Gene 549 (2014) 161–170

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

The evolutionary history of the genes involved in the biosynthesis of the antioxidant ergothioneine

Gary W. Jones, Sean Doyle, David A. Fitzpatrick *

Department of Biology, The National University of Ireland Maynooth, Maynooth, County Kildare, Ireland

ARTICLE INFO

Article history:
Received 24 June 2014
Received in revised form 22 July 2014
Accepted 25 July 2014
Available online 26 July 2014

Keywords:
Ergothioneine
Antioxidant
Fungal
Prokaryotes
Phylogenetics
Fusion gene

ABSTRACT

Ergothioneine (EGT) is a histidine betaine derivative that exhibits antioxidant action in humans. EGT is primarily synthesized by fungal species and a number of bacterial species. A five-gene cluster (*egt*A, *egt*B, *egt*C, *egt*D & *egt*E) responsible for EGT production in *Mycobacteria smegmatis* has recently been identified. The first fungal biosynthetic EGT gene (*NcEgt-1*) has also been identified in *Neurospora crassa*. *NcEgt-1* contains domains similar to those found in *M. smegmatis egt*B and *egt*D. EGT is biomembrane impermeable. Here we inferred the evolutionary history of the EGT cluster in prokaryotes as well as examining the phyletic distribution of Egt-1 in the fungal kingdom. A genomic survey of 2509 prokaryotes showed that the five-gene EGT cluster is only found in the Actinobacteria. Our survey identified more than 400 diverse prokaryotes that contain genetically linked orthologs of *egt*B and *egt*D. Phylogenetic analyses of Egt proteins show a complex evolutionary history and multiple incidences of horizontal gene transfer. Our analysis also identified two independent incidences of a fusion event of *egt*B and *egt*D in bacterial species. A genomic survey of over 100 fungal genomes shows that Egt-1 is found in all fungal phyla, except species that belong to the Saccharomycotina subphylum. This analysis provides a comprehensive analysis of the distribution of the key genes involved in the synthesis of EGT in prokaryotes and fungi. Our phylogenetic inferences illuminate the complex evolutionary history of the genes involved in EGT synthesis in prokaryotes. The potential to synthesize EGT is a fungal trait except for species belonging to the Saccharomycotina subphylum.

1. Introduction

Ergothioneine (EGT) is a histidine betaine derivative with a thiol group located on the C_2 atom of the imidazole ring (Genghof et al., 1956; Hartman, 1990; Melville et al., 1957). EGT was first isolated from the ergot fungus *Claviceps purpurea* (Tanret, 1909) and subsequent observational studies suggested that it is primarily synthesized by fungal species and a number of bacterial species particularly those belonging to the Actinobacterial and Cyanobacterial phyla (Genghof, 1970; Genghof and Vandamme, 1964; Genghof et al., 1956; Pfeiffer et al., 2011).

Although EGT is present in plants and animals they do not synthesize it but rather obtain it from nutrients. For example, plants acquire EGT through their roots and sometimes via actinomycete symbionts (Park et al., 2010). Animals acquire EGT from food including mushrooms, garlic, wheat, oats and beans which have been shown to have concentrations of EGT ranging from 210 ng mg^{-1} to 2600 ng mg^{-1} (Dubost et al., 2007). After ingestion EGT is retained with minimal metabolism (Pfeiffer et al., 2011). In humans, EGT is concentrated in erythrocyte progenitor cells, monocytes, the intestine and kidneys. EGT is cell membrane impermeable and recent human studies have identified a specific plasma membrane bound organic cation transporter known as OCTN1, which is encoded by the gene SLC22A4 (Grundemann et al., 2005). Polymorphisms in SLC22A4 have been associated with diabetes (Santiago et al., 2006) and susceptibility to chronic inflammatory diseases such as Crohn's disease (Fisher et al., 2006; Leung et al., 2006; Peltekova et al., 2004). A specific EGT transporter suggests a beneficial role for EGT and multiple studies support an antioxidant function for EGT but its specific mode of action remains unclear (Cheah and Halliwell, 2012; Hartman, 1990).

Mycobacteria do not synthesize the thiol glutathione (GSH), which is known for its efficient detoxification of free radicals as well as reactive oxygen and nitrogen species. Instead it produces two low-molecular-weight thiols, mycothiol (MSH) and EGT (Genghof and Vandamme, 1964; Newton et al., 1995, 1996). In *Mycobacterium tuberculosis* evidence suggests that MSH is involved in detoxifying reactive oxygen species (Vilcheze et al., 2008). In *Mycobacteria smegmatis* MSH deficient mutants, the levels of the organic hydroperoxide resistance protein and ERG are elevated, suggesting that ERG may partly compensate for the loss of MSH and thus have a role as an antioxidant (Ta et al.,

Abbreviations: EGT, ergothioneine; HSP, highest scoring pair; GSH, glutathione; MSH, mycothiol.

* Corresponding author.
E-mail addresses: Gary.Jones@nuim.ie (G.W. Jones), Sean.Doyle@nuim.ie (S. Doyle), david.fitzpatrick@nuim.ie (D.A. Fitzpatrick).

http://dx.doi.org/10.1016/j.gene.2014.07.065
0378-1119/

2011). ERG has also been implicated in modulating the immune response (Rahman et al., 2003) and in the inhibition of metalloenzymes, preventing oxidation of DNA and protein due to its metal-chelating properties (Zhu et al., 2010), this implies that it may also act as a virulence factor. Initial investigations into the function of EGT in fungi showed that in *Neurospora crassa* it helps protect conidia during the quiescent period between conidiogenesis and germination, and protects conidia during the germination process from the toxicity of peroxide (Bello et al., 2012).

The genes (*egt*A, *egt*B, *egt*C, *egt*D & *egt*E) responsible for EGT production in *M. smegmatis* have recently been identified (Seebeck, 2010). These genes are found adjacent to one another in a five-gene cluster but are predicted not to be essential for growth of *M. tuberculosis* laboratory strain H37Rv (Griffin et al., 2011; Sassetti et al., 2003). Seebeck (2010) cloned EgtD and has shown it to be a histidine methyltransferase that converts histidine to hercynine in an *S*-adenosyl methionine (SAM) dependent manner (Fig. 1-A). Sequence similarity of EgtA to γ-glutamylcysteine ligase suggests that γ-Glu-Cys rather than Cys is a sulphur donor (Seebeck, 2010). This observation was confirmed when cloned EgtB (contains an Fe(II) binding site) was assayed with hercynine and γ-Glu-Cys in the presence of $FeSO_4$ and shown to produce *S*-hercynyl-γ-glutamylcysteine (Fig. 1-A) (Seebeck, 2010). Addition of cloned EgtC to the reaction generated hercynylcysteine sulfoxide and cloned EgtE (β-lyase) in the presence of pyridoxal-5′-phosphate produced EGT (Fig. 1-A) (Seebeck, 2010).

Recently the first fungal biosynthetic EGT gene (*NcEgt-1*) was identified in *N. crassa* (Bello et al., 2012). *NcEgt-1* catalyzes the first two steps of EGT biosynthesis from histidine to hercynine to hercynylcysteine sulfoxide (Fig. 1-B). Comparisons between wild type and *NcEgt-1* indicate that EGT plays an important protective role against the toxicity of peroxide in conidia during germination (Bello et al., 2012). Interestingly *NcEgt-1* contains domains similar to those found in *M. smegmatis egt*B and *egt*D and is most likely the result of a fusion between ancestral fungal EgtB and EgtD genes (Bello et al., 2012).

No analysis to date has attempted to fully uncover the evolutionary history of the 5-gene *egt* cluster in prokaryotes or indeed that of Egt-1 in the fungal kingdom. As providing a comprehensive analysis of the phyletic distribution of these genes in the tree of life, we have performed in-depth phylogenetic analyses of these genes. Our results give a detailed overview of the distribution of the key genes involved in the synthesis of EGT in prokaryotes and fungi. Our phylogenetic inferences illuminate the complex evolutionary history of the genes involved in EGT synthesis in prokaryotes.

