kong, J.S.; Kang, J.-Y.; Kim, H.L.; Kwon, O.-S.; Lee, K.H.; Park, Y.S. 6-Pyruvoyltetrahydropterin synthase orthologs of either a single or dual domain structure are responsible for tetrahydrobiopterin synthesis in bacteria. *FEBS Lett.* **2006**, *580*, 4900–4904. [CrossRef] [PubMed]
153. Dittrich, S.; Mitchell, S.L.; Blagborough, A.M.; Wang, Q.; Wang, P.; Sims, P.F.G.; Hyde, J.E. An atypical orthologue of 6-pyruvoyltetrahydropterin synthase can provide the missing link in the folate biosynthesis pathway of malaria parasites. *Mol. Microbiol.* **2008**, *67*, 609–618. [CrossRef] [PubMed]
154. Pribat, A.; Jeanguenin, L.; Lara-Núñez, A.; Ziemak, M.J.; Hyde, J.E.; de Crécy-Lagard, V.; Hanson, A.D. 6-pyruvoyltetrahydropterin synthase paralogs replace the folate synthesis enzyme dihydroneopterin aldolase in diverse bacteria. *J. Bacteriol.* **2009**, *191*, 4158–4165. [CrossRef] [PubMed]
155. McCarty, R.M.; Somogyi, A.; Bandarian, V. *Escherichia coli* QueD is a 6-carboxy-5,6,7,8-tetrahydropterin synthase. *Biochemistry* **2009**, *48*, 2301–2303. [CrossRef] [PubMed]
156. Reader, J.S.; Metzgar, D.; Schimmel, P.; de Crécy-Lagard, V. Identification of four genes necessary for biosynthesis of the modified nucleoside queuosine. *J. Biol. Chem.* **2004**, *279*, 6280–6285. [CrossRef]
157. Markowitz, V.M.; Chen, I.-M.A.; Palaniappan, K.; Chu, K.; Szeto, E.; Pillay, M.; Ratner, A.; Huang, J.; Woyke, T.; Huntemann, M.; et al. IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Res.* **2014**, *42*, D560–D567. [CrossRef]
158. Markowitz, V.M.; Chen, I.-M.A.; Chu, K.; Pati, A.; Ivanova, N.N.; Kyrpides, N.C. Ten years of maintaining and expanding a microbial genome and metagenome analysis system. *Trends Microbiol.* **2015**, *23*, 730–741. [CrossRef] [PubMed]
159. Aziz, R.K.; Bartels, D.; Best, A.A.; DeJongh, M.; Disz, T.; Edwards, R.A.; Formsma, K.; Gerdes, S.; Glass, E.M.; Kubal, M.; et al. The RAST Server: Rapid annotations using subsystems technology. *BMC Genom.* **2008**, *9*, 75. [CrossRef] [PubMed]
160. Brettin, T.; Davis, J.J.; Disz, T.; Edwards, R.A.; Gerdes, S.; Olsen, G.J.; Olson, R.; Overbeek, R.; Parrello, B.; Pusch, G.D.; et al. RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Sci. Rep.* **2015**, *5*, 8365. [CrossRef] [PubMed]
161. Tatusova, T.; DiCuccio, M.; Badretdin, A.; Chetvernin, V.; Ciufo, S.; Li, W. Prokaryotic genome annotation pipeline. In *The NCBI Handbook [Internet]*; National Center for Biotechnology Information (US): Bethesda, MD, USA, 2013.
