Evolution and Bioinformatics Analysis of the Drug: H+ Antiporter Family 1 (DHA1) in the hemiascomycetes yeasts

André Miguel Moreira Machado

Under supervision of Prof. Dr. Isabel Maria de Sá Correia Leite de Almeida

and supervision of Dr. Paulo Jorge Moura Pinto da Costa Dias

IST, Lisbon, Portugal

July, 2016

Extended Abstract

The *Saccharomyces cerevisiae* 12-spanner drug: H+ antiporters (DHA1) and 14- spanner drug: H+ antiporters (DHA2) of the Major Facilitator Superfamily (MFS) are involved the in Multidrug/Multixenobiotic resistance (MDR/MXR) phenomenon. The aim of the present work is to reconstruct and characterize the evolution of the DHA1 genes encoded in 33 hemiascomycetes strains classified in the Saccharomycetaceae taxonomic family, (corresponding to a total of 29 yeast species). The DHA1 and DAG (DHA2, ARN, GEX) proteins encoded in the genomes of 61 additional strains, spanning more than 15 hemiascomycetous taxonomic families, were also identified and briefly analysed. The constraining and traversing of a network representing the blastp pairwise similarity relationships established between more than one million hemiascomycetous translated ORFs allowed the identification of 1382 bona fide full-size DHA1 transporters (after correction of problematic translated ORFs). The evolutionary history of the DHA1 genes encoded in the genome sequences of 33 Saccharomycetaceae strains was reconstructed by combining phylogenetic and gene neighbourhood approaches. Twenty-six DHA1 phylogenetic clusters were identified and nine DHA1 gene lineages reported in previously published studies were revised and extended. State-of-the-art methodologies on phylogeny, comparative genomics and protein evolution were used to advance the existing knowledge on the still poorly biochemically characterized DHA1 transporters, allowing obtaining new insights on how the functional diversification of these proteins is related to ancestral genomic events, such as the Whole Genome Duplication (WGD), local gene duplications and losses, Horizontal Gene Transfers (HGT) between yeast species, chromosomal rearrangements, and other genome reshaping phenomena.

Keywords: DHA1, DAG, *Saccharomyces cerevisiae*, Multidrug resistance, Saccharomycetaceae, Whole Genome Duplication.

1. Introduction

The yeasts are one of the most important microorganisms in biotechnology. These microorganisms are used in a wide diversity of applications, from industrial fermentations to environmental and agricultural research. The wide application of these yeasts has been a driving force to the comprehension of their characteristics and survival mechanisms. Also, detailed comprehension of these mechanisms has given us knowledge about the resistance of these microorganisms to the surrounding environment, which is related to the capacity to process or expel a wide range of structurally and functionally unrelated cytotoxic chemicals, a phenomenon known as Multidrug / Multixenobiotic resistance (MDR/MXR) (dos Santos *et al.* 2014; dos Santos and Sá-Correia 2015). In the Hemiascomycetes, two superfamilies of transporters are known to be involved in the Multidrug Resistance phenomenon: the ATP-Binding-Cassette (ABC-PDR) and the Major Facilitator Superfamily (MFS-MDR) efflux pumps (Sá-Correia and Tenreiro 2002). The MFS-MDR transporters are energized by an electrochemical proton-motive force established between the cell cytoplasm and the external medium. After the identification of 28 MFS-MDR genes encoded in the genome of *Saccharomyces cerevisiae*, it was proposed dividing the corresponding proteins into two families, the 12-spanner Drug: H+ Antiporter family 1 (DHA1), with 12 members, and the 14-spanner Drug: H+ Antiporter family 1 (DHA2), with 16 members (Paulsen *et al.* 1998; Sá-Correia and Tenreiro 2002; Sá-Correia *et al.* 2009). The members of the DHA1 family encoded in the S. cerevisiae genome are the *AQR1, DTR1, QDR1, QDR2, QDR3, TPO1, TPO2, TPO3, TPO4, HOL1, FLR1* and *YHK8* genes. The members of the DHA1 family are described being involved in a range of different physiological functions, such as polyamine transport, spore wall biosynthesis and stress resistance (Sá-Correia *et al.* 2009). More recently, it has been shown that the DHA2 family should be renamed and divided into three sub-families (Dias and Sá-Correia 2013): the DHA2 subfamily, with 10 members, corresponding to the ATR1, AZR1, SGE1, VBA1, VBA2, VBA3, VBA4, VBA5 and AMF1 genes and the still uncharacterized translated ORF YMR279c, the ARN subfamily, with 4 members, corresponding to the ARN1, ARN2, SIT1, and ENB1 genes, and the GEX subfamily, with 2 members, corresponding to the GEX1 and GEX2 genes. The majority of the members of the DHA2 subfamily do not have assigned function. On the other hand, the members of the GEX and ARN subfamilies encode glutathione; H+ antiporters and siderophore-iron chelate influx pumps, respectively (Lesuisse, Simon-casteras and Labbe 1998; Sá-Correia *et al.* 2009; Dhaoui *et al.* 2011). The identification and in silico characterization of the DHA1 and DAG genes encoded in the Hemiascomycetes is an important goal since the yeast MFS-MDR genes have shown promising applications in a different variety of fields in Biotechnology.

The present work focuses the reconstruction of the evolution of the DHA1 family members in the sub-phylum Saccharomycotina, aiming to extend the results reported in previously published studies to 94 additional yeast strains whose genome sequences were recently determined. The identification of the DAG genes encoded in these yeast strains will be also performed in this work and a brief characterization in a subset of 74 hemiascomycetes strains will be attempted. The evolution of the DHA1 genes encoded in the genomes of yeast species belonging to the Saccharomycetaceae taxonomic family will be scrutinized in detail, with this goal being pursued based on complete bioinformatics and comparative genomic approaches developed by the BSRG group in the last 5 years.