2. Methods

2.1. Sequence data and database searches

Amino acid sequences from all completely sequenced prokaryotic genomes were obtained from the NCBI ftp site. The list of the genomes utilized and their taxonomic affiliations are listed in Additional file 1-A. Complete bacterial genomes were utilized to ensure that potential EGT clusters could be identified. In total our dataset contained 7,850,632 amino acid sequences from 2509 genomes (Additional file 1-A). The Proteobacterial phylum is the most densely sampled accounting for ~42% of all genomes, followed by the Firmicutes phylum ~21% and the Actinobacteria phylum ~10.5% (Additional file 1-A).

The five genes (*egt*ABCDE) from the *Mycobacterium smegmatis* JS623 EGT cluster were used as database query sequences (GenBank GI numbers 433650597, 433650596, 433650595, 433650594 and 433650593 respectively). Taking one EGT protein at a time, putative orthologs were identified using a reciprocal BlastP (Altschul et al., 1997) search with a cutoff expectation (E) value of 10^{-5}. Each EGT gene was searched against an individual bacterial genome. The top significant hit was recorded and searched back against the *M. smegmatis* genome to ensure a reciprocal top hit. Putative orthologs are listed in Additional file 1-B.

Our fungal protein dataset consisted of 103 genomes and 1001217 individual genes (Additional file 1-C). Where available, data was obtained

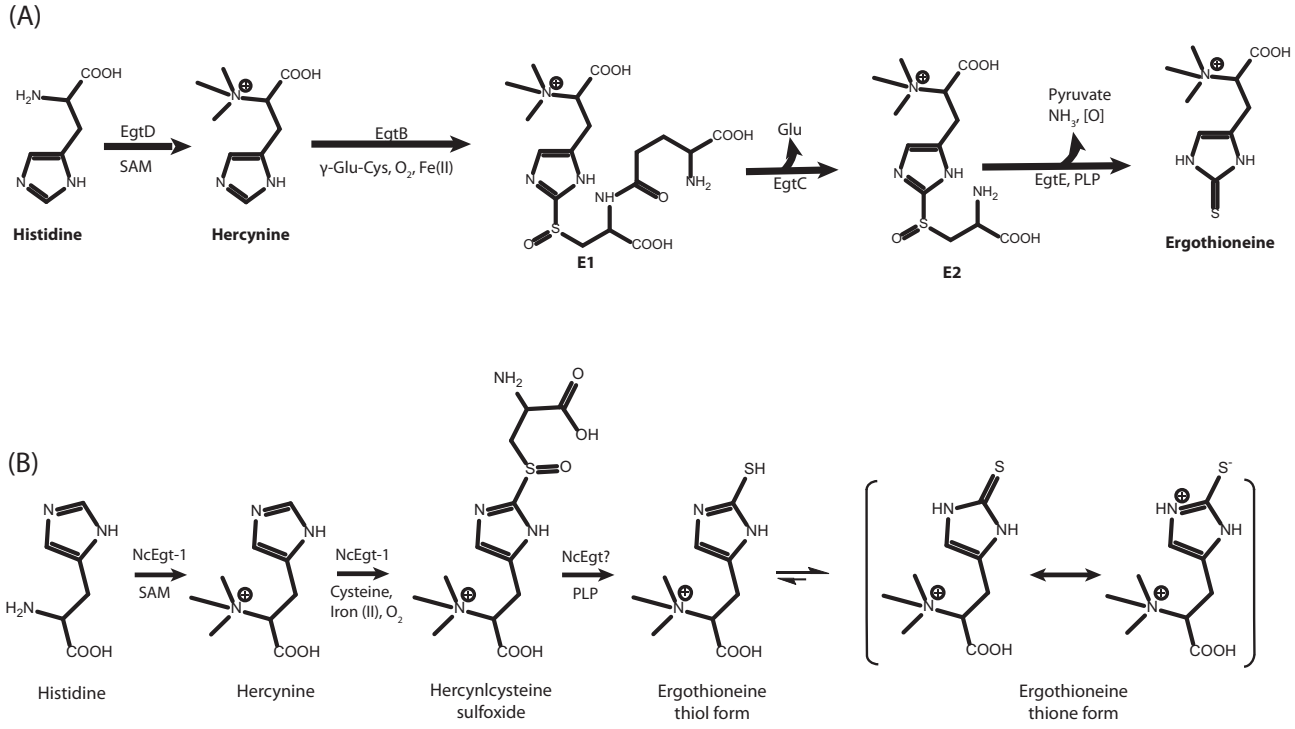


Fig. 1. A) Reaction sequence of ergothioneine biosynthesis in *Mycobacterium smegmatis*, redrawn from (Seebeck, 2010). B) The proposed ergothioneine biosynthetic pathway in *Neurospora crassa*.
Redrawn from Bello et al. (2012).

from the NCBI fungal genome FTP site. The remaining data was downloaded from the relevant sequencing centres (Additional file 1-C). The *N. crassa* EGT (NcEgt-1) gene was searched against each individual fungal genome. The top significant hit was recorded and searched back against the *N. crassa* genome to ensure a reciprocal top hit. We set a cut-off of 70% for the length of the highest scoring pair (HSP). Top significant database hits above this cutoff were deemed to be putative orthologs and are listed in Additional file 1-C. To account for incomplete genome annotation, assemblies that did not contain an ortholog of EGT were searched again using a tblastn strategy.

2.2. Phylogenetic methods

All EGT protein alignments were aligned using MUSCLE (v3.6) (Edgar, 2004), with the default settings. Obvious alignment ambiguities were manually corrected.

Phylogenetic relationships were inferred for full and representative datasets using maximum likelihood methods. Appropriate protein models of substitution were selected for each gene family using the Bayesian information criterion implemented in ProtTest3 (Darriba et al., 2011). Optimum models and associated parameters for all protein families are summarized in Additional file 1-D. One hundred bootstrap replicates were then carried out with the appropriate protein model using the software programme PHYML (Guindon and Gascuel, 2003) and summarized using the majority-rule consensus method.

2.3. Analysis of the EGT cluster in prokaryotes

We utilized previously described perl scripts (Martin and McInerney, 2009) to identify clusters of EGT genes. In brief, protein sequences for EgtA, EgtB, EgtC, EgtD and EgtE from *M. smegmatis* where queried using BlastP against each of the 438 bacterial genomes in our dataset that contain a putative ortholog of *egt*B & *egt*D (Additional file 1-B). If two *egt* genes were found to have no more than five intervening genes between them, then such genes were considered to be a linked pair. In total, 404 genomes that have at least one pair of genes linked were identified (Additional file 1-B).

3. Results and discussion

3.1. The 5-gene EGT cluster is only found in the Actinomycetes

Previous studies have shown that EgtB and EgtD are key enzymes in the production of ergothioneine (Seebeck, 2010; Sao Emani et al., 2013). From our original dataset of 2509 prokaryotic genomes (Additional file 1-A), 438 were found to contain a putative ortholog of *egt*B and *egt*D (Additional file 1-B). Interestingly 404 of these displayed evidence of linkage as they were found to have no more than five intervening genes between them (Additional file 1-B). Closer inspection shows that the EGT cluster is specific to members of the Actinobacteria as most species outside this phylum are missing orthologs of *egt*A, *egt*C and *egt*E (Additional file 1-B). Most non-Actinobacterial species, which contain an ortholog of one of these three genes, do not display linkage relative to *egt*B or *egt*D. However, there are some exceptions to this observation as 9 Cyanobacterial species, 3 Proteobacterial species and one member of the Chloroflexi phylum contain an ortholog of *egtC*, which is clustered beside *egt*B and *egt*D (Additional file 1-B and Fig. 2).

Previous studies have shown that bacterial species that are missing orthologs of *egt*A, *egt*C and *egt*E are still able to produce EGT (Pfeiffer et al., 2011) indicating that they may not be universally essential enzymes in its synthesis. Conceivably, their enzymatic function, may be complemented by other enzymes or chemical means. We examined the expression profiles of both *egt*B and *egt*D in *M. tuberculosis* H37Rv using the TB database (Galagan et al., 2010). While the expression of both genes positively correlates with one another (0.47242), there is no evidence to suggest that *egt*A, *egt*C or *egt*E expression is positively correlated with that of either *egt*B or *egt*D, thereby lending further support to the hypothesis that their functions in other species could be complemented by other means.

3.2. Phylogenetic analysis of prokaryote EGT genes

The wide phyletic distribution of *egt*B and *egt*D suggests that ergothioneine biosynthesis is a common trait amongst Actinobacteria and Cyanobacteria and also occurs in certain Bacteroidetes, Proteobacteria, Acidobacteria and even Archaeal species (Additional file 1-B). Such a wide phyletic distribution suggests that ergothioneine may play important physiological roles in many bacterial species. However, while the phyletic distribution of these enzymes is wide, it does appear to be patchy and is most likely the result of rampant horizontal gene transfer (HGT). To try and uncover the evolutionary history of the genes involved in the production of ergothioneine we undertook an in-depth phylogenetic analysis.

3.3. EgtA phylogeny

A BlastP analysis identified 203 EgtA orthologs in our bacterial database (Additional file 1-B). Phylogenetic reconstruction infers two strongly supported (100% Bootstrap support (BP)) monophyletic clades (Additional file 2). There is also a single representative from the Planctomycetes phylum (*Isosphaera pallida*).

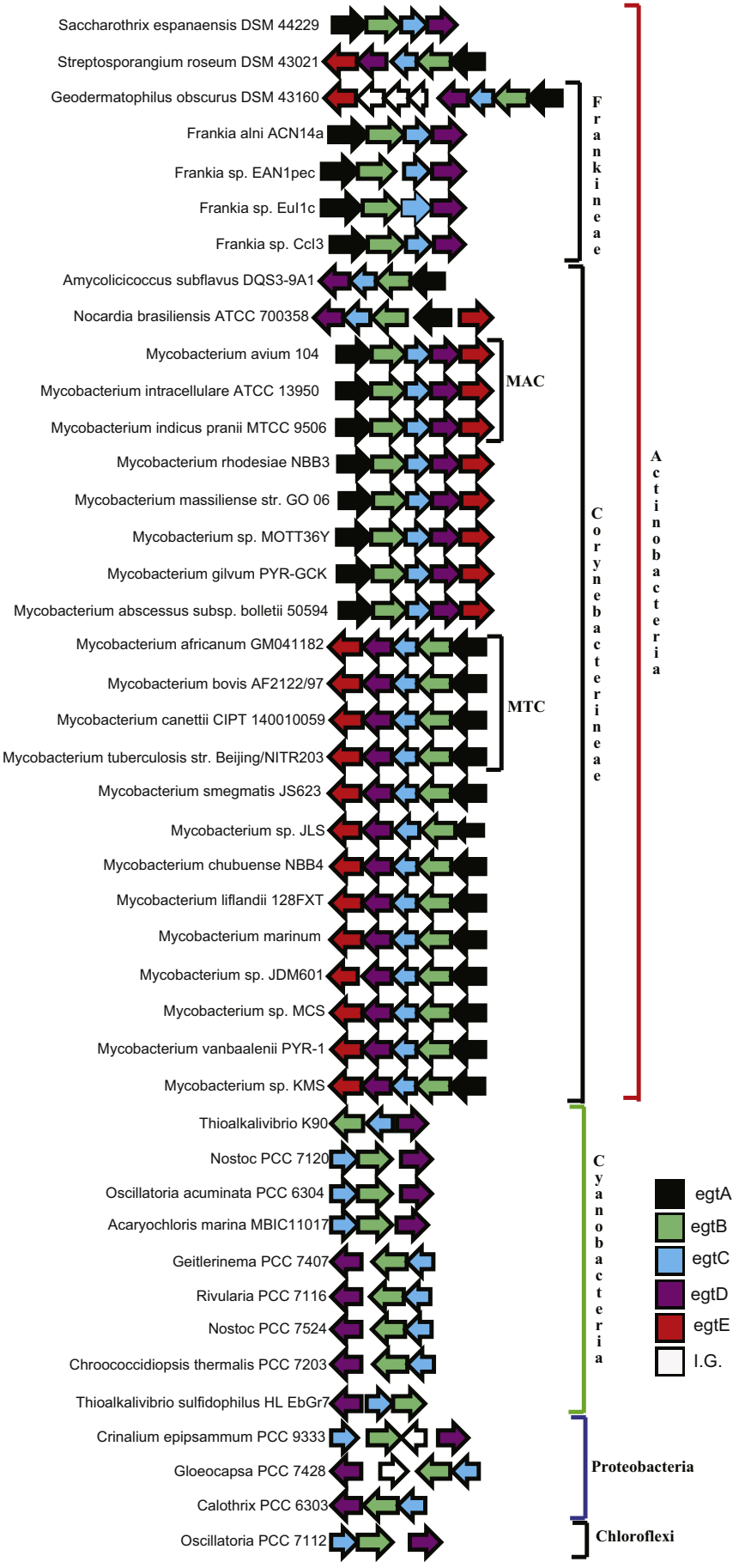
The first monophyletic clade is Actinobacterial specific (Additional file 2). Of the 121 Actinobacterial genomes that contain a copy of *egt*B and *egt*D, 112 have an ortholog of *egtA* (Additional file 1-B). For the 10 Actinobacterial species that lack the *egt*A ortholog, gene loss is evident for individual species but is not specific for a particular Actinobacterial subclass. The second clade is Proteobacterial specific and contains representatives from the Proteobacteria phylum, specifically the γ-, δ-proteobacteria but primarily the α-proteobacteria classes. Of the 210 Proteobacterial genomes that contain a copy of *egt*B and *egt*D, 90 have an ortholog of *egt*A (Additional file 1-B). We did not locate putative *egt*A orthologs for β or ε-proteobacterial classes (Additional file 1-B).

There are a number of scenarios that can be invoked to explain the evolutionary history of *egt*A. The patchy phyletic distribution of *egt*A within the Proteobacteria phylum, coupled with its widespread distribution within the Actinobacterial phylum is indicative of an ancient gene transfer event from an ancestral Actinobacterial species into an ancestral α-proteobacterial species, followed by subsequent α to γ- and δ-proteobacterial transfers. Other scenarios include a proteobacterial origin for *egt*A and transfer into an ancestral Actinobacterial species. However, this is a less parsimonious proposition as it requires gene losses from the β and ε-proteobacterial lineages.

3.4. EgtC phylogeny

A BlastP analysis identified 205 EgtC orthologs in our bacterial database (Additional file 1-B). Phylogenetic reconstruction infers a strongly supported (74% BP) Actinobacterial, Cyanobacterial (98%) and Proteobacterial clades (100% BP) (Additional file 3). Of the 121 Actinobacterial genomes that contain a copy of *egt*B and *egt*D, 114 have an ortholog of *egt*C (Additional file 1-B). The 7 species (*Acidimicrobidae bacterium*, *Corynebacterium halotolerans*, *Gordonia bronchialis*, *Gordonia* KTR9, *Segniliparus rotundus*, *Nocardioides* JS614 and *Conexibacter woesei*) that are missing *egt*C are also missing *egt*A and 3 of these (*C. halotolerans*, *Nocardioides* JS614 and *C. woesei*) are also missing *egt*E (Additional file 1-B). Loss of these genes seems to be species rather than lineage specific as they are not specific to one particular Actinobacterial subclass.

The Cyanobacterial clade (98% BP (Additional file 3)) contains members from the Oscillatoriophycideae subclass and Nostocales, Pleurocapsales and Gloeobacteria orders (Additional file 1-B). Of the 58 Cyanobacterial genomes that contain a copy of *egt*B and *egt*D, 38


Saccharothrix espanaensis DSM 44229
Streptosporangium roseum DSM 43021
Geodermatophilus obscurus DSM 43160
Frankia alni ACN14a
Frankia sp. EAN1pec
Frankia sp. Eul1c
Frankia sp. Ccl3
Amycolicicoccus subflavus DQS3-9A1
Nocardia brasiliensis ATCC 700358
Mycobacterium avium 104
Mycobacterium intracellulare ATCC 13950
Mycobacterium indicus pranii MTCC 9506
Mycobacterium rhodesiae NBB3
Mycobacterium massiliense str. GO 06
Mycobacterium sp. MOTT36Y
Mycobacterium gilvum PYR-GCK
Mycobacterium abscessus subsp. bolletii 50594
Mycobacterium africanum GM041182
Mycobacterium bovis AF2122/97
Mycobacterium canettii CIPT 140010059
Mycobacterium tuberculosis str. Beijing/NITR203
Mycobacterium smegmatis JS623
Mycobacterium sp. JLS
Mycobacterium chubuense NBB4
Mycobacterium liflandii 128FXT
Mycobacterium marinum
Mycobacterium sp. JDM601
Mycobacterium sp. MCS
Mycobacterium vanbaalenii PYR-1
Mycobacterium sp. KMS
Thioalkalivibrio K90
Nostoc PCC 7120
Oscillatoria acuminata PCC 6304
Acaryochloris marina MBIC11017
Geitlerinema PCC 7407
Rivularia PCC 7116
Nostoc PCC 7524
Chroococcidiopsis thermalis PCC 7203
Thioalkalivibrio sulfidophilus HL EbGr7
Crinalium epipsammum PCC 9333
Gloeocapsa PCC 7428
Calothrix PCC 6303
Oscillatoria PCC 7112
Frankineae
MAC
MTC
Corynebacterineae
Actinobacteria
Cyanobacteria
Proteobacteria
Chloroflexi
egtA
egtB
egtC
egtD
egtE
I.G.

have an ortholog of *egt*C (Additional file 1-B). The species that are missing *egt*C are not specific to one particular subclass indicating species-specific losses.

The Proteobacteria clade (100% BP, (Additional file 3)) contains representatives from the Proteobacterial phylum, specifically the γ-, α- but primarily β-proteobacteria classes (Additional file 1-B). Of the 210 Proteobacterial genomes that contain a copy of *egt*B and *egt*D, 52 have an ortholog of *egt*C (Additional file 1-B). We did not locate putative *egt*C orthologs for δ or ε-proteobacterial classes (Additional file 1-B).

Overall there is little evidence of recent inter phyla HGT of *egt*C orthologs. There is one exception as there is evidence of interphylum HGT from a Proteobacterial source into the Actinobacterial species *Gordonia polyisoprenivorans*. Our phylogeny infers that the *G. polyisoprenivorans egt*C ortholog is nested within the Proteobacterial clade and specifically beside a β-proteobacterial clade ((Additional file 3), 75% BP). Closer inspection of the gene order in *G. polyisoprenivorans* shows it contains four *egt* orthologs (*egt*B, C, D and E). However, unlike most other Actinobacteria, there is only evidence of clustering for two of the *egt* genes (*egt*B and *egt*D). Interestingly, there are two other *Gordonia* species present in our analysis (*G. bronchialis* and *Gordonia* KTR9) (Additional file 1-A). Neither of these species contains an ortholog of *egt*C suggesting an ancestral loss of the Actinobacterial *egt*C from a *Gordonia* ancestor and a subsequent independent HGT event of a β-proteobacterial ortholog into *G. polyisoprenivorans*.

The patchy phyletic distribution of *egt*C within the Proteobacterial and Cyanobacterial phyla, coupled with its widespread distribution within the Actinobacteria phylum is indicative of an ancient HGT event from an ancestral Actinobacterial species into an ancestral Proteobacterial and Cyanobacterial species followed by species-specific gene losses in these phyla.

3.5. egtE phylogeny

Our BlastP analysis identified 68 *egt*E orthologs in our bacterial database (Additional file 1-B). All of these orthologs are exclusively from the Actinobacteria phylum (Additional file 1-B, Additional file 4). The majority of species belonging to the Frankineae and Corynebacterineae suborders contain an ortholog of *egt*E whereas species belonging to the Micromonosporineae, Propionibacterineae, Pseudonocardineae, Streptomycineae and Streptosporangineae suborders are all missing the *egt*E gene (Additional file 1-B).

3.6. EgtB & EgtD phylogenies

Because of their importance in the synthesis of ergothioneine (Seebeck, 2010; Sao Emani et al., 2013), the EgtB and EgtD phylogenies will be discussed in tandem.

A BlastP analysis identified 438 species that contain an ortholog of both *egt*B and *egt*D. Of the 263 Actinobacterial genomes in our original dataset (2509 genomes, Additional file 1-A), 122 (~46%) contain a copy of *egt*B and *egt*D. Similarly 210 of the 1034 (~20%) Proteobacterial genomes and 58 of the 263 (~22%) Cyanobacterial genomes in our original dataset contain copies of both *egt*B and *egt*D (Additional file 1-B). These three phyla account for ~73% of all species considered in our phylogenetic analyses (Additional file 1-B).

We reconstructed phylogenetic trees for EgtB and EgtD (Additional files 5 and 6). Both protein families appear to be heterogeneous, with an average pairwise identity of 39.51% and 41.77% for EgtB and EgtD proteins respectively. As a result many clades are poorly supported making inferences regarding the evolutionary history of these gene families difficult.

To help visualize the data, smaller representative datasets of 230 species were generated for each gene family and phylogenetic trees reconstructed (Figs. 3 & 4). For ease we will refer to the smaller representative phylogenies from this point forward, as they are congruent with the phylogenies derived from all 438 taxa (Additional files 5 & 6).

Of the 154 Archaeal genomes in our dataset only four (*Cenarchaeum symbiosum*, *Candidatus Nitrosopumilus koreensis*, *Candidatus Nitrosopumilus* and *Nitrosopumilus maritimus*) contain copies of *egt*B and *egt*D (Additional file 1-B). Both genes are genetically linked in all four species (Additional file 1-B). These four species are all members of the Thaumarchaeota phylum and are grouped together in a strongly supported monophyletic clade (100% BP monophyletic clades, Clade-1 Figs. 3 & 4). There is a fifth species (*Nitrososphaera gargensis*) from this phylum represented in our dataset but it does not contain a copy of either *egt*B or *egt*D (Additional file 1-A & B). This patchy phyletic distribution within the Archaeal lineage would indicate a putative horizontal gene transfer event into the Thaumarchaeotal ancestor of these species, followed by gene loss in *N. gargensis*. Due to the poorly supported inferences in parts of our phylogenies it is difficult to confidently locate the putative donor of *egt*B & *egt*D into the Thaumarchaeota phylum. Furthermore, searching the Archaeal copies of EgtB & EgtD back against our bacterial dataset we do not see a consistent non-Archaeal top BlastP search hit.

Of the 95 Bacteroidetes genomes represented in our dataset 27 have copies of *egt*B & *egt*D and in all cases they are genetically linked (Additional file 1-A & B). Three Bacteroidetes classes (Cytophagia, Flavobacteriia and Sphingobacteriia) are represented. All Bacteroidetes species are found in monophyletic clades (71% & 89% BP respectively) in both EgtB & EgtD phylogenies (Clade-2, Figs. 3 & 4). Both phylogenies support a strongly supported (100% BP) sister group relationship with 3 Proteobactrerial species (*Shewanella woodyi*, *Methylophaga* JAM1 and *Teredinibacter turnerae*). All 3 species are γ-proteobacterial. This is an example of putative HGT of EgtB & EgtD in tandem but it is not possible to confidently infer the donor species. The sister group relationships amongst other clades in this part of the phylogeny are poor. One possible explanation is an ancient HGT from a Proteobacterial species into the ancestor of a Bacteroidetes species followed by a more recent HGT from the Bacteroidetes represented here back into *S. woodyi*, *Methylophaga* JAM1 and *T. turnerae*. Another explanation is that the Bacteroidetes ancestor already had *egt*B & *egt*D but they have been lost multiple times through speciation events and that the γ-proteobacteria species recently gained the Bacteroidetes orthologs.

Our phylogenetic analyses of EgtB & EgtD infer two strongly supported clades composed primarily of Actinobacterial species (Clade 3, Figs. 3 & 4). With respect to EgtB, a small number of species (*S. rotundus*, *C. halotolerans*, *G. bronchialis*, *A. bacterium* and *Nocardioides*) are found grouped outside this monophyletic clade (Fig. 3, Clade-4). These species all belong to the order Actinomycetales but no unique sub-order is observed. The fact that they are grouped with proteobacterial species and specifically two α-proteobacteria species (*Beijerinckia indica* and *Parvibaculum lavamentivorans*) would suggest that they may be α-proteobacterial in origin, however bootstrap support values are very low (31%) and make it impossible to confidently infer a donor species. Interestingly these species are found in a monophyletic Actinobacterial clade in our EgtD phylogeny (Fig. 4, Clades 3 & 4). Furthermore even though the *egt*B & *egt*D orthologs in these species have a mosaic history they are genetically linked (Additional file 1-B). This provides further evidence that there is a selective pressure to maintain these genes beside one another in the genome even following HGT.