162. Kersey, P.J.; Allen, J.E.; Armean, I.; Boddu, S.; Bolt, B.J.; Carvalho-Silva, D.; Christensen, M.; Davis, P.; Falin, L.J.; Grabmueller, C.; et al. Ensembl Genomes 2016: More genomes, more complexity. *Nucleic Acids Res.* **2016**, *44*, D574–D580. [CrossRef] [PubMed]
163. Angiuoli, S.V.; Gussman, A.; Klimke, W.; Cochrane, G.; Field, D.; Garrity, G.; Kodira, C.D.; Kyrpides, N.; Madupu, R.; Markowitz, V.; et al. Toward an online repository of Standard Operating Procedures (SOPs) for (meta)genomic annotation. *OMICS* **2008**, *12*, 137–141. [CrossRef] [PubMed]

164. Huntemann, M.; Ivanova, N.N.; Mavromatis, K.; Tripp, H.J.; Paez-Espino, D.; Palaniappan, K.; Szeto, E.; Pillay, M.; Chen, I.-M.A.; Pati, A.; et al. The standard operating procedure of the DOE-JGI Microbial Genome Annotation Pipeline (MGAP v.4). *Stand. Genom. Sci.* **2015**, *10*, 86. [CrossRef] [PubMed]
165. Mavromatis, K.; Ivanova, N.N.; Chen, I.-M.A.; Szeto, E.; Markowitz, V.M.; Kyrpides, N.C. The DOE-JGI Standard Operating Procedure for the Annotations of Microbial Genomes. *Stand. Genom. Sci.* **2009**, *1*, 63–67. [CrossRef] [PubMed]
166. Markowitz, V.M.; Chen, I.-M.A.; Palaniappan, K.; Chu, K.; Szeto, E.; Grechkin, Y.; Ratner, A.; Anderson, I.; Lykidis, A.; Mavromatis, K.; et al. The integrated microbial genomes system: An expanding comparative analysis resource. *Nucleic Acids Res.* **2010**, *38*, D382–D390. [CrossRef] [PubMed]
167. Meyer, F.; Overbeek, R.; Rodriguez, A. FIGfams: Yet another set of protein families. *Nucleic Acids Res.* **2009**, *37*, 6643–6654. [CrossRef] [PubMed]
168. Pedruzzi, I.; Rivoire, C.; Auchincloss, A.H.; Coudert, E.; Keller, G.; de Castro, E.; Baratin, D.; Cuche, B.A.; Bougueleret, L.; Poux, S.; et al. HAMAP in 2015: Updates to the protein family classification and annotation system. *Nucleic Acids Res.* **2015**, *43*, D1064–D1070. [CrossRef] [PubMed]
169. Jones, P.; Binns, D.; Chang, H.-Y.; Fraser, M.; Li, W.; McAnulla, C.; McWilliam, H.; Maslen, J.; Mitchell, A.; Nuka, G.; et al. InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **2014**, *30*, 1236–1240. [CrossRef] [PubMed]
170. Mitchell, A.; Chang, H.-Y.; Daugherty, L.; Fraser, M.; Hunter, S.; Lopez, R.; McAnulla, C.; McMenamin, C.; Nuka, G.; Pesseat, S.; et al. The InterPro protein families database: The classification resource after 15 years. *Nucleic Acids Res.* **2015**, *43*, D213–D221. [CrossRef] [PubMed]
171. Vilella, A.J.; Severin, J.; Ureta-Vidal, A.; Heng, L.; Durbin, R.; Birney, E. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* **2009**, *19*, 327–335. [CrossRef] [PubMed]
172. Crooks, G.E.; Hon, G.; Chandonia, J.-M.; Brenner, S.E. WebLogo: A sequence logo generator. *Genome Res.* **2004**, *14*, 1188–1190. [CrossRef] [PubMed]
173. Vacic, V.; Iakoucheva, L.M.; Radivojac, P. Two Sample Logo: A graphical representation of the differences between two sets of sequence alignments. *Bioinformatics* **2006**, *22*, 1536–1537. [CrossRef] [PubMed]
174. Brown, S.D.; Gerlt, J.A.; Seffernick, J.L.; Babbitt, P.C. A gold standard set of mechanistically diverse enzyme superfamilies. *Genome Biol.* **2006**, *7*, R8. [CrossRef] [PubMed]
175. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
176. Markowitz, V.M.; Chen, I.M.A.; Palaniappan, K.; Chu, K.; Szeto, E.; Grechkin, Y.; Ratner, A.; Jacob, B.; Huang, J.; Williams, P.; et al. IMG: The integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res.* **2012**, *40*, D115–D122. [CrossRef]
177. Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T.J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J.; et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **2011**, *7*, 539. [CrossRef] [PubMed]
178. Wattam, A.R.; Abraham, D.; Dalay, O.; Disz, T.L.; Driscoll, T.; Gabbard, J.L.; Gillespie, J.J.; Gough, R.; Hix, D.; Kenyon, R.; et al. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.* **2014**, *42*, D581–D591. [CrossRef] [PubMed]
179. Larsson, A. AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **2014**, *30*, 3276–3278. [CrossRef] [PubMed]
180. Zhao, S.; Sakai, A.; Zhang, X.; Vetting, M.W.; Kumar, R.; Hillerich, B.; San Francisco, B.; Solbiati, J.; Steves, A.; Brown, S.; et al. Prediction and characterization of enzymatic activities guided by sequence similarity and genome neighborhood networks. *eLife* **2014**, *3*. [CrossRef] [PubMed]
181. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N.S.; Wang, J.T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13*, 2498–2504. [CrossRef] [PubMed]