2. Materials and Methods

2.1. Selection of the Hemiascomycetes yeasts used in this study.

Based on published literature (Dias *et al.* 2010; Dias and Sá-Correia 2014) and using a local genome database (LDB) created in the MySQL relational database environment, 94 hemiascomycetous yeast strains were selected. The ninety-four pre-selected yeast strains correspond to sixty-four species that are phylogenetically subdivided into 53 strains belonging to the *Saccharomyces* complex; 20 strains of the CTG complex; 6 strains classified as are early divergent hemiascomycetous strains and other 15 species belonging other taxonomic families.

2.2. Extraction of DHA1 and DAG gene sequences from a Local Database

The identification of members of DHA1 and DAG families was based in silico methodologies already develop and published studies by BSRG group (Dias *et al.* 2010; Dias and Sá-Correia 2013, 2014) . In the present study twenty-

three DHA1 genes phylogenetically classified into different clusters (Dias and Sá-Correia 2014) were used as starting nodes in a blastp network traversal approach for the identification of DHA1 family members. Regarding the DAG family, also twenty genes were selected from the twenty phylogenetically classified into different clusters (Dias and Sá-Correia 2013) and were used as starting nodes in a similar traversal approach.

2.3. Sequence clustering of all DHA1 and DAG Proteins.

Using the local genome and blastp database and R scripting already developed, a blast all-against-all pairwise comparison between 43 reference genes and LDB was performed. For this purpose, the blastp algorithm made available in blast2 package (Tatusova and Madden 1999) was used. The traversing of this blastp network at different e-values made possible the gathering of all proteins belong to the DHA1 and Dag protein families.

2.4. Pre-analysis of the DHA1 and DAG genes.

After the identification of the potential DHA1 and DAG genes encoded in the analysed genomes, a number of approaches were used to confirm the obtained results. To dismiss the presence of false positive in the gathered DHA1 and DAG gene set, the length of sequences, topology, phylogeny and protein alignment were verified. In early validation stage, HMMTOP, TMHMM 2.0 and TOPPRED and TOPPRED 2 were used for topology, Muscle for multiple alignment, JALVIEW 2.0 for alignment visualization, PROTDIST/ NEIGHBOUR functions of the PHYLIP package for phylogeny data processing , and DENDROSCOPE tool for the visualization of phylogenetic tree (Plotree and Plotgram 1989; Von Heijne 1992; Tusnady and Simon 2001; Edgar 2004; Huson *et al.* 2007; Bernsel *et al.* 2009; Waterhouse *et al.* 2009). The genes comprising the DHA1 and DAG families were scrutinized with different depth.

In the case of DAG family, a brief analysis was performed. The number of transmembrane segments and the length of the amino acid sequence was used as an initial criterion to aim deciding whether the potential members of this protein family was a bona fide DAG transporter and if it comprised a full size protein or a fragment. Importantly, to consider the proteins in the study were established that all amino acid sequences belonging to the DAG family, should contain, at least 14 transmembrane segments in topology and 420 residues in length. The proteins that initially had these pre-requirements were used for the phylogenetic tree construction of the representing the DAG proteins. In contrast with the previous analysis, to consider the DHA1 proteins in the study were established that all amino acid sequences belonging to this family, should contain, at least 12 transmembrane segments in topology and 431 residues in length. Sequence errors deriving from genome assembly annotation as well as a loss of DHA1 genes led us to attempt the corrections of fragments and frameshifts. The amino acid length, topology and the place of residence of the DHA1 genes in the phylogenetic tree was used to detect and select amino acid sequences for error corrections.

2.5. Final analysis of DHA1 and DAG genes.

Once the pre-analysis was completed, a final validation was necessary. In case of 2D-topology prediction, additional software were used to complement the initial analyses. These software include POLYPHOBIUS, OCTOPUS, PHYLIUS, SCAMPI, SPOCTOPUS, and still the web server TOPCONS (Käll, Krogh and Sonnhammer 2007; Bernsel *et al.* 2008, 2009; Viklund and Elofsson 2008; Viklund *et al.* 2008).The Muscle and Jalview software were used to perform the multiple alignments and to confirm the occurrence of highly hydrophobicity portions in the protein sequences (Edgar 2004; Waterhouse *et al.* 2009).The hydrophobicity associated with each TMS predicted to occur in protein sequences under analyses was verified using the software TOPPRED II(Claros, M.G.; Von Heijne 1994) . Complementing the Phylip Package-PROTDIST/NEIGHBOUR algorithms , two other methodologies of phylogenetic tree construction were used (Plotree and Plotgram 1989). The PHYML 3.0 software (Guindon *et al.* 2010) based on the maximum-likelihood approach, together with approximate likelihood-ratio test (aLRT), and more classical non-parametric bootstrapping method for trees validation were used (Felsenstein 1985; Anisimova and Gascuel 2006). Mr Bayes 3.2, a bayesian methodology for statistical inference was also used for tree validation based on the calculation of the probability of node position (Ronquist *et al.* 2012).

2.6. Construction and analyses of syntenic block diagrams representing genome regions of 33 Hemiascomycetous strains.

The analysis of the chromosome environment where the *S. cerevisiae* DHA1 members reside was based on the fifteen neighbouring genes immediately upstream and downstream in the genome sequence. Neighbour genes were retrieved from the MySQL database using the package 'sqldf' and complementary scripting in R language (Rice, Longden and Bleasby 2000). E-value thresholds of E- 50 and E-40 were used to define clusters of protein similarity. The synteny output was visualized and analysed in the Cytoscape software (Lotia *et al.* 2013).

2.7. Identification of positive and purifying extended over the amino acid sequences of the FLR1p homologs proteins during the evolution of Saccharomycetaceae yeasts.

A protein alignment sequence conservation score and positive and purifying selection methods were used to improve our understanding of evolutionary forces acting on the FLR1 homologs genes encoded in the genome of yeasts species belonging to the Saccharomycetaceae family. Three distinct methodologies were used – a " Fast, Unconstrained Bayesian AppRoximation" (FUBAR) (Murrell *et al.* 2013),a "mixed effects model of evolution" (MEME) (Murrell *et al.* 2012) and a "fixed effects likelihood" (FEL) (Kosakovsky Pond *et al.* 2005) implemented in Hyphy 2.2 Package for selection and BILD ''Bayesian Integral Log-odds'' for protein sequence conservation analyses (Altschul *et al.* 2010).

3. Results

3.1. Identification of the DHA1 and DAG transporters encoded in the genomes of hemiascomycetes yeasts.

The constraining and traversing of a network representing the blastp pairwise similarity relationships established between more than one million hemiascomycetous translated ORFs allowed the identification of DHA1 and DAG subfamilies members encoded in 94 and 78 Hemiascomycetous strains, respectively. The twenty-three DHA1 and twenty DAG proteins identified in the previous published studies of BSRG group were used as starting node. The analysis of the plot representing the number of sequences retrieved at different e-values allowed choosing of E-48 to constrain the pairwise similarity network. In the final of this approach was possible collect 1466 DHA1 and 1706 DAG sequences to analyse in further analyses (Figure.1).

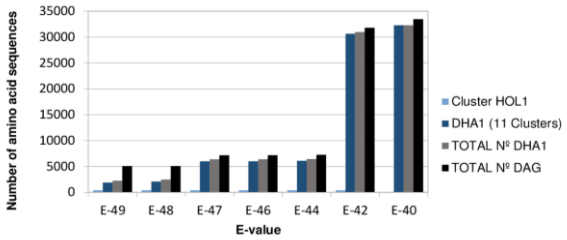

Figure. 1 Identification of the MFS-MDR DHA1 and DAG genes encoded in 94 yeast strains. Plot representing the number of sequences retrieved after constraining and traversing the pairwise similarity network at different e-values using 43 reference genes as starting nodes.

3.2. A brief analysis of the DAG transporters encoded in the genomes of 78 hemiascomycetes strains.

Although the DAG proteins are not the main focus of this study, the gathered amino acid sequences were used to construct a global phylogenetic tree representing the DAG transporters. However, in the analysis of these proteins, only 78 yeast genomes were considered, corresponding to 63 different hemiascomycetes species. The analysis of the topology prediction results, the aligned amino acid sequences and protein length allowed excluding 715 false positives from the initially identified 1706 potential DAG transporters. Of the remaining 991 bona fide DAG proteins, 818 were full-size 14-spanner transporters and 173 were fragments. Based on a published study of Dias et al. (2013) the full-size DAG transporters were classified into 20 phylogenetic clusters, labelled from A to T**.**

The total number of DAG proteins encoded in each hemiascomycetous genomes was determined and used to calculate the average number of these transporters encoded in the genomes of the distinct taxonomic families/groups (see section 2.1) considered in this study. The DHA2 subfamily comprises the major part of the DAG proteins and clusters with 564 of the 818 total number of full-size DAG transporters. The cluster B (SGE1/AZR1/ VBA3/VBA5) is the more populated with 145 members, while the G and A clusters have the smallest number of proteins with 3 and 1 members, respectively. Interestingly, a few members such as *S. stipites*, *L. thermotolerans* and *L. kluyvery* encoded a high number of clusters B members in their genome sequences. The GEX transporters comprise a total of thirteen members in all 78 yeasts

strains, being predominantly encoded in genomes of the *Saccharomyces sensu stricto* group. The ARN subfamily comprises a total of 241 members. The homologs of the *S. cerevisiae,* ARN1 and ARN2 genes are predominantly found in *Lachancea* and *Kluyveromyces species*. Our results data still allowed corroborate, some cases of lateral gene transference inside the Saccharomycetaceae family, in clusters, B, F, J, N, and S previously proposed in Dias et al. (2013).

3.3. Analysis of the DHA1 transporters in the Hemiascomycetes.

The main focus of this master's thesis is the characterization of the DHA1 transporters in the yeasts commonly known as the Hemiascomycetes (subphylum Saccharomycotina). This goal required an exhaustive and complete validation of the gathered 1466 translated ORF's identified using the blastp network traversal approach, allowing deciding whether these amino acid sequences corresponded to bona fine full-size DHA1 proteins or not. The validation of problematic translated ORFs was based on sequence analysis at the DNA level, on 2D-topology prediction software and on tblastn queries against the corresponding yeast genomes.

3.3.1. Correction of sequencing and annotation errors in the initial DHA1 protein dataset.

An initial analysis gathered 1466 amino acid sequences shows that, in average, the potential DHA1 transporters have a size of 574 amino acids, ranging from 431 to 800 amino acids in length. Besides the size of each protein, the analysis of these amino acid sequences using 2D-topology prediction software, the corrections of sequence errors and the inspection of the global multiple alignments of the gathered protein set confirmed that 1382 were bona fine DHA1 proteins.

3.3.2. Analysis of the 2D-topology of the DHA1 transporters.

In order to complement the pre-analyses of topology, and to predict a more accurate number of transmembrane segments (TMS) for the gathered DHA1 proteins, three additional software were used to complement the HMMTOP and TMHMM - PHOBIUS, SCAMPI and TOPPRED II (GES scale and KD Scale). The web server TOPCONS 2.0 was also used to test the prediction of TMS's based on five individual software - OCTOPUS, PHILIUS, POLYPHOBIUS, SPOCTOPUS, SCAMPI. This allowed the identification of 1284 DHA1 proteins with twelve transmembrane segments and 98 with eleven TMS's. Interestingly all of these 98 proteins show a high amino acid similarity to the *S. cerevisiae* HOL1 gene.

3.4. Global phylogenetic analysis of the DHA1 transporters encoded in the genomes of the hemiascomycetes yeast

With the goal of extending the published results regarding the evolution of the DHA1 gene family in the subphylum Saccharomycotina (Dias *et al.* 2010; Dias and Sá-Correia 2014), the 1382 DHA1 proteins were used to construct a phylogenetic tree representing the diversity of these transporters in 94 hemiascomycetes strains, corresponding to 63 different yeast species. The construction of these phylogenetic trees allowed the identification of 26 clusters.

3.4.1. Phylogenetic analysis of the DHA1 transporters encoded in species of the *Saccharomyces* complex.

When compared with the previously published literature on the DHA1 proteins, this master's thesis has an important advantage since the number of yeast species and strains analysed in the present study increased to 63 and 94, respectively. The strong increase in the number of yeast genomes considered in this master's thesis and, consequently, the number of DHA1 transporters, has the benefit of improving the power and resolution of the phylogenetic analysis of these proteins.

The phylogenetic analysis of the 654 DHA1 full-size proteins found encoded in the 53 yeast strains classified in the *Saccharomyces* complex in the present study recovered two additional clusters (K and D) not observed in the published 2010 and 2014 papers. Previous literature reported that the genome of the *S. cerevisiae* S288c reference strain encodes 12 DHA1 transporters. The analysis of 18 different *S. cerevisiae* strains, showed that the genome sequence of the *S. cerevisiae* RM11-1a, vin13 and Y12 strains had variations in the total DHA1 gene number. Strain RM11-1a, encodes only 11 full-size DHA1 proteins since the amino acid sequence of the HOL1 gene have accumulated stop codons, being considered a pseudogene in this work. Strain vin13 also encodes 11 DHA1 proteins, lacking the TPO2 gene. Although the S. cerevisiae Y12 strain (sace_37) encodes 12 DHA1 proteins, the phylogenetic analysis showed that the TPO1 gene is missing in its genome sequence and that an extra copy of the TPO4 gene has been acquired. The

phylogenetic analysis of phylogeny still allowed identifying a number of strong discrepancies regarding the DHA1 gene number of yeast species classified in the same taxonomic genus. The *Kazachstania* genus species show a total difference of 4 DHA1 proteins encoded in their genomes. The hemiascomycetes yeast classified in the *Tetrapisispora* genus show a total difference of 5 DHA1 proteins encoded in their genomes. Consistent with the high DHA1 gene number reported for *Zygosaccharomyces rouxii* in the 2010 paper (Dias *et al.* 2010), totalling 21 genes, the three strains of the related yeast species *Zygosaccharomyces bailii* (ISA1307, IST302, CLIB 213) considered in this master's thesis were found encoding ,39, 21 and 18 DHA1 proteins in their genome sequences.

3.4.2. Phylogenetic analysis of the DHA1 transporters encoded in species of the CTC complex.

The analysis of 17 yeast species residing in the CTG complex and corresponding to a total of 20 hemiascomycetes strains, allowed the identification of 447 DHA1 genes in this phylogenetic clade. The genomes of the yeast species classified in the CTG complex, when compared with those of the Saccharomyces complex, encode a higher DHA1 gene number, with an average of 0.86 genes per cluster. The analysis of DHA1 proteins showed the extension of DHA1 members in 19 different phylogenetic clusters.

3.4.3. Phylogenetic analysis of the DHA1 transporters encoded in species classified in other late-divergent taxonomic families in the Hemiascomycetes

In the other late-divergent taxonomic families, the *D. bruxellensis* and *O. parapolymorpha* strains revealed an intermediate number of DHA1 genes. The remaining species of this group have a total number of DHA 1 proteins that range from 7 to 22 proteins, with J cluster (homologs of *S. cerevisiae* HOL1 gene) being the cluster comprising the largest number of DHA1 proteins, with 37 members. On the other hand, the members of the phylogenetic clusters T and E encoded in these yeast species contain a lower number of members comparatively with others gene sets, nine and six. This suggests that these homologs of *S. cerevisiae* DTR1 and FLR1 genes are not essential, for the fluconazole or others multidrug resistance (Cluster T) or in spore wall synthesis (Cluster E).

3.4.4. Phylogenetic analysis of the DHA1 transporters encoded in species classified in early-divergent taxonomic families in the Hemiascomycetes

The early-divergent species are the most heterogeneous group regarding the total DHA1 gene members. Showing an average of 0.54 DHA1 genes per cluster, these yeast species have a high number of members, the case of *Y. lipolytica* with 32 members, a medium number, *C. caseinolytica* with 18 members, and with a low number, *Komagataella* genus with an average of 8 genes per strain.

3.5. Phylogenetic, gene neighbourhood and protein evolution analyses of the DHA1 transporters encoded in the genomes of Saccharomycetaceae yeast species.

The main focus of this master's thesis is the characterization of the DHA1 transporters. However, the analysis of all the 1382 bona fide full-size DHA1 proteins found encoded in 94 Hemiascomycetes strains (comprised in more than 15 taxonomic families) using the state-of-the-art methodologies would fall beyond the limited scope of a master's thesis. Thus, only 438 DHA1 proteins encoded in the genomes of 33 yeast strains, belong to the Saccharomycetaceae family, were used for a more detailed analyses with phylogenetic and syntenic approaches. Four hundred and nineteen full-size genes were used for the construction of phylogenetic tree and subsequent division of the tree into twenty clusters. The identity and similarity, between the proteins of each cluster, ranged from 47.2 to 87% and 62.6 to 92%, respectively. Although the DHA1 protein population density is maintained inside the clusters, the protein ratio between the clusters change. In the case of E, R and S clusters (DTR1, YHK8, and TPO4 representative genes), the DHA1 genes are uniformly spread by all species. Without *S. cerevisiae* representatives, the A+B, D, I, K1, K2, N2, O, Q and U+V clusters comprise a maximum and minimum of thirteen and two genes, being the lowest populated clusters. The syntenic approach used in the following section was based on the published work of Dias et al. (2010). During the syntenic assessment of the DHA1 genes, all 438 members (fragments and full-size genes) were considered. Moreover, just 433 proteins were detected as having common genes in their vicinity.

3.6. Analysis of the homologs of the S. cerevisiae FLR1 gene (cluster T, lineage 10)

The published literature reports that the DHA1 lineage 10, representing the homologs of the S. cerevisiae FLR1 gene, comprises the most complex evolutionary history from the set of DHA1 genes encoded in the genomes of the Saccharomycetaceae yeast species (Dias et al., 2010; Dias and Sá-Correia, 2014). The phylogenetic analyses of 71 full-size cluster T proteins encoded in 33 Saccharomycetaceae yeast strain allowed the subdivision of the phylogenetic tree in 5 sub-clusters. The gene neighbourhood approach allowed confirmed the previously published report (Dias et al., 2010) that gene duplication was a frequent phenomenon in the evolution of the FLR1 homolog genes (**Figure.2**).

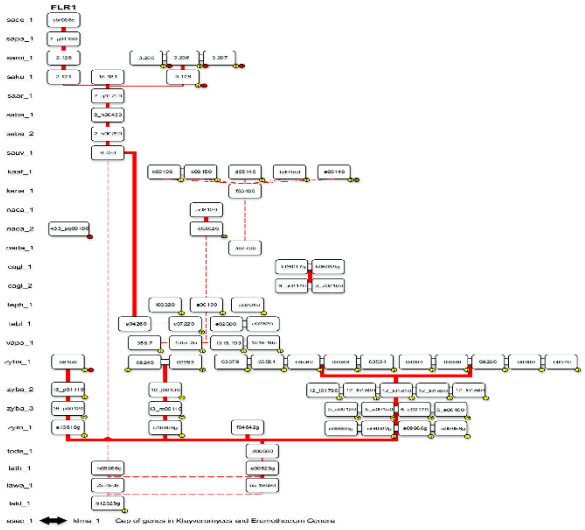

Figure. 2 Lineage 10 (homologs of S. cerevisiae FLR1 gene). Each box represents a gene. Lines connect genes sharing common neighbours. E, C, and T indicates that the corresponding gene was classified as a fragment, corrected or subtelomeric gene, respectively. The dashed line encompasses groups of proteins more similar in amino acid sequence (inferred from the analysis of the phylogenetic tree). HGT represents the plausible occurrence of events of horizontal gene transfer between species.

The gene neighbourhood analysis also showed that the chromosome environment where FLR1 homologues reside was poorly conserved and that lineage 10 does not exhibit the normal WGD gene pattern where the ancestral gene originates two ohnologs genes that can be either subsequently lost or not. The S. cerevisiae FLR1 gene, as well as the corresponding orthologues encoded in the genomes of the *Saccharomyces sensu stricto* group, share strong synteny with the tebl_1_a04260, a translated ORF encoded in an early-divergent post-WGD yeast species of the *Tetrapisispora* genus. In addition, the absence of neighbourhood conservation of FLR1 homologues belonging to successive post-WGD species suggests the occurrence either of multiple gene transpositions in the corresponding genomes or multiple gene loss and gain. This potential gain of cluster T members would require at least five lateral transference of genes between different yeast species.

3.7. Positive/ Negative Selection and Conservation analysis.

The Log-OddsLogo and Fast Unconstrained Bayesian AppRoximation (FUBAR) software were used to analyse of cluster T. Regarding as expected, the results of sequence conservation and negative selection obtained from these two methodologies show a high association, with the portions where the protein sequence is strongly conserved also showing a high probability of the parameter representing the synonymous substitution rate (**α**) being bigger than the parameter representing the non-synonymous substitution rate (**β**). In multiple alignment of the amino acid sequence of cluster T allowed the identification of 15 residues under positive selection that showed a high mutation rate

3.8. Phylogenetic and gene neighbourhood analyses of the remaining homologs of the *S. cerevisiae* genes encoded in the genomes of the Saccharomycetaceae species.

3.8.1. Lineage 1 (Homologs of the S. cerevisiae DTR1 gene)

Lineage 1 comprises a linear evolutionary history of 33 DHA1 genes. The gene neighbourhood analysis was able to identify the formation of two sub-lineages with origin in the WGD event. One of these sub-lineages comprises the orthologs of the S. cerevisiae DTR1 gene, being present in all genomes of yeast species evolving after the divergence of the Tetrapisispora genus. The other sub-lineage is composed by two translated ORFs, vapo_1_1048.45, and teph_1_l01760 encoded in the genomes of V. polyspora and T. phaffii species. This sublineage was lost during the evolution of yeast species evolving after the divergence of *Tetrapisispora* species.

3.8.2. Lineage 9 (Homologs of the S. cerevisiae TPO4 gene)

The 29 homologs of the S. cerevisiae TPO4 gene form a simple evolutionary history. The gene neighbourhood analysis revealed that one *T. phaffii* translated ORF (teph_1_g01430) and one *V. polyspora* translated ORF (vapo_1_413.2) encoding cluster S members are paralogs of the S. cerevisiae TPO4 gene, with origin in the WGD

event. These two translated ORFs form a DHA1 gene sub-lineage that after the divergence of the *Tetrapisispora* species is discontinued, being lost in the late-divergent post-WGD species.

3.8.3. Lineage 2 (Homologs of the S. cerevisiae QDR1/QDR2 and AQR1 genes)

The cluster F members are organized in the DHA1 gene lineage 2. Although this lineage comprises a great number of DHA1 genes, the corresponding evolutionary history fits well the commonly observed WGD pattern where the protoploid gene lineage is split into two sub-lineages, one originating the S. cerevisiae QDR1/QDR2 genes, and the other the S. cerevisiae AQR1 gene (Dias et al., 2010). During the evolutionary history of lineage 2, the occurrence of the WGD event is clear, with the presence of two independent and full lineages of genes. In the sublineage comprising the AQR1 gene, in the major part species, just is shown one gene per strain with the high number of neighbour connections. The cases T. *blattae, N castellii CBS 4309 and N. castellii NRRL Y-12630 are* the exception, since these yeasts species show horizontal expansions of AQR1 homologs. In the QDR1/QDR2 sub-lineage, the evolutionary history after the WDG is quite different, since various tandem repeat genes are observed, mainly in *Saccharomyces sensu stricto* species. From *V. polyspora* (vapo_1) to K. naganishii (kan_1), excluding T. blattae (tebl_1) species, just one member per strain is observed, although these genes share strong synteny among them.

3.8.4. Lineage 3 (Homologs of the S. cerevisiae QDR3)

Lineage 3 comprises the members of the phylogenetic cluster G, encompassing a total of 25 DHA1 genes. The analysis of the gene neighbourhood results showed the existence of strong synteny linking the cluster G members encoded in the yeast species classified in the *Saccharomyces* genus share synteny with those encoded in the protoploid Saccharomycetaceae species. However, the absence of genes in so many successive genera suggest that the cluster G members were lost immediately after the WGD event, being re-acquired by the yeast ancestral, from a horizontal gene transfer event, that originated the *Saccharomyces* species.

3.8.5. Lineage 4 (Homologs of the S. cerevisiae HOL1)

Lineage 4 comprises twenty-two members of the phylogenetic cluster J. The *S. cerevisiae* member of this cluster, the HOL1 gene, is not very well biochemically characterized. The cluster J members encoded in the genomes of the protoploid Saccharomycetaceae species reside in a highly conserved chromosome region. As reported in the previously published study focusing the evolution of the DHA1 genes in the Saccharomycetaceae family, lineage 4 show an evident gap between the pre- and post-WGD species. The lack of homologs of the *S. cerevisiae* HOL1 gene in six successive taxonomic genera reinforces the proposed hypothesis of the ancestral yeast originating the *Saccharomyces* species having acquired a cluster J member (Dias *et al.* 2010).

3.8.6. Lineage 5 (Homologs of the S. cerevisiae TPO2/TPO3 genes)

Cluster N1 (Tpo2p, Tpo3p) is composed of 40 full-size DHA1 genes, plus three fragmented and one corrected gene. The *S. cerevisiae* TPO2 and TPO3 genes have been reported as ohnologs. The origin of these two genes was assigned to the WGD event, where two distinct DHA1 sub-lineages were formed. Two evolutionary scenarios can explain these results. The first scenario consists of the *S. cerevisiae* TPO2 and TPO3 genes being true ohnologs and in the evolutionary history of the TPO3 gene, having been a consistent gene loss of TPO3 orthologs. This scenario requires accepting a high number of consecutive and independent gene loss events. The second scenario consists of the rejection of the S. cerevisiae TPO2 and TPO3 genes being true ohnologs. Through the synteny analyses in the vertical scheme of blocks was possible to obtain evidences supporting the first scenario, since the genes residing in the neighbourhood of each subcluster, after the WGD, were clearly distinct. On the other hand, the phylogenetic analyses suggested the second scenario as more probable, being open the question which of these scenarios are correct.

3.8.7. Lineage 8(Homologs of the S. cerevisiae YHK8)

Lineages 8 comprise 21 members of the phylogenetic cluster R, with the *S. cerevisiae* YHK8 being the prototype of this DHA1 phylogenetic cluster. Regarding the yeast species evolving after the WGD event, the analysis of the corresponding genome sequences shows that all of them encode a single cluster R member. In contrast with the suggestion proposed in the previously published study (Dias et al. 2010a), the evaluation of chromosome environment where the cluster R member resides in the

post-WGD species supports the existence of a single DHA1 gene sub-lineage.

3.8.8. Lineage 7 (Homologs of the S. cerevisiae TPO1)

Lineage 7 comprises a total of 43 genes encoding members of the phylogenetic cluster P, with 42 corresponding to full-size DHA1 transporters and 1 to a protein fragment. With the exception of *S. bayanus* MCYC 623 strain and the *Eremothecium* species, all yeast strains analysed in this work encode at least one cluster P member in their genome sequences. The reconstruction of the evolutionary history of the homologs of the *S. cerevisiae* TPO1 gene was consistent with the previous one proposed by Dias et al. (2010). The cluster P members encoded in the genomes of the protoploid Saccharomycetaceae species reside in a highly conserved chromosome environment. The gene neighbourhood analysis shows that the WGD event originated two sub-lineages, each comprising ohnologs of the DHA1 genes comprised in the other sub-lineage.

4. Discussion and Conclusions

The main goal of this MSc thesis was to reconstruct the evolution of the genes encoding drug: H+ antiporters of family 1 using 33 yeast strains classified in the Saccharomycetaceae taxonomic family. In complement, the diversity of the DAG family members encoded in the genomes of 78 hemiascomycetes strains was also attempted. The pursue of these two objectives this MSc thesis is based on two local databases developed by the BSRG group in the past five years. The first database comprises the genomic information gathered on these yeast ORFs and, therefore, is named Genome DB. The second database comprises pairwise amino acid sequence similarity information obtained using an all-against-all comparison between these yeast ORFs, being named Blastp DB.

Due to the high number of DHA1 genes identified in the yeast strains under analysis in this work (1382 proteins from 94 hemiascomycetous strains), it was decided to limit the scope of this study and only attempt to extend the published studies performed by Dias and co-workers (2010, 2014), focusing only on the Saccharomycetaceae taxonomic family.

In the previous studies focusing the evolution of the DHA1 and DAG genes and in this MSc thesis, the members of each of these gene families were identified by constraining and traversing of a blastp pairwise similarity network (Dias et al., 2013; Dias et al. (2014). However, although the past studies were able to identify the DHA1 and DAG proteins using a sole member of each of these two families as starting node for network traversal, in this MSc thesis, it was required using one reference gene for each cluster identified in the previous phylogenetic analyses for the traversal of the blastp network. The fundamental reason for the necessity of multiple rounds of network traversal from different members of each of these gene families was the increase in the genome DB and, consequently, of the Blastp DB.

The use of a wide range of different 2D-topology prediction software also allowed establishing the number of the TMS found in the 1382 full-size DHA1 proteins identified in this study. Interestingly, not all DHA1 transporters seem to obey the presumed rule of all of the members of this gene family spanning twelve times the cellular membrane onto which the encoded proteins are embedded. All the 98 DHA1 transporters residing in the phylogenetic cluster J, comprising the homologs of the S. cerevisiae HOL1 protein, were predicted having only eleven TMS. This result was based on the consistent report of the absence of one TMS in the N-terminal extremity of the Hol1 homologs by six software suites used to predict the 2D-topology of these transmembrane proteins (in a total of 10 different software's considered).

The construction of the phylogenetic tree allowed the identification of two additional clusters, labelled X and Y, when compared with the last published study focusing the DHA1 gene family (Dias and Sá-Correia 2014). This result is not unexpected considering that the number of hemiascomycetous genomes increased since then, spanning 63 additional strains corresponding to 38 additional species, many belonging to taxonomic families not sampled in the referred 2014 paper. Two phylogenetic clusters, D and K2, already described in the previously published studies (Dias *et al.* 2010; Dias and Sá-Correia 2014)) but not reported to exist in the Saccharomycetaceae yeast, were found encoded in the *Kluyveromyces wickerhamii*, *Kluyveromyces aestuarii* and *Kluyveromyces marxianus* species. Regarding the number of DHA1 proteins encoded in the genomes of the

Saccharomycetaceae species, some divergences in the number of DHA1 proteins residing in clusters F, P and T (homologs of the S. cerevisiae QDR1/QDR2/AQR1, TPO1, and FLR1, respectively) were also found between yeast species classified in the same taxonomic. Although these results were not expected, these can be explained when species of the same genus have a remote common ancestral, thus not sharing a close phylogenetic relation (Dujon 2010) or when their ecological niches or their geographic origin are very distinct. The number of DHA1 proteins found encoded in the hemiascomycetes yeast belonging to the CTG complex was high. In some cases, the reason for the abundance of these transporters resides in strong horizontal gene expansions. Our study allowed the identification two additional cases of horizontal gene expansions, in relation to the previous study of (Dias and Sá-Correia, 2014), where the same phenomena occurred - *P. sorbitophila*, *C. orthopsilosis*. Two reasons can explain these cases of strong horizontal gene expansion: 1) these additional genes may be used as backups of the original FLR1 and TPO1 homolog in these species; 2) these horizontal expansions can be interpreted as a reservoir of genes for the creation of functional novelty.

Consistent with the previous studies focusing the DHA1 genes (Dias et al., 2010; Dias and Sá-Correia, 2014), gene duplication was a frequent event in the majority of DHA1 gene lineages reconstructed in this MSc thesis. The synteny and phylogenetic analyses of 33 yeasts strains performed in this study also suggest the occurrence of HGT in lineage 3 (QDR3 homologs), lineage 4 (HOL1 homologs), lineage 8 (YHK8 homologs) and lineage 10 (FLR1 homologs). In cluster F two independent sublineages were observed, one with the homologs of QDR1/QDR2 gene and another one with the homologs of AQR1. This showed an unexpected evolutionary scenario since the early divergent post-WDG species just presented one gene in this sublineage. To explain the evolutionary history of this cluster two scenarios are possible. First, the initial creation of tandem repeat genes in early post-WGD species followed by the independent gene losses of at least eleven non-sequential species; or second, the loss of one gene in *V. polyspora* and just at the begin of *Saccharomyces sensu stricto* group occurred a local duplication (*S. uvarum* species), being the previous tandem repeat cases punctual and independent events. Observing the gene neighbourhood environment exist facts supporting both hypotheses equally, since there are more or less the same number of connections and shared neighbours between the tandem repeat genes of *Saccharomyces sensus stricto* and the remaining genes of the sublineage. Despite this, the loss of at least eleven gene members in eleven strains continued to be less likely, than the initial loss in *V. polyspora* and the local duplication ten strains after. The cluster N1 is composed by two independent sublineages, the first one comprise the TPO2 homologs genes while the second one comprise the homologs of TPO3 genes. On the other hand, the evolutionary origin of both sublineages caused some doubts and opened the possibility of another scenario, distinct of WGD scenario, to explain this cluster of genes. In order to explain this five lineage 3 scenarios were proposed. The *S. cerevisiae* TPO2 and TPO3 genes are true ohnologs in WGD event originate, or alternatively, in second and third scenario there is a rejection of the S. cerevisiae TPO2 and TPO3 genes as being true ohnologs. The remaining lineages, excepting the case of Lineage 10 that shows a general absence of synteny between the FLR1 homologous, showed a generally high level of conservation in their chromosomic environments. The Total lack of synteny in cluster T jointly with the non-conventional WDG pattern makes this lineage one of the most interesting here studied. In addition, both similarities and syntenic analyses support the idea of at least five HGT occurrences in this clusters, which is coherent with the subtelomeric localization of the most part of the genes comprised in this cluster.

The construction of a phylogenetic tree representing the full-size DAG proteins allowed the classification of the gathered transporters into twenty clusters. The analysis of this phylogenetic tree showed that the DAG members of the phylogenetic clusters B (Sge1/Azr1/Vba3/Vba5), C (Vba1/Vba2), D (Vba4), E (Atr1/YMR279C) and P (Arn3) were found in the majority of the hemiascomycetous species considered in this MSc thesis. In addition, the DAG members of the phylogenetic cluster S, containing the *S. cerevisiae* glutathione exchangers (Gex1/Gex2), were found encoded only in the genome of yeast belonging to the Saccharomycetaceae family. The higher number of yeast species considered in our work allowed to validate the cases of HGT in clusters B, F, J, N, S. Furthermore, this work suggests existence of new cases of HGT involving DAG genes in the phylogenetic clusters C, E, F,

J, and T that, in the future, should be confirmed using the more resolving analyses based on synteny

Overall, the results obtained in this MSc thesis sheds light in some interesting evolutionary patterns involving DHA1 genes. The detailed analysis of the evolution of this gene family in the Saccharomycetaceae taxonomic family allowed understanding dubious cases that could not be resolved by the available tools and genome sequences available in the 2010 paper. Here, the gene neighbourhood analyses clearly proved to be one of the most powerful methodology to understand the evolution of the DHA1 gene family. Notwithstanding, in the future, the DHA1 and DAG members of some specific phylogenetic clusters should be submitted to more detailed analyses using methods able to evaluate cases of positive and negative selection with the goal of linking specific amino acids to functional diversification of these proteins in cases where solid evidence exists that the origin of the encoding genes were on duplication or transfer event. This could enhance our understanding of the physiological role of specific DHA1 and DAG transporters and open the door for using the phenotypic features encoded by these important genes in Biotechnological, Pharmaceutical and Clinically applications.

5. Bibliography

Altschul SF, Wootton JC, Zaslavsky E *et al.* The construction and use of log-odds substitution scores for multiple sequence alignment. *PLoS Comput Biol* 2010;**6**:e1000852.

Anisimova M, Gascuel O. Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative. *Syst Biol* 2006;**55**:539–52.

Bernsel A, Viklund H, Falk J *et al.* Prediction of membrane-protein topology from first principles. *Proc Natl Acad Sci U S A* 2008;**105**:7177–81.

Bernsel A, Viklund H, Hennerdal A *et al.* TOPCONS: Consensus prediction of membrane protein topology. *Nucleic Acids Res* 2009;**37**:465–8.

Claros, M.G.; Von Heijne G. TopPred II : An improved software for membrane protein structure predictions Claros, Manuel G.; von Heijne, Gunnar CABIOS. *CABIOS* 1994;**10**:685–6.

Dhaoui M, Auchère F, Blaiseau P-L *et al.* Gex1 is a yeast glutathione exchanger that interferes with pH and redox homeostasis. *Mol Biol Cell* 2011;**22**:2054–67.

Dias PJ, Sá-Correia I. The drug:H^+ antiporters of family 2 (DHA2), siderophore transporters (ARN) and glutathione:H^+ antiporters (GEX) have a common evolutionary origin in hemiascomycete yeasts. *BMC Genomics* 2013;**14**:901.

Dias PJ, Sá-Correia I. Phylogenetic and syntenic analyses of the 12-spanner drug:H(+) antiporter family 1 (DHA1) in pathogenic Candida species: evolution of MDR1 and FLU1 genes. *Genomics* 2014;**104**:45–57.

Dias PJ, Seret M-L, Goffeau A *et al.* Evolution of the 12-spanner drug:H+ antiporter DHA1 family in hemiascomycetous yeasts. *OMICS* 2010;**14**:701–10.

Dujon B. Yeast evolutionary genomics. *Nat Rev Genet* 2010;**11**:512–24.

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004;**32**:1792–7.

Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution (N Y)* 1985;**39**:783–91.

Guindon S, Dufayard JF, Lefort V *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst Biol* 2010;**59**:307–21.

Von Heijne G. Membrane protein structure prediction: hydrophobicity analysis and the positive-inside rule. *J Mol Biol* 1992;**225**:487–94.

Huson DH, Richter DC, Rausch C *et al.* Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics* 2007;**8**:460.

Käll L, Krogh A, Sonnhammer ELL. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. *Nucleic Acids Res* 2007;**35**:W429–32.

Kosakovsky Pond SL, Pond SLK, Frost SDW *et al.* Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection. *Mol Biol Evol* 2005;**22**:1208–22.

Lesuisse E, Simon-casteras M, Labbe P. Siderophore-mediated iron uptake in Saccharomyces cerevisiae : the S/Tl gene encodes a ferrioxamine B permease that belongs to the major facilitator superfamily. *Microbiology* 1998;**144**:12379.

Lotia S, Montojo J, Dong Y *et al.* Cytoscape app store. *Bioinformatics* 2013:btt138.

Murrell B, Moola S, Mabona A *et al.* FUBAR: A Fast, Unconstrained Bayesian AppRoximation for Inferring Selection. *Mol Biol Evol* 2013;**30**:1196–205.

Murrell B, Wertheim JO, Moola S *et al.* Detecting individual sites subject to episodic diversifying selection. *PLoS Genet* 2012;**8**:e1002764.

Paulsen IT, Sliwinski MK, Nelissen B *et al.* Unified inventory of established and putative transporters encoded within the complete genome of Saccharomyces cerevisiae. *FEBS Lett* 1998;**430**:116–25.

Plotree D, Plotgram D. PHYLIP-phylogeny inference package (version 3.2). *cladistics* 1989;**5**:163–6.

Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. *Trends Genet* 2000;**16**:276–7.

Ronquist F, Teslenko M, van der Mark P *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol* 2012;**61**:539–42.

Sá-Correia I, dos Santos SC, Teixeira MC *et al.* Drug:H+ antiporters in chemical stress response in yeast. *Trends Microbiol* 2009;**17**:22–31.

Sá-Correia I, Tenreiro S. The multidrug resistance transporters of the major facilitator superfamily, 6 years after disclosure of Saccharomyces cerevisiae genome sequence. *J Biotechnol* 2002;**98**:215–26.

dos Santos SC, Sá-Correia I. Yeast toxicogenomics: lessons from a eukaryotic cell model and cell factory. *Curr Opin Biotechnol* 2015;**33**:183–91.

dos Santos SC, Teixeira MC, Dias PJ *et al.* MFS transporters required for multidrug/multixenobiotic (MD/MX) resistance in the model yeast: Understanding their physiological function through post-genomic approaches. *Front Physiol* 2014;**5**:1–15.

Tatusova TA, Madden TL. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett* 1999;**174**:247–50.

Tusnady GE, Simon I. The HMMTOP transmembrane topology prediction server. *Bioinformatics* 2001;**17**:849–50.

Viklund H, Bernsel A, Skwark M *et al.* SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. *Bioinformatics* 2008;**24**:2928–9.

Viklund H, Elofsson A. OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics* 2008;**24**:1662–8.

Waterhouse AM, Procter JB, Martin DMA *et al.* Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 2009;**25**:1189–91.