There is evidence that the ancestor of five δ-proteobacterial species (*Myxococcus fulvus*, *Myxococcus stipitatus*, *Myxococcus xanthus*,

Fig. 2. Analysis of the five-gene ergothioneine cluster in prokaryotes. The cluster is widespread in Actinobacterial species and representative genomes are shown. *egt*A, *egt*C and *egt*E are missing from most prokaryotes but a small number of Cyanobacterial, Proteobacterial and Chloroflexi species have orthologs of *egt*C and they are genetically linked with *egt*B and *egt*D. Each arrow represents a gene, with the name of the gene given in the legend. I.G. refers to an intervening gene, which is a gene in the cluster that is not involved in the biosynthesis of ergothioneine. MAC refers to species belonging to the *Mycobacterium avium* complex and MTC refers to species belonging to the *Mycobacterium tuberculosis* complex.

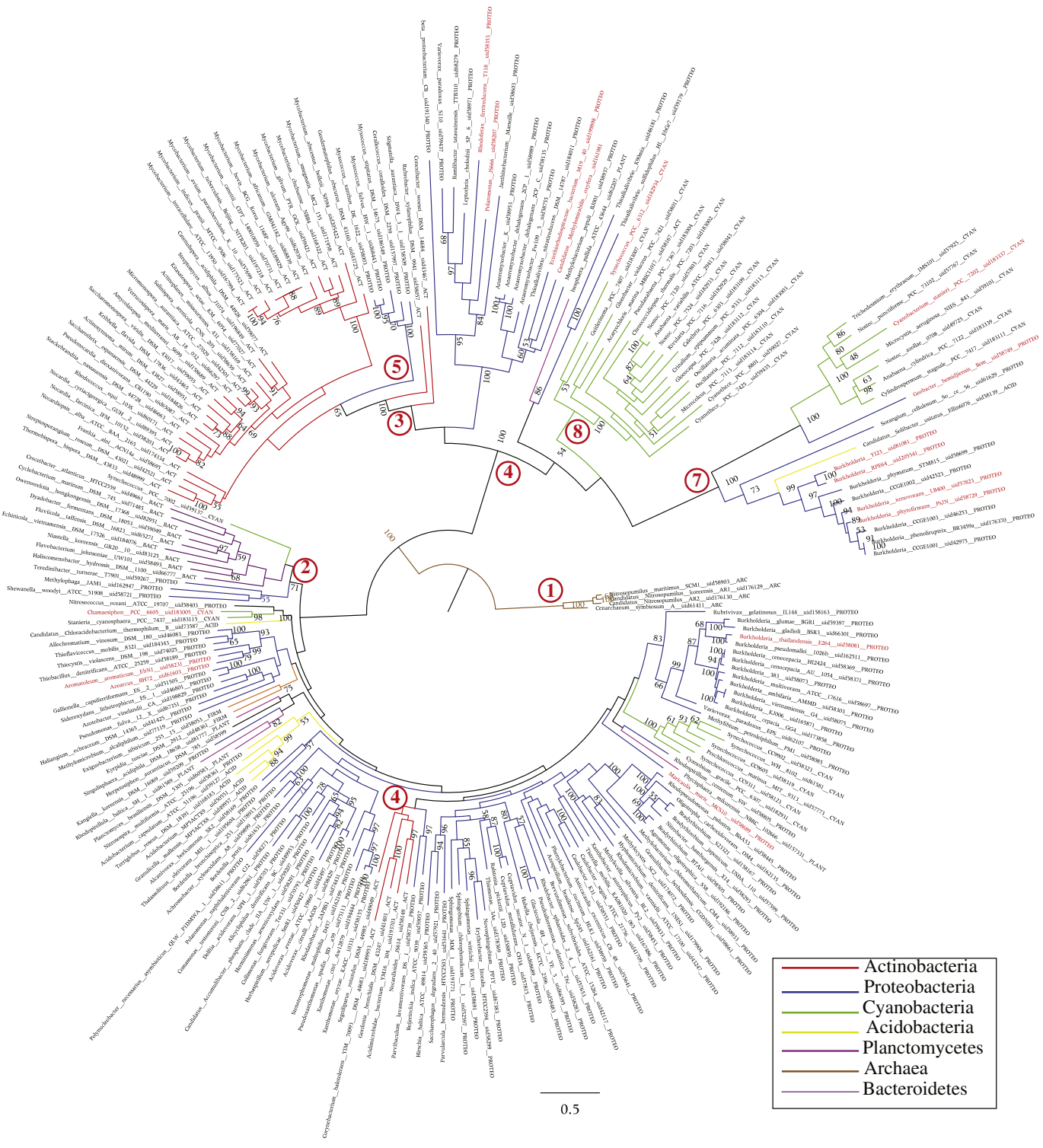


Fig. 3. EgtB maximum likelihood phylogeny. Tree is rooted around Archaeal species. Bootstrap scores greater than 50% are displayed for selected branches. Red numbers in red circles relate to specific clades discussed in the main text. Branches are coloured based on the phylum the species belong to. Species name highlighted in red show no genetic linkage between *egt*B and *egt*D.

Corallococcus coralloides and *Stigmatella aurantiaca*) obtained orthologs of *egt*B & *egt*D via HGT from an Actinobacterial source. All five species form a monophyletic clade with the predominant Actinobacterial clade (Figs. 3 & 4, Clade-5) with strong bootstrap support (100% & 81% respectively). As with the majority of species that contain copies of *egt*B & *egt*D, both genes are genetically linked (Additional file 1-B). The remaining 6 δ-proteobacterial species that contain orthologs of *egt*B & *egt*D are found grouped with other Proteobacterial species (Figs. 3 & 4).

Examining the phylogenetic relationships between the Cyanobacterial species in our dataset we fail to locate a single monophyletic Cyanobacterial clade illustrating the complex evolutionary history of *egt*B and *egt*D. For *egt*B, we see that the majority of Cyanobacterial

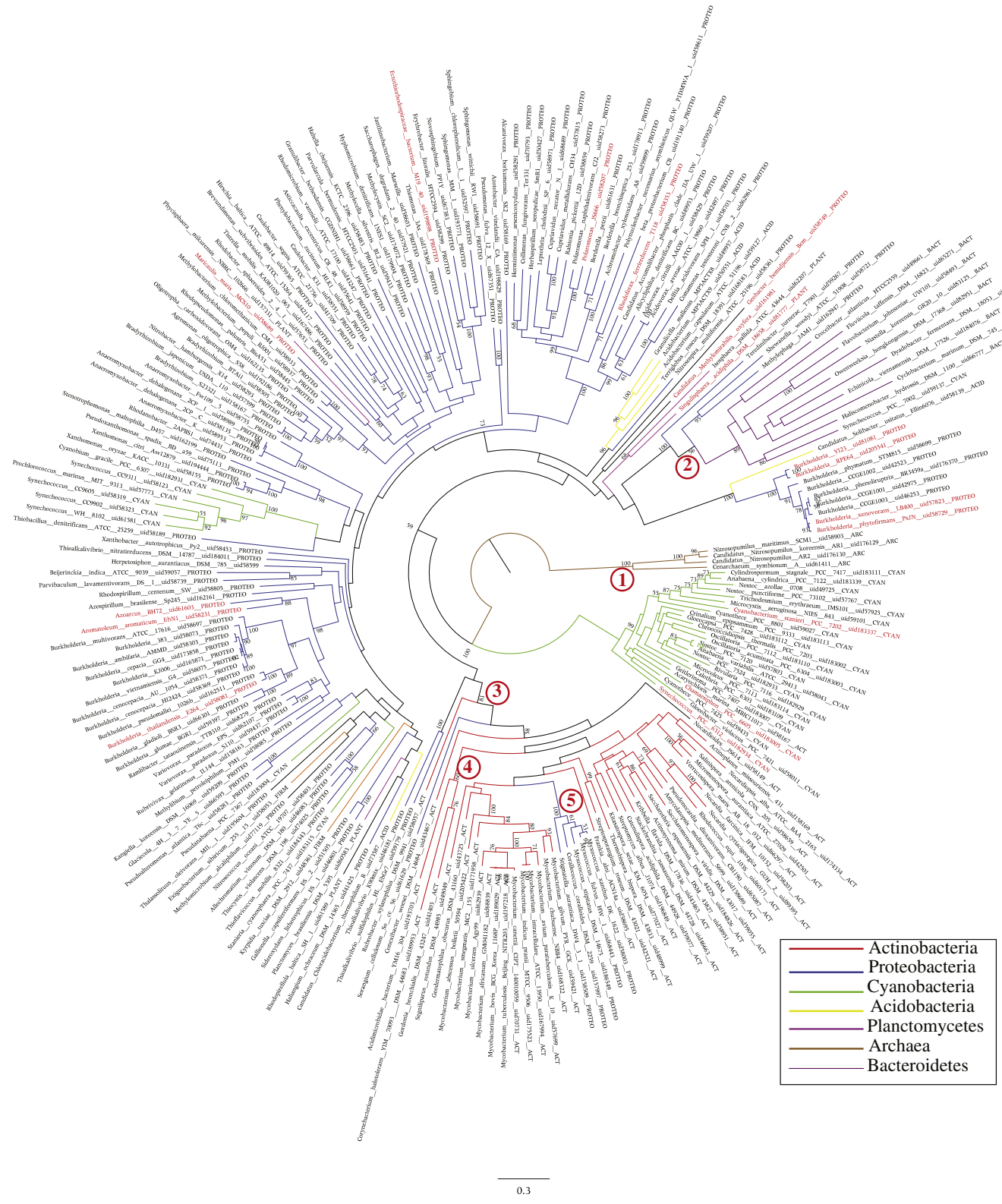


Fig. 4. EgtD maximum likelihood phylogeny. Tree is rooted around Archaeal species. Bootstrap scores greater than 50% are displayed for selected branches. Red numbers in red circles relate to specific clades discussed in the main text. Branches are coloured based on the phylum the species belong to. Species name highlighted in red show no genetic linkage between *egt*B and *egt*D.

species are located in a clade (54% BP) along with Actinobacterial and Proteobacterial species exclusively (with the exception of one Planctomycetes & one Acidobacteria species) (Fig. 3, Clade-6). These relationships are not observed in our EgtD phylogeny (Fig. 4). With respect to EgtB, there is a strongly supported (100% BP) Cyanobacterial/Proteobacterial clade (Fig. 3, Clade-7); these form a weak sister group relationship with a monophyletic Cyanobacterial clade (Fig. 3, Clade-8). This would suggest that these species have obtained a Cyanobacterial ortholog

of EgtB via HGT. The majority of Proteobacterial species in Clade-8 (Fig. 3) are from the *Burkholderia* genus. Interestingly four of these species (*Burkholderia* YI23, *Burkholderia* RPE64, *Burkholderia xenovorans* LB400 and *Burkholderia phytofirmans* PsJN) do not display linkage of *egt*B and *egt*D (Additional file 1-B and Figs. 3 and 4). This would indicate a recent HGT event into the ancestor of these *Burkholderia* species and absence of genome reassortment to link *egt*B and *egt*D since speciation.

There are numerous possibilities to the origin of *egt*B and *egt*D. Based on its patchy phyletic distribution within the prokaryotes it is parsimonious to assume that they were not ubiquitous characters that were lost in a species-specific manner through evolutionary time. Furthermore, based on the genomic data currently available, the most parsimonious scenario is that they arose within the Actinobacteria and were transferred into the ancestors of other phyla such as the Cyanobacteria and Proteobacteria before being subsequently transferred into other phyla in a species by species manner.

3.7. Evidence of fusion of egtB and egtD in selected bacterial species

Genes primarily evolve via the mechanisms of mutation, duplication and recombination. Recombination can generate proteins with novel domain architecture. Fusion or fission of genes is a type of recombination that generates either a composite protein or two (or more) smaller split proteins (Kummerfeld and Teichmann, 2005). Gene fission involves the gain of regulatory regions whereas gene fusion involves the loss of the terminal regions of one gene and initial regulatory regions of another gene. Genome analyses have shown a predominance of fusions relative to fissions (Kummerfeld and Teichmann, 2005; Snel et al., 2000). This is unsurprising as fusion is simpler to achieve genetically (Kummerfeld and Teichmann, 2005; Stechmann and Cavalier-Smith, 2002). There is also an inherent benefit to fusion as it allows for the physical amalgamation of functions that are biologically coupled (Marcotte et al., 1999). For example it has been shown that the vast majority of bacterial fusion genes are either part of the same complex or function in the same pathway (Marcotte et al., 1999; Snel et al., 2000).

Our analysis showed that there is a strong selective pressure for the genetic clustering of *egt*B or *egt*D (Additional file 1-B and Fig. 2). Furthermore, our analysis of prokaryote genomes identified two independent incidences where fusion of these genes has occurred. Five representatives (*Asticcacaulis excentricus*, *Brevundimonas subvibrioides*, *Caulobacter* K31, *Caulobacter segnis* ATCC 21756, *Phenylobacterium zucineum* HLK1) of the α-Proteobacteria order Caulobacterales have a protein that is the product of a fusion event between *egt*B and *egt*D. The resultant proteins have an N-terminal region with a DinB_2 and an FGE-sulfatase (homologous to *egt*B) and a C-terminal SAM-dependent methyltransferase (homologous to *egt*D). There are two other Caulobacterales species in our original bacterial dataset (*Caulobacterales crescentus* CB15 and *C. crescentus* CB15N, Additional file 1-A). Both contain an ortholog of *egt*B, but do not have a copy of *egt*D. A second fusion event has occurred in a single Cyanobacterial species, Cyanothece PCC 7425 (Additional file 1-B). The domain architecture is identical to that seen in the Caulobacterales species.

3.8. Phyletic distribution of egt genes in the fungal kingdom

M. smegmatis EGT proteins were searched against 103 individual fungal genomes across the fungal kingdom (Additional file 1-C & Fig. 5). No significant BlastP hits were recovered for *M. smegmatis* EgtA, EgtC or EgtE. A previous analysis found that *N. crassa* Egt-1 (NcEgt-1) contains domains found in both *M. smegmatis* EgtB and EgtD (Bello et al., 2012). For completeness the NcEGT-1 protein was searched against our fungal database. Best BlastP search hits were deemed orthologs if they had an E-value less than 10^{-5}, had NcEGT-1 as their top bidirectional search hit when searched back against the *N. crassa* genome and had a HSP that spanned 70% of the original NcEGT-1 protein.

Overall we identified 73 fungal Egt-1 orthologs (Additional file 1-C & Fig. 5). The average sequence pairwise identity between all orthologs is approximately 47%. Interestingly, all Saccharomycotina species investigated are missing Egt-1 (Fig. 5). This infers that the last common ancestor of all Saccharomycotina species in our analysis lost Egt-1 before speciation occurred. Egt-1 orthologs were found for all members of the Pezizomycotina. These findings concur with and expand those reported by Bello et al. (2012). Orthologs could not be identified for the Basidiomycete species *Moniliophthora perniciosa* or *Malassezia globosa*. The remaining 17 Basidiomycetes have a copy of Egt-1. Due to their phylogenetic distance (Fig. 5), this would suggest that both *M. perniciosa* or *M. globosa* have lost Egt-1 independently.

Previous in silico analyses have shown that NcEgt-1 is a multidomain protein, with an N-terminal SAM-dependent methyltransferase (27% identity to *M. smegmatis* EgtD) and a C-terminal region with a DinB_2 and a FGE-sulfatase (24% identity to *M. smegmatis* EgtB) (Seebeck, 2010). All 77 fungal Egt-1 orthologs we identified contain these three domains, with the same domain architecture. Using fungal copies of Egt-1 and bacterial copies of egtB and egtD we reconstructed phylogenetic trees (not shown) to determine if any recent interkingdom HGT has occurred. However, in both phylogenetic trees bacterial species and fungal species were also found in monophyletic clades indicating no recent HGT. Based on these inferences it is impossible to unambiguously determine the origin of EGT biosynthesis.

4. Conclusions

We investigated the phyletic distribution of a characterized 5-gene EGT cluster in Bacterial and Archaeal species. Overall we found that the gene cluster is specific for the Actinobacteria. Our analysis found that two of the genes, *egt*B and *egt*D, are found in a number of diverse phyla. Furthermore in the vast majority of cases both of these genes are genetically linked as they are found in close proximity to one another. Previous biochemical analyses have shown that Cyanobacterial species produce high levels of EGT even though they are missing orthologs of *egt*A, *egt*C and *egt*E (Pfeiffer et al., 2011). This indicates that EgtB and EgtD are key enzymes in the production of EGT and the enzymatic functions of EgtA, EgtC and EgtE may be performed by other as of yet unknown enzymes. Interestingly when we examined the expression data of *M. tuberculosis* H37Rv we observed that *egt*B and *egt*D are positively correlated in their expression levels, conversely there is no evidence to suggest that *egt*A, *egt*C or *egt*E is expressed in tandem with *egt*B and *egt*D lending further support to the hypothesis that other enzymes may be complementing the enzymatic steps of EgtA, EgtB and EgtC in the synthesis of EGT. We reconstructed phylogenetic trees in order to elucidate the evolutionary history of the five-*egt* genes. Due to the patchy phyletic distribution of *egt*A, *egt*C and *egt*E it is impossible to definitively infer their origin. *egt*A is nearly exclusively found in the Actinobacteria and Proteobacteria. *egt*C is only found in the Actinobacteria, Proteobacteria and Cyanobacteria while *egt*E is Actinobacteria specific. *egt*B and *egt*D are found across a number of diverse bacterial phyla but predominantly in Actinobacterial, Proteobacterial and Cyanobacterial species. Assuming that the 5-gene *egt* cluster contains the original enzymes necessary to synthesize EGT it follows that it has an Actinobacterial origin. Alternatively the 5-gene cluster may have evolved independently in the Actinobacteria after HGT of *egt*B and *egt*D from a Cyanobacterial or Proteobacterial source. There also may be the issue of a sampling bias in our dataset. Actinobacterial and Proteobacterial species account for approximately 53% of the species in out bacterial database. It is plausible that *egt*B and *egt*D may have arisen in a sparsely sampled bacterial phylum. However based on the data available to us we feel that the most parsimonious explanation is an Actinobacterial origin for the 5-gene EGT cluster (and *egtB* and *egtD* specifically) followed by HGT into ancestral Proteobacterial and Cyanobacterial species as well as multiple independent HGT events into individual species.

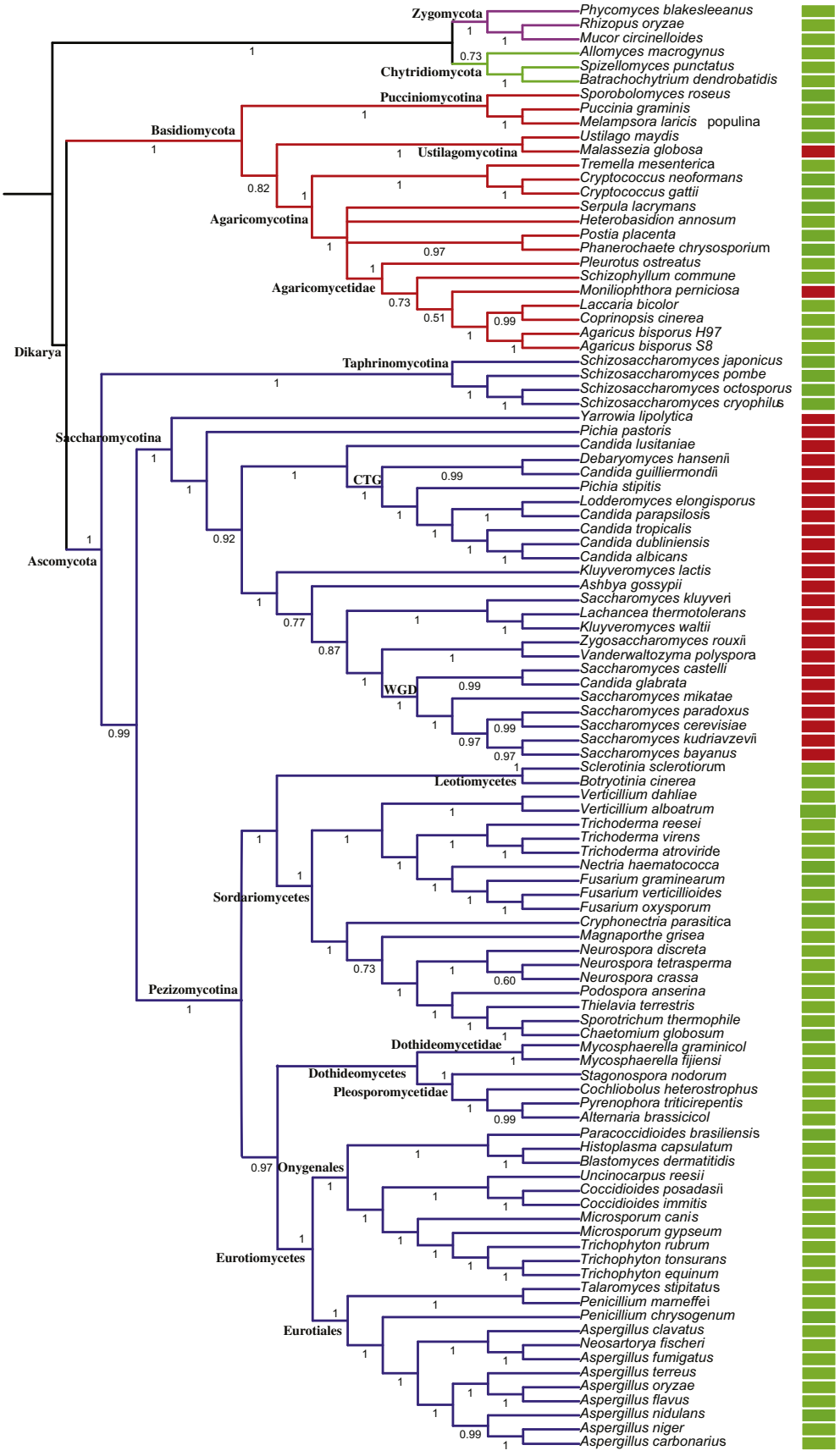


Fig. 5. Phyletic distribution of Egt-1 in the Fungal kingdom. Green rectangles indicate the presence of an Egt-1 ortholog while red rectangles indicate that Egt-1 is missing. All Saccharomycotina species are missing Egt-1. Accession numbers for orthologs are presented in Additional file 1-C.
Phylogeny modified from Medina et al. (2011).

EGT is also produced by fungi and recently Egt-1 has been shown to be the key enzyme in its synthesis (Bello et al., 2012). Egt-1 has domains homologous to bacterial EgtB and EgtD and is the likely result of an ancient gene fusion event. We surveyed over 100 fungal genomes and found that Egt-1 is found across the fungal kingdom with the notable exception of all Saccharomycotina species. This indicates that Egt-1 was in the last common ancestor of all fungal species examined but was lost in the ancestral Saccharomycotina species. Database BlastP searches and phylogenetic trees did not show any recent interdomain HGT between fungal or bacterial species for either Egt-1 or *egt*B/*egt*D. Therefore, it is not possible to confidently infer the origin of EGT biosynthesis. There are three possible scenarios. Firstly it is may be a fungal innovation that was transferred into a bacterial phyla (the Actinobacteria for example), fission of Egt-1 occurred to give *egt*B and *egt*D and these were subsequently spread in prokaryotes via HGT. Alternatively, it may be a bacterial innovation, *egt*B and *egt*D may have been transferred into the last common fungal ancestor followed by a fusion event to give Egt-1. The final scenario is that both bacteria and fungi have independently evolved mechanisms to synthesize EGT. However based on sequence similarity and the relatedness of pathways involved (Fig. 1) we feel that this scenario is the least likely of the three.

Our analysis provides a comprehensive dissection of the evolutionary history and distribution of the genes involved in EGT production in the tree of life. We have shown, based on the presence of *egt*B/*egt*D or Egt-1 that a diverse range of bacteria and fungi can potentially synthesize EGT.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors were involved in the design phase. DAF performed all in silico analysis and drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was in part funded by a Science Foundation Ireland Principal Investigator Award (PI/11/1188) and FIRM grant 13/F/463 from the Irish Department of Agriculture Food and the Marine. We thank Mr. Stephen Dolan from the Biotechnology Laboratory, National University of Ireland, Maynooth, Ireland, for generating Fig. 1.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gene.2014.07.065.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. (Online) 25, 3389–3402.

Bello, M.H., Barrera-Perez, V., Morin, D., Epstein, L., 2012. The *Neurospora crassa* mutant NcDeltaEgt-1 identifies an ergothioneine biosynthetic gene and demonstrates that ergothioneine enhances conidial survival and protects against peroxide toxicity during conidial germination. Fungal Genet. Biol. 49, 160–172.

Cheah, I.K., Halliwell, B., 2012. Ergothioneine; antioxidant potential, physiological function and role in disease. Biochim. Biophys. Acta 1822, 784–793.

Darriba, D., Taboada, G.L., Doallo, R., Posada, D., 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27, 1164–1165.

Dubost, N.J., Ou, B., Beelman, R.B., 2007. Quantification of polyphenols and ergothioneine in cultivated mushrooms and correlation to total antioxidant capacity. Food Chem. 105, 727–735.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.

Fisher, S.A., Hampe, J., Onnie, C.M., Daly, M.J., Curley, C., Purcell, S., Sanderson, J., Mansfield, J., Annese, V., Forbes, A., Lewis, C.M., Schreiber, S., Rioux, J.D., Mathew, C.G., 2006. Direct or indirect association in a complex disease: the role of SLC22A4 and SLC22A5 functional variants in Crohn disease. Hum. Mutat. 27, 778–785.

Galagan, J.E., Sisk, P., Stolte, C., Weiner, B., Koehrsen, M., Wymore, F., Reddy, T.B., Zucker, J.D., Engels, R., Gellesch, M., Hubble, J., Jin, H., Larson, L., Mao, M., Nitzberg, M., White, J., Zachariah, Z.K., Sherlock, G., Ball, C.A., Schoolnik, G.K., 2010. TB database 2010: overview and update. Tuberculosis (Edinb) 90, 225–235.

Genghof, D.S., 1970. Biosynthesis of ergothioneine and hercynine by fungi and Actinomycetales. J. Bacteriol. 103, 475–478.

Genghof, D.S., Vandamme, O., 1964. Biosynthesis of ergothioneine and hercynine by mycobacteria. J. Bacteriol. 87, 852–862.

Genghof, D.S., Inamine, E., Kovalenko, V., Melville, D.B., 1956. Ergothioneine in microorganisms. J. Biol. Chem. 223, 9–17.

Griffin, J.E., Gawronski, J.D., Dejesus, M.A., Ioerger, T.R., Akerley, B.J., Sassetti, C.M., 2011. High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism. PLoS Pathog. 7, e1002251.

Grundemann, D., Harlfinger, S., Golz, S., Geerts, A., Lazar, A., Berkels, R., Jung, N., Rubbert, A., Schomig, E., 2005. Discovery of the ergothioneine transporter. Proc. Natl. Acad. Sci. U. S. A. 102, 5256–5261.

Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52, 696–704.

Hartman, P.E., 1990. Ergothioneine as antioxidant. Methods Enzymol. 186, 310–318.

Kummerfeld, S.K., Teichmann, S.A., 2005. Relative rates of gene fusion and fission in multi-domain proteins. Trends Genet. 21, 25–30.

Leung, E., Hong, J., Fraser, A.G., Merriman, T.R., Vishnu, P., Krissansen, G.W., 2006. Polymorphisms in the organic cation transporter genes SLC22A4 and SLC22A5 and Crohn's disease in a New Zealand Caucasian cohort. Immunol. Cell Biol. 84, 233–236.

Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O., Eisenberg, D., 1999. Detecting protein function and protein–protein interactions from genome sequences. Science 285, 751–753.

Martin, F.J., McInerney, J.O., 2009. Recurring cluster and operon assembly for Phenylacetate degradation genes. BMC Evol. Biol. 9, 36.

Medina, E.M., Jones, G.W., Fitzpatrick, D.A., 2011. Reconstructing the fungal tree of life using phylogenomics and a preliminary investigation of the distribution of yeast prion-like proteins in the fungal kingdom. J. Mol. Evol. 73, 116–133.

Melville, D.B., Eich, S., Ludwig, M.L., 1957. The biosynthesis of ergothioneine. J. Biol. Chem. 224, 871–877.

Newton, G.L., Bewley, C.A., Dwyer, T.J., Horn, R., Aharonowitz, Y., Cohen, G., Davies, J., Faulkner, D.J., Fahey, R.C., 1995. The structure of U17 isolated from *Streptomyces clavuligerus* and its properties as an antioxidant thiol. Eur. J. Biochem. 230, 821–825.

Newton, G.L., Arnold, K., Price, M.S., Sherrill, C., Delcardayre, S.B., Aharonowitz, Y., Cohen, G., Davies, J., Fahey, R.C., Davis, C., 1996. Distribution of thiols in microorganisms: mycothiol is a major thiol in most actinomycetes. J. Bacteriol. 178, 1990–1995.

Park, E.J., Lee, W.Y., Kim, S.K., Ahn, J.K., Bae, E.K., 2010. Ergothioneine accumulation in a medicinal plant *Gastrodia elata*. J. Med. Plant Res. 4, 1141–1147.

Peltekova, V.D., Wintle, R.F., Rubin, L.A., Amos, C.I., Huang, Q., Gu, X., Newman, B., Van Oene, M., Cescon, D., Greenberg, G., Griffiths, A.M., St George-Hyslop, P.H., Siminovitch, K.A., 2004. Functional variants of OCTN cation transporter genes are associated with Crohn disease. Nat. Genet. 36, 471–475.

Pfeiffer, C., Bauer, T., Surek, B., Schomig, E., Grundemann, D., 2011. Cyanobacteria produce high levels of ergothioneine. Food Chem. 129, 1766–1769.

Rahman, I., Gilmour, P.S., Jimenez, L.A., Biswas, S.K., Antonicelli, F., Aruoma, O.I., 2003. Ergothioneine inhibits oxidative stress- and TNF-alpha-induced NF-kappa B activation and interleukin-8 release in alveolar epithelial cells. Biochem. Biophys. Res. Commun. 302, 860–864.

Santiago, J.L., Martinez, A., de la Calle, H., Fernandez-Arquero, M., Figueredo, M.A., de la Concha, E.G., Urcelay, E., 2006. Evidence for the association of the SLC22A4 and SLC22A5 genes with type 1 diabetes: a case control study. BMC Med. Genet. 7, 54.

Sao Emani, C., Williams, M.J., Wiid, I.J., Hiten, N.F., Viljoen, A.J., Pietersen, R.D., van Helden, P. D., Baker, B., 2013. Ergothioneine is a secreted antioxidant in *Mycobacterium smegmatis*. Antimicrob. Agents Chemother. 57, 3202–3207.

Sassetti, C.M., Boyd, D.H., Rubin, E.J., 2003. Genes required for mycobacterial growth defined by high density mutagenesis. Mol. Microbiol. 48, 77–84.

Seebeck, F.P., 2010. In vitro reconstitution of Mycobacterial ergothioneine biosynthesis. J. Am. Chem. Soc. 132, 6632–6633.

Snel, B., Bork, P., Huynen, M., 2000. Genome evolution. Gene fusion versus gene fission. Trends Genet. 16, 9–11.

Stechmann, A., Cavalier-Smith, T., 2002. Rooting the eukaryote tree by using a derived gene fusion. Science 297, 89–91.

Ta, P., Buchmeier, N., Newton, G.L., Rawat, M., Fahey, R.C., 2011. Organic hydroperoxide resistance protein and ergothioneine compensate for loss of mycothiol in *Mycobacterium smegmatis* mutants. J. Bacteriol. 193, 1981–1990.

Tanret, M.C., 1909. Sur une base nouvelle retiree du siegle ergote, l' ergothioneine. C. R. 49, 22–224.

Vilcheze, C., Av-Gay, Y., Attarian, R., Liu, Z., Hazbon, M.H., Colangeli, R., Chen, B., Liu, W., Alland, D., Sacchettini, J.C., Jacobs, W.R., 2008. Mycothiol biosynthesis is essential for ethionamide susceptibility in *Mycobacterium tuberculosis*. Mol. Microbiol. 69, 1316–1329.

Zhu, B.Z., Mao, L., Frei, B., 2010. Ergothioneine prevents copper-induced oxidative damage to DNA and protein by formation of a redox-inactive ergothioneine–copper complex. Free Radic. Biol. Med. 49, S205-S205.