

APPORT DE LA SYNTÉNIE POUR L'IDENTIFICATION DES
ORTHOLOGUES CHEZ LES LEVURES :
MÉTHODOLOGIE ET APPLICATION À L'ÉTUDE DE
L'ÉVOLUTION DES TRANSPORTEURS

Marie-Line Seret

Thèse présentée en vue de l'obtention du grade de
docteur en sciences agronomiques
et ingénierie biologique

Jury:

Prof. Jacques Mahillon, président (Université catholique de Louvain)

Prof. Philippe Baret, promoteur (Université catholique de Louvain)

Prof. Michel Ghislain, secrétaire (Université catholique de Louvain)

Prof. Jean-Luc Souciet (Université de Strasbourg)

Prof. Xavier Draye (Université catholique de Louvain)

Louvain-la-Neuve, Septembre 2012

ABSTRACT

Comparative sequence analysis is widely used to infer gene function and study genome evolution and requires proper ortholog identification. We have developed a program for the inference of homology relationships and the identification of positional orthologs between closely related species. The algorithm combines two levels of evidence to determine co-ancestrality at the genome scale: sequence similarity and shared chromosomal neighborhood.

The method called IONS (for Identification of Orthologs by Neighborhood and Similarity) was first applied on some transporters family or subfamily, in particular on the Pleiotropic Drug Resistant (PDR) transporters which are involved in the export of antifungals. These small scale tests allowed validating and improving the method.

The algorithm was then applied to the complete genomes of nine hemiascomycetous yeasts (about 50,000 genes) studied within the Génolevures Consortium: *Saccharomyces cerevisiae*, *Candida glabrata*, *Zygosaccharomyces rouxii*, *Kluyveromyces thermotolerans*, *Saccharomyces kluyveri*, *Kluyveromyces lactis*, *Eremothecium gossypii*, *Debaryomyces hansenii* and *Yarrowia lipolytica*. Results were highly consistent with a manually curated database. The main advantages of the method are the automation and the absence of dataset-dependent parameterization, facilitating consistent integration of newly released genomes.

The inference of homology relationships were used to study the evolution of gene's families and for two other goals.

First, subsets of positional orthologs were used as anchor points to build syntenic blocks at the genome scale.

Second, subsets of homologs were used jointly with the Génolevures' families to proceed to the propagation of functional annotation based on the hierarchical classification of the Transport Classification database (TCDB) within a total of 5,173 transporters from ten yeasts (the nine mentioned above plus *Arxula adeninivorans*).

REMERCIEMENTS

Je tiens tout d'abord à remercier Philippe Baret, mon promoteur, pour sa disponibilité, sa patience, ses encouragements, sa compréhension et ses conseils avisés.

Merci à André Goffeau pour le temps qu'il a passé à me transmettre une infime partie de ses connaissances.

Merci à Jean-Luc Souciet et Bernard Dujon pour la confiance qu'ils m'ont accordée et leur enthousiasme vis-à-vis de la méthode développée.

Merci à Xavier Draye, Michel Ghislain et Fredj Tekaia pour leur accompagnement vers la réalisation de cette thèse et leurs remarques judicieuses qui m'ont permis d'améliorer la qualité du présent document.

Merci également à Jacques Mahillon d'avoir accepté de présider le jury de cette thèse.

Cette thèse a été réalisée en tant qu'assistante au cadre de l'UCL et a bénéficié du soutien des Fonds de soutien à la recherche (FSR) de l'Université catholique de Louvain, que je remercie également.

Au sein de Génolevures, un merci tout particulier à Tiphaine Martin pour sa gentillesse et le temps qu'elle a passé lorsque j'avais des soucis de base de données. Merci aussi à Thomas Rolland, Véronique Leh Louis, Anne Friedrich, Laurence Despons, David Sherman, Joseph Schacherer et l'ensemble des membres du consortium Génolevures pour les fructueuses collaborations et/ou leurs conseils.

Je tiens également à remercier vivement Julie Diffels pour sa collaboration, les discussions porteuses que nous avons partagées et pour sa compagnie agréable.

Merci à Margarida Palma, Paulo Jorge Dias et Isabel Sá Correia pour l'intérêt qu'ils ont porté à mon travail et pour notre collaboration.

Merci à mes mémorantes, Inês Azevedo Isidro et Cécile Desseille pour leur engagement et l'excellent travail qu'elles ont réalisé.

Merci à mes collègues : Steve pour sa disponibilité et sa gentillesse, Adeline, toujours prête à donner un coup de main, merci ;-), Julie, Pierre-Yves, Jean-François et Laurent (lors des festivités notamment), Olivier, Frédéric, Béatrice, Thérèse, Drissa, Laure-Anne, Carolin, Christel, Christian, Marc, Olivier, Anne-Laure, Muriel ou anciens collègues notamment Julien, Monique, Christel, Delphine, Gaëtan, Isabelle, Julie et tous les autres du

plateau pour leur bonne humeur et pour tout ce que l'on partage depuis sept ans pour certains.

Merci à mes amis qui, par les bons moments passés ensemble, donnent de belles couleurs à ma vie et me fournissent l'énergie pour continuer. Et un merci particulier à Céline pour son coup de main à la mise en page.

Je tiens aussi à exprimer toute ma gratitude envers ma famille et belle-famille, pour leurs encouragements et aides diverses, notamment la garde des enfants dans la dernière ligne droite.

Un merci tout particulier à ma maman pour tout le temps qu'elle a passé et passe encore pour nous ainsi qu'à mon papa pour lui avoir permis de le faire et pour son aide à lui aussi.

Enfin, merci à mon mari, pour son soutien (et le reste) et à mes enfants, Louis et Manon, les petits soleils qui illuminent ma vie.

TABLES DES MATIÈRES

Abstract	i
Remerciements	iii
Tables des matières	v
Liste des figures	ix
Liste des tableaux	xi
Liste des publications	xiii
Publications scientifiques	xiii
Mémoire	xiv
Acronymes	xv
I. Introduction	1
Présentation et objectifs de la thèse	1
Contexte de la problématique de l'identification des gènes orthologues pour l'annotation fonctionnelle	4
Homologie, orthologie, paralogie : quelques définitions	4
Notion de groupe d'orthologues dans un contexte multi-espèces	8
Qu'en est-il de l'aspect fonctionnel ?	10
Petite précision terminologique	12
II. Matériel	13
Les levures hémiacomycètes	13
Le projet Génolevures et ses bases de données	14
Recherche des orthologues : partie méthodologique	25
III. Les méthodes et bases de données existantes	27
Recherche des orthologues par comparaison de séquences	32
Recherche des orthologues basée sur la phylogénie	36
Recherche des orthologues utilisant le voisinage des gènes	39
Comparaison et benchmarking des méthodes	42
IV. Identification des orthologues par voisinage et similarité	47
Genèse et objectifs de la méthode	47
La méthode IONS	50
IONS: Identification of Orthologs by Neighborhood and Similarity an Automated Method to Identify Orthologs in Chromosomal Regions of	

Common Evolutionary Ancestry and its Application to Hemiascomycetous	
Yeasts-----	51
Abstract-----	51
Introduction-----	52
Material and Methods-----	53
Results-----	59
Discussion-----	64
Conclusions-----	67
Acknowledgements-----	68
Disclosures-----	68
References-----	68
Additional files-----	71
V. Utilisation des SONS comme point d’ancrage pour la synténie -	79
Définition de la synténie-----	79
Intérêts de la conservation de la synténie-----	80
VI. Discussion et perspectives de la partie méthodologique-----	83
Complémentarité des critères pour l’inférence de l’homologie-----	83
Forces de la méthode-----	83
Limites de la méthode-----	89
Perspectives-----	92
<i>Recherche des orthologues : application aux transporteurs chez les</i>	
<i>levures-----</i>	96
VII. Les protéines de transport-----	97
Rôles des transporteurs-----	97
Types de transporteurs-----	98
Système de classification-----	99
VIII. Etude de l’évolution d’une famille de transporteurs-----	101
Combined phylogeny and neighborhood analysis of the evolution of the ABC	
transporters conferring multiple drug resistance in hemiascomycete yeasts	103
Abstract-----	103
Background-----	104
Results-----	106
Discussion-----	114
Conclusion-----	117
Methods-----	118
Authors' contributions-----	119
Acknowledgements-----	119

References -----	120
Additional material -----	126
IX. Extension de la classification des transporteurs aux cinq nouvelles espèces séquencées-----	133
Classification des transporteurs-----	133
Extension de la classification-----	134
Methodology for the classification of transporter proteins and its application to ten hemiascomycete yeasts -----	137
Abstract -----	137
Introduction -----	138
Material and method-----	139
Results-----	146
Discussion -----	152
Acknowledgements-----	154
References -----	154
Additional files -----	156
X. Discussion et perspectives de la partie appliquée-----	159
Intérêts de la classification des transporteurs -----	159
Problématique des mises à jour des bases de données -----	160
Orthologie, homologie et conservation de la fonction-----	162
Perspectives pour l'amélioration de la classification -----	169
Conclusion générale -----	171
XI. Conclusion et perspectives -----	173
Les points forts des méthodes proposées-----	173
Contributions-----	175
Utilisation des ensembles d'orthologues définis par la méthode IONS pour l'annotation fonctionnelle -----	176
Perspectives -----	178
Conclusion-----	179
Références-----	181
Annexes -----	193
Annexe 1 : Comparative genomics of protoploid Saccharomycetaceae -----	195
Annexe 2 : Combined phylogentic and neighbourhood analysis of the hexose transporters and glucose sensors in yeasts	
Annexe 3 : Evolution of the 12-spanner Drug :H ⁺ Antiporter DHA1 Family in Hemiascomycetous Yeasts	

LISTE DES FIGURES

Figure 1. Schéma des différentes parties de la thèse. _____	3
Figure 2. Illustration de différents types de relations d'homologie et notion de groupes d'orthologues et inparalogues. _____	5
Figure 3. Illustration de la fonction attendue dans chacune des copies dupliquées selon le scénario d'évolution. _____	7
Figure 4. Illustration de différents cas de relation d'orthologie selon qu'il y ait présence ou non d'inparalogues dans chacun des organismes. _____	9
Figure 5. Vue d'ensemble des génomes de levures séquencés complètement ou partiellement (Dujon, 2010). _____	21
Figure 6. Illustration de la nomenclature utilisée par Génolevures. _____	22
Figure 7. Illustration des piliers et pistes du navigateur visuel du YGOB. _____	40
Figure 8. Histoire évolutive de deux génomes Π et Γ (Fu et al., 2007). _____	42
Figure 9. Illustration de la numérotation hiérarchique des sous-ensembles de gènes homologues. _____	87
Figure 10. Illustration d'un regroupement incorrect par transitivité. _____	89
Figure 11. Pourcentage des gènes de <i>S. cerevisiae</i> classés comme Ortho1 en fonction du pourcentage d'identité moyen entre orthologues de la paire d'espèces considérées. _____	90
Figure 12. Utilisation des différents types d'inférence de l'homologie pour la propagation de l'annotation fonctionnelle des transporteurs. _____	167

Figures de l'article "IONS: Identification of Orthologs by Neighborhood and Similarity - an automated method to identify orthologs in chromosomal regions of common evolutionary ancestry and its application to Hemiascomycetous yeasts".

Figure 1. Flowchart of the IONS algorithm. _____	54
Figure 2. Clustering by transitivity. _____	56
Figure 3. Improvement of the classification by the sequential procedure. _____	58
Figure 4. Cladogram of the nine hemiascomycetous yeasts. _____	59

Figures de l'article "Combined phylogeny and neighborhood analysis of the evolution of the ABC transporters conferring multiple drug resistance in Hemiascomycete yeasts"

Figure 1. Phylogenetic tree of the 56 PDR produced by Neighbor-Joining. _____	109
---	-----

Figure 2. Manually curated Subsets of Orthologs by Neighborhood and Similarity of Pdrp. _____	111
Figure 3. Relationships between the PDR sensu stricto based on the neighborhood analysis. _____	115

Figures du projet d'article "Methodology for the classification of transporter proteins and its application to ten Hemiascomycete yeasts".

Figure 1. Flowchart of the different steps of the update of the TC ID of genes of the YETI database. _____	141
Figure 2. Steps for the automatic assignment of a TC ID. _____	144
Figure 3. Working basis of the automatic assignment method. _____	145

LISTE DES TABLEAUX

Table 1. Comparaison de quelques méthodes d'identification et bases de données d'orthologues.	28
---	----

Tables de l'article "IONS: Identification of Orthologs by Neighborhood and Similarity - an automated method to identify orthologs in chromosomal regions of common evolutionary ancestry and its application to Hemiascomycetous yeasts"

Table 1. Comparison of the 7,927 <i>Génolevures</i> families (similarity) to the subsets produced by the IONS method.	60
Table 2. Types of subsets produced by the IONS method for the nine <i>Génolevures</i> species.	61
Table 3. Comparison of the classification of the 12,691 genes in subsets of orthologs by the IONS method, in the YGOB database, and by SYNERGY.	63
Table 4. Adjusted Rand index of the SONS to the YGOB pillars and SYNERGY orthogroups.	63

Table de l'article "Combined phylogeny and neighborhood analysis of the evolution of the ABC transporters conferring multiple drug resistance in Hemiascomycete yeasts"

Table 1: The members of the GL03C0025 family comprising all SACE Pdrp.	107
--	-----

Tables du projet d'article "Methodology for the classification of transporter proteins and its application to ten Hemiascomycete yeasts"

Table 1. Update of the gene content of the YETI database, due to modifications in databases since 2005.	140
Table 2. Number of genes treated by the 3 major steps of the procedure to update the TC ID of genes of the YETI database.	143
Table 3. Global statistics of transporter proteins currently in the TC and YETI databases and automatically assigned in ten Hemiascomycete species.	147
Table 4. Transporter proteins in ten Hemiascomycete species per subclasses and per species.	148
Table 5. Number of TC ID3 per type of evolution pattern and listing of the TC ID3 for which the chi-square test was significant at $p=0.05$.	149

LISTE DES PUBLICATIONS

PUBLICATIONS SCIENTIFIQUES

- Leh, V, Despons, L, Friedrich, A, Martin, T, Durrens, P, Casaregola, S, Neuveglise, C, Fairhead, C, Marck, C, Cruz, JA, Straub, ML, Kugler, V, Sacerdot, C, Uzunov, Z, Thierry, A, Weiss, S, Bleykasten, C, De, MJ, Jacques, N, Jung, P, Lemaire, M, Mallet, S, Morel, G, Richard, GF, Sarkar, A, Savel, G, Schacherer, J, Seret, ML, Talla, E, Samson, G, Jubin, C, Poulain, J, Vacherie, B, Barbe, V, Pelletier, E, Sherman, DJ, Westhof, E, Weissenbach, J, Baret, PV, Wincker, P, Gaillardin, C, Dujon, B and Souciet, JL. 2012. "*Pichia sorbitophila*, an Interspecies Yeast Hybrid, Reveals Early Steps of Genome Resolution After Polyploidization." *G3*.(Bethesda). 2(2):299-311.
- Seret, ML and Baret, PV. 2011. "IONS: Identification of Orthologs by Neighborhood and Similarity-an Automated Method to Identify Orthologs in Chromosomal Regions of Common Evolutionary Ancestry and its Application to Hemiascomycetous Yeasts." *Evolutionary Bioinformatics*. 7:123-133.
- Dias, PJ, Seret, ML, Goffeau, A, Correia, IS and Baret, PV. 2010. "Evolution of the 12-Spanner Drug:H⁺ Antiporter DHA1 Family in Hemiascomycetous Yeasts." *Omics-A Journal of Integrative Biology*. 14:701-710.
- Souciet, JL, Dujon, B, Gaillardin, C, Johnston, M, Baret, PV, Cliften, P, Sherman, DJ, Weissenbach, J, Westhof, E, Wincker, P, Jubin, C, Poulain, J, Barbe, V, Segurens, B, Artiguenave, F, Anthouard, V, Vacherie, B, Val, ME, Fulton, RS, Minx, P, Wilson, R, Durrens, P, Jean, G, Marck, C, Martin, T, Nikolski, M, Rolland, T, Seret, ML, Casaregola, S, Despons, L, Fairhead, C, Fischer, G, Lafontaine, I, Leh, V, Lemaire, M, De Montigny, J, Neuveglise, C, Thierry, A, Blanc-Lenfle, I, Bleykasten, C, Diffels, J, Fritsch, E, Frangeul, L, Goeffon, A, Jauniaux, N, Kachouri-Lafond, R, Payen, C, Potier, S, Pribylova, L, Ozanne, C, Richard, GF, Sacerdot, C, Straub, ML, Talla, E and Genolevures Consortium. 2009. "Comparative genomics of protoploid Saccharomycetaceae." *Genome Research*. 19:1696-1709.
- Seret, ML, Diffels, JF, Goffeau, A and Baret, PV. 2009. "Combined phylogeny and neighborhood analysis of the evolution of the ABC transporters conferring multiple drug resistance in hemiascomycete yeasts." *BMC Genomics*. 10:459.

- Palma, M, Seret, ML and Baret, PV. 2009. "Combined phylogenetic and neighbourhood analysis of the hexose transporters and glucose sensors in yeasts." *Fems Yeast Research*. 9:526-534.
- Diffels, JF, Seret, ML, Goffeau, A and Baret, PV. 2006. "Heavy metal transporters in Hemiascomycete yeasts." *Biochimie*. 88:1639-1649.

MÉMOIRE

- Seret, ML. 2007. "Génomique comparative des levures: identification des orthologues par étude de l'environnement chromosomique." Mémoire réalisé en vue de l'obtention du Diplôme d'Etudes Approfondies en Sciences Agronomiques et Ingénierie Biologique, Université Catholique de Louvain. Promoteur : P. Baret.

ACRONYMES

ABC : ATP-binding cassettes

ARAD : *Arxula adenivorans*

BBH : best bidirectional hit

CAGL : *Candida glabrata*

CDS : coding sequence

COG : cluster of orthologous group

DEHA : *Debaryomyces hansenii*

DHA1: drug:H⁺ antiporter

ERGO : *Eremothecium gossypii*

GL : Génolevures

GO : Gene Ontology

InPara: groupe d'inparalogues produit par la méthode IONS et contenant plusieurs gènes appartenant tous à la même espèce (Exemple de pp : 2,0,0).

IONS : Identification of Orthologs by Neighbourhood and Similarity

KLLA : *Kluyveromyces lactis*

KLTH : *Kluyveromyces thermotolerans*

KOG : euKaryotic Orthologous Groups

NJ : neighbour-joining

Ortho1 : groupe d'orthologues positionnels produit par la méthode IONS et comprenant exactement un gène par espèce (Exemple de pp : 1,1,1)

Orthox : groupe d'orthologues positionnels produit par la méthode IONS et dans lequel une espèce ou plus n'est pas représentée (Exemple de pp : 1,1,0)

PDR : Pleiotropic Drug Resistant

PISO : *Pichia sorbitophila*

pp : phyletic pattern

SONS : Subsets of Orthologs defined by Neighborhood and Similarity

SACE : *Saccharomyces cerevisiae*

SAKL : *Saccharomyces kluyveri*

TC : Transport classification

TC ID : identification number of the transport classification

Tsing : gène se retrouvant seul suite à l'analyse IONS

Undet : groupe d'homologues produit par la méthode IONS et contenant plus d'un gène dans un espèce au minimum (Exemple de pp : 1,2,0)

WGD : Whole Genome Duplication (duplication génomique complète)

YALI : *Yarrowia lipolytica*

YETI : Yeast transport information

YGOB : Yeast gene order browser

ZYRO: *Zygosaccharomyces rouxii*

I. INTRODUCTION

PRÉSENTATION ET OBJECTIFS DE LA THÈSE

Le travail que j'ai réalisé a été initié dans la continuité du travail de Benoît de Hertogh et Anne-Catherine Lantin sur les transporteurs transmembranaires chez les levures. Il est le fruit de la collaboration entreprise par l'ex-unité GENA, dirigée par le professeur Philippe Baret, avec le consortium français Génolevures à l'instigation du professeur André Goffeau. Ce dernier, pionnier du séquençage génomique, qui a dirigé le premier séquençage d'un organisme eucaryote entier, la levure *Saccharomyces cerevisiae*, est à la base du choix qui a été fait de travailler sur les levures et plus particulièrement sur les transporteurs transmembranaires des levures à l'étude desquels il a consacré l'essentiel de sa carrière scientifique.

Reflétant cette double influence des professeurs Philippe Baret et André Goffeau, ma thèse peut être divisée en deux grandes parties : une partie méthodologique et une partie plus appliquée (Figure 1).

L'objectif visé par la partie méthodologique est de définir et de programmer une méthode simple d'inférence de l'orthologie qui puisse facilement être appliquée à de multiples génomes. Cette méthode est présentée sous forme d'article dans sa langue de publication, l'anglais.

La seconde partie utilise l'outil créé dans la première pour réaliser des inférences d'homologie et d'orthologie en vue d'étudier l'évolution de familles de gènes et de réaliser une classification des transporteurs transmembranaires chez les levures hémiascomycètes.

Préalablement à sa publication, la méthode d'inférence des homologues et orthologues a été testée sur trois grandes familles de transporteurs : les transporteurs PDR (Pleiotropic Drug Resistant), les transporteurs d'hexose et la famille DHA1 (drug:H⁺ antiporter), faisant chacune l'objet d'une publication. Deux d'entre elles ont été écrites suite à une collaboration : avec Margarida Palma du Centro de Recursos Microbiológicos de la New University of Lisbon, pour les transporteurs d'hexose, et avec Paulo Jorge Dias et Isabel Sá Correia de l'Instituto Superior Técnico de la Technical University of Lisbon, pour la famille DHA1. La troisième, relative à l'évolution des transporteurs PDR (Pleiotropic Drug Resistant) est présentée au chapitre VIII et permet de faire le lien entre la partie méthodologique et la partie appliquée relative à la classification des transporteurs.

La classification des transporteurs constitue l'objectif principal de cette seconde partie. Le but est de mettre à jour la base de données YETI (Yeast Transport Information) en identifiant et annotant les transporteurs des cinq nouvelles espèces de levures séquencées par le consortium Génolevures. Cette partie s'appuie sur le travail méthodologique réalisé par Inês Azevedo Isidro dans le cadre de son mémoire de fin d'étude (Inês Azevedo Isidro, 2009). Elle est rédigée en anglais également, sous forme d'un projet d'article.

	Nombre d'espèces	Nombre de gènes	Page		
IONS	Recherches d'homologues dans une région chromosomique conservée	9	47874	51	Partie méthodologique
	Orthologues positionnels utilisés comme points d'ancrage pour la synténie	9	14481	79	
	Etude de l'histoire évolutive de familles de gènes:				Partie appliquée
	- Hexoses	9	75	101	
	- DHA1	13	189	101	
- PDR	9	62	103		
	Classification des transporteurs transmembranaires	10	5173	133	

Figure 1. Schéma des différentes parties de la thèse.

CONTEXTE DE LA PROBLÉMATIQUE DE L'IDENTIFICATION DES GÈNES ORTHOLOGUES POUR L'ANNOTATION FONCTIONNELLE

Depuis 1996, année de publication du premier génome eucaryote séquencé, *Saccharomyces cerevisiae* (Goffeau *et al.*, 1996), l'amélioration des méthodes expérimentales et le développement des outils assistés par ordinateur ont conduit à une efficacité croissante dans le séquençage de génomes. Cette efficacité accrue s'est rapidement traduite par une augmentation significative du nombre de génomes entièrement séquencés, qui atteint actuellement 170 espèces en ce qui concerne les eucaryotes et dépasse les 3000 génomes d'espèces différentes tous règnes confondus (<http://www.genomesonline.org>).

Avec l'augmentation du nombre de génomes séquencés, la caractérisation individuelle de chaque gène appartenant à chaque génome est rapidement devenue une tâche phénoménale via les analyses informatiques et totalement impossible de manière expérimentale. L'annotation fonctionnelle des gènes telle qu'elle a été réalisée pour *Saccharomyces cerevisiae* est et restera une exception. En effet, la grande majorité des gènes des génomes séquencés ne seront jamais étudiés de manière expérimentale. Donc pour la plupart des gènes, le seul espoir de caractérisation fonctionnelle vient de la génomique comparative, via le transfert de l'information fonctionnelle des gènes caractérisés expérimentalement à leurs orthologues.

Face à ce constat, on comprend aisément l'intérêt d'une méthode fiable d'identification des gènes orthologues.

HOMOLOGIE, ORTHOLOGIE, PARALOGIE : QUELQUES DÉFINITIONS

Peu après la parution du livre de Darwin « The Origin of Species », le concept d'homologie a acquis le sens qui lui est actuellement attribué pour décrire une origine évolutive commune : deux gènes sont homologues s'ils dérivent d'un même gène ancestral. La transitivité est une caractéristique de l'homologie qui implique que si A et B sont homologues et que B et C sont homologues, alors A et C sont homologues également.

Les termes « orthologues » et « paralogues » sont quant à eux introduits en 1970 par Fitch (Fitch, 1970).

Le préfixe « ortho » vient du grec ancien et convoie l'idée de droiture, de justesse. Il a été choisi car l'histoire des gènes orthologues reflète l'histoire des espèces (Fitch, 1970). Les gènes orthologues sont des gènes appartenant à des espèces différentes qui dérivent d'un seul et unique gène ancestral présent chez le dernier ancêtre commun des espèces comparées. Cette simple définition inclut deux déclarations explicites importantes. Premièrement, le fait que les gènes descendent d'un seul gène ancestral est essentiel. Si l'on découvre que le génome ancestral possédait deux gènes qui ont donné naissance aux gènes d'intérêt, il est incorrect de les considérer comme des orthologues, même si dans certains cas, la perte d'une copie différente dans chacune des espèces puisse donner l'apparence de l'orthologie (exemple Figure 2b, si le gène B et le gène C' sont perdus). Deuxièmement, la définition spécifie que le gène ancestral doit être présent chez le dernier ancêtre commun aux espèces comparées et pas dans un ancêtre commun quelconque, ce qui peut exclure les cas de transfert horizontal de gène, relativement fréquent chez les procaryotes. Par exemple à partir de la Figure 2a, si on imagine une situation hypothétique (non représentée sur la figure) où le gène A de l'espèce A serait acquis par l'espèce B par transfert horizontal, le gène A acquis par l'espèce B et le gène C de l'espèce C ne seraient pas orthologues car ils ne descendraient pas d'un même gène ancestral présent chez le dernier ancêtre commun (S_2), mais bien d'un gène ancestral présent chez un ancêtre commun plus éloigné (S_1).

Par contre la définition n'inclut pas certaines conditions qui pourraient paraître naturelles mais qui ne font pas partie du concept d'orthologie. Par exemple, la définition ne requiert pas qu'il s'agisse d'une relation de type un-un. Un gène peut avoir plusieurs orthologues dans une même espèce (exemple Figure 2a, le gène A possède deux orthologues dans l'espèce B, les gènes B et B').

Les gènes paralogues sont des gènes issus d'un événement de duplication. Le préfixe « para » vient également du grec et signifie « à côté de ». Suite à l'événement de duplication, les deux copies évoluent en parallèle dans l'histoire évolutive. La définition n'inclut aucune condition relative à l'appartenance des gènes à un même génome, ni aucune indication quant à l'âge de la duplication ayant conduit à l'émergence des gènes paralogues. On pourra cependant distinguer les paralogues anciens des paralogues plus récents par les termes outparalogues et inparalogues (Sonnhammer et Koonin, 2002), selon que l'événement de duplication ait eut lieu avant ou après l'événement de spéciation menant au dernier ancêtre commun des gènes considérés (Figure 2a et b).

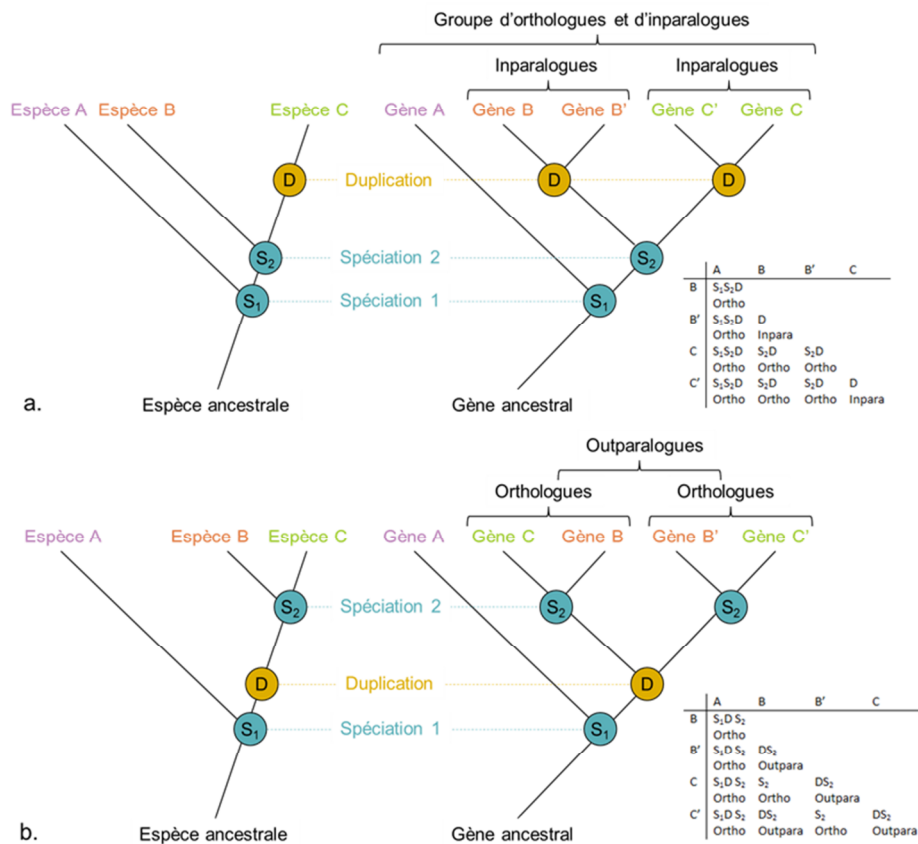


Figure 2. Illustration de différents types de relations d'homologie et notion de groupes d'orthologues et d'inparalogues.

A gauche, l'arbre phylogénétique de trois espèces. Au centre, celui d'un ensemble de gènes homologues. A droite, le tableau des relations d'homologie pour chaque paire de gènes. Le tableau reprend les différents événements depuis le dernier ancêtre commun de la paire de gènes et le type d'homologie : Ortho pour les orthologues, Inpara pour les inparalogues et Outpara pour les outparalogues.

a. L'événement de duplication ayant eu lieu après l'ensemble des événements de spéciations (S₁ et S₂), les gènes B et B' sont des inparalogues, tout comme C et C'.

b. L'événement de duplication ayant eu lieu avant le second événement de spéciation (S₂), les paires de gènes B-B', C-C', B-C' et B'-C sont des outparalogues.

L'apparition de gènes paralogues peut se faire de plusieurs manières : duplication d'un gène individuel, duplication d'un segment chromosomique, duplication d'un chromosome ou encore du génome entier (on parlera alors d'ohnologues). Chez les eucaryotes, le taux de duplication d'un gène est de 0.01 duplication par gène par million d'années (Lynch et Conery, 2000).

Une fois dupliquée, une séquence peut être fixée ou perdue. La probabilité de fixation d'un gène est liée à sa fonction. Les gènes dupliqués pour

lesquels le fait de présenter deux copies du gène présente une valeur sélective neutre par rapport au fait d'avoir une seule copie de ce gène ont une probabilité très faible d'être retenus en deux copies. Ceci suggère que pour une proportion importante de ces gènes dupliqués, une des copies va être rapidement perdue.

Les deux copies des gènes qui ont été dupliqués et maintenus en double exemplaire peuvent évoluer selon plusieurs scénarios (voir Figure 3):

– **la néofonctionnalisation** : une copie peut acquérir une nouvelle fonction bénéfique (Y dans la Figure 3) et dès lors être préservée par la sélection naturelle tandis que l'autre copie conserve la fonction originelle (X).

– **la subfonctionnalisation** : ou «partage de fonctions ancestrales». Si le gène originel avait deux "fonctions" (X et Y dans la Figure 3), les deux copies peuvent devenir partiellement compromises par l'accumulation de mutations de telle sorte qu'une des copies dupliquées peut perdre une des fonctions (Figure 3, la fonction Y est perdue par la copie B), tandis que l'autre copie perd la fonction complémentaire (Figure 3, la fonction X est perdue par la copie B'). Au final, l'organisme est toujours capable d'assurer les deux fonctions, mais par le biais de deux protéines différentes.

– **la conservation de la fonction** : les deux copies conservent la même fonction (X dans la Figure 3) que le gène originel. Pour certains gènes l'augmentation du nombre de copies et donc de la quantité de protéines a un effet bénéfique. Dans certains cas, les deux gènes gardent leur fonction originelle mais une des copies s'exprime dans un tissu spécifique ou acquiert une localisation cellulaire différente

– **la pseudogénisation** : une des deux copies du gène peut perdre sa fonction (la copie B' dans la Figure 3) et ne plus être exprimée, se transformant ainsi en pseudogène par l'accumulation de mutations dégénératives.

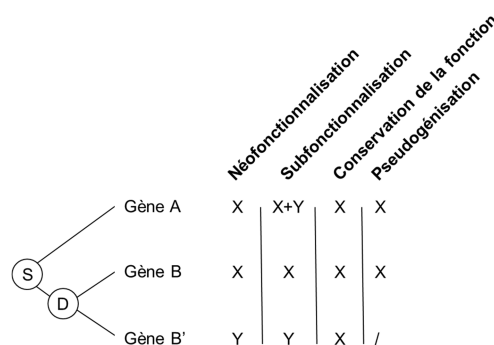


Figure 3. Illustration de la fonction attendue dans chacune des copies dupliquées selon le scénario d'évolution.

X et Y représentent deux fonctions distinctes, le backslash (/) représente une perte de fonction.

NOTION DE GROUPE D'ORTHOLOGUES DANS UN CONTEXTE MULTI-ESPÈCES

Il est important de noter que la relation d'orthologie ou de paralogie se définit par rapport à un triplet : deux gènes et un événement de spéciation de référence. Cette relation entre deux gènes s'étend à leurs descendants s'ils évoluent par de nouveaux événements de spéciation et/ou duplication. Dès lors, lorsque l'on s'intéresse à plus de deux organismes, la notion de groupe d'orthologues, et pas seulement de paires d'orthologues, permet de mieux représenter les relations ancestrales des gènes. Cependant la définition de ces groupes est complexe car contrairement à l'homologie, l'orthologie n'est pas transitive (par exemple Figure 2a, le gène A est orthologue du gène B, le gène A est orthologue du gène B', or les gènes B et B' ne sont pas orthologues mais bien inparalogues).

L'approche la plus commune est de former des groupes d'orthologues et d'inparalogues (Altenhoff et Dessimoz, 2009). Ces groupes sont construits de telle manière que chaque paire de gènes du groupe présente soit une relation d'orthologie, soit une relation d'inparalogie par rapport au dernier événement de spéciation de leur clade, c'est-à-dire que les inparalogues inclus dans le groupe sont des gènes d'une même espèce qui résultent d'un événement de duplication ayant eu lieu après l'ensemble des événements de spéciation du clade.

On parlera dès lors de groupes d'orthologues de type un-un, un-plusieurs ou plusieurs-plusieurs selon qu'il y ait absence totale d'inparalogues, présence d'inparalogues dans une espèce ou dans les deux (Figure 4).

En conséquence, dans ces groupes d'orthologues et d'inparalogues, deux gènes sont orthologues s'ils appartiennent à des espèces différentes et paralogues s'ils appartiennent à la même espèce. Suivant cette définition, l'ensemble des gènes illustrés à la Figure 2a forment un groupe d'orthologues et d'inparalogues. Il est à noter que ce type de groupement montre ses limites quand un ou plusieurs événements de duplication se sont produits après le premier et avant le dernier événement de spéciation.

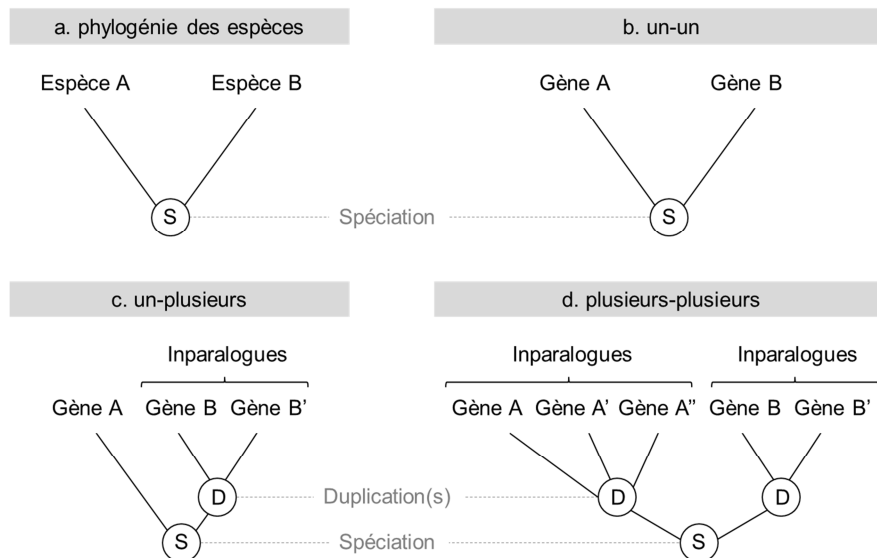


Figure 4. Illustration de différents cas de relation d'orthologie selon qu'il y ait présence ou non d'inparalogues dans chacun des organismes.

(a) Phylogénie des espèces. (b) Relation de type un-un où chacune des deux espèces possède une seule copie du gène ancestral. (c) Relation de type un-plusieurs dans laquelle le gène de l'espèce A est orthologues avec chacune des deux copies du gène ancestral présentes dans l'espèce B suite à un événement de duplication subséquent à l'événement de spéciation S. Par rapport à cet événement de spéciation, les deux copies présentes dans l'espèce B sont des inparalogues. (d) Relation de type plusieurs-plusieurs dans laquelle chacune des copies du gène ancestral présentes dans les espèces A et B suite à deux événements de duplication subséquents à l'événement de spéciation S sont orthologues. Par rapport à cet événement de spéciation, les trois copies présentes dans l'espèce A sont des inparalogues, de même que les deux copies présentes dans l'espèce B.

Dans ces cas, qui ne sont pas rares chez les eucaryotes, la nature non transitive de l'orthologie rend impossible la partition des gènes dans de tels groupes sans perdre de relation d'orthologie. En effet, si on prend l'exemple de la Figure 2b, on observe six paires de gènes orthologues (A-B, A-C, A-B', A-C', B-C et B'-C') et quatre paires d'outparalogues (B-B', C-C', B-C' et B'-C). L'ensemble des gènes ne forment donc pas un groupe d'orthologues et d'inparalogues puisqu'il existe des relations d'outparalogie. Si on tente de partitionner, on peut par exemple regrouper les gènes orthologues A, B et C, mais dans ce cas on perd les relations d'orthologie entre les paires de gènes A-B' et A-C'.

QU'EN EST-IL DE L'ASPECT FONCTIONNEL ?

Il est à noter que les définitions ci-dessus ne font nullement mention de la fonction des gènes.

Cependant, il est communément admis, que les gènes orthologues remplissent typiquement des fonctions équivalentes dans les organismes respectifs auxquels ils appartiennent (Koonin, 2005). Cette conception est à la fois théoriquement plausible et a pu être observée de nombreuses fois de manière expérimentale (Dolinski et Botstein, 2007).

Au niveau des paralogues, l'idée de la différenciation fonctionnelle des paralogues a été développée par Ohno (Ohno, 1970). Ohno propose que la duplication, par la formation de gènes paralogues, soit le principal processus responsable de l'émergence de nouvelles fonctions durant l'évolution. Selon lui, suite à la duplication, un des paralogues échappait à la pression des contraintes existantes au niveau de la fonction du gène ancestral et devenait libre d'évoluer vers une nouvelle fonction.

La différenciation fonctionnelle des paralogues est un sujet complexe qui a été adressé par de nombreuses études théoriques et empiriques (Ohno, 1970; Hughes, 1994; Lynch et Conery, 2000; Zhang, 2003; Koonin, 2005). Si bon nombre de biologistes s'intéressant à l'évolution considèrent que les paralogues ont des fonctions biologiques distinctes bien que généralement proches, la question de savoir à quel point la fonction des paralogues diffère et si la fonction du gène ancestral est généralement mieux conservée par les orthologues que par les paralogues ayant un degré de divergence de séquence similaire fait encore débat (Studer et Robinson-Rechavi, 2009; Nehrt *et al.*, 2011; Dessimoz *et al.*, 2012).

Si on se place au niveau d'un gène, les deux mécanismes qui, en augmentant le taux d'évolution, peuvent causer une modification de la fonction sont : (i) un relâchement de la sélection négative* ; et (ii) une augmentation de la sélection positive. Ces deux mécanismes peuvent tous deux résulter d'une modification de la taille de la population, de contraintes fonctionnelles, ... qui touchent aussi bien les paralogues que les orthologues. Donc de ce point de vue, l'évolution des paralogues et des orthologues peut être affectée par les mêmes mécanismes (Studer, Robinson-Rechavi, Cell press 2009).

Une étude récente (Nehrt *et al.*, 2011) a tenté de comparer le pouvoir prédictif des relations d'orthologie et de paralogie au niveau de la fonction

* Sélection contre un allèle qui diminue la valeur sélective (fitness) d'un individu, par opposition à la sélection positive qui se fait en faveur d'un allèle qui améliore la fitness

chez les mammifères. Cette étude est basée sur la comparaison des termes de l'ontologie des gènes (relatifs au processus biologique et à la fonction moléculaire) attribués de manière expérimentale aux gènes de l'homme et de la souris. Cette étude conclut que, en général, la similarité de fonctions entre paralogues est plus élevée qu'entre orthologues, bien qu'elle dépende de plusieurs facteurs (type de paralogie : inparalogues, outparalogues intra-espèce, outparalogues inter-espèces ; pourcentage d'identité de séquence ; localisation chromosomique relative des gènes dupliqués). Pour ces auteurs, les paralogues intra-espèces sont donc considérés comme les meilleurs prédicteurs de la fonction tandis que les orthologues et les outparalogues inter-espèces ont un pouvoir prédictif pratiquement identique et relativement faible ; l'inférence de la fonction d'un gène réalisée par propagation de l'information fonctionnelle disponible pour ses orthologues ou outparalogues inter-espèces, s'avérant juste dans moins de 70% des cas.

D'un point de vue méthodologique, cette étude de Nehrt *et al.* a fait l'objet d'une controverse au sein de la communauté scientifique, (Thomas *et al.*, 2012) critiquant l'utilisation des termes de l'ontologie des gènes (GO) pour évaluer la similarité de fonction et mettant en évidence le fait que les annotations GO sont souvent incomplètes, potentiellement biaisées (notamment par le type d'expériences réalisées sur différents organismes modèles) et sujettes à une supposition de « monde ouvert » (l'absence d'une annotation, n'impliquant pas l'absence de la fonction).

De façon plus générale, d'autres études (Peterson *et al.*, 2009; Henricson *et al.*, 2010; Forslund *et al.*, 2011; Huerta-Cepas *et al.*, 2011a) ont été entreprises récemment pour comparer les propriétés des orthologues et des paralogues et semblent généralement soutenir l'importance de distinguer les orthologues des paralogues (Dessimoz *et al.*, 2012), les orthologues restant une base fiable pour la propagation de l'annotation fonctionnelle.

Il est à noter que si la duplication, par la formation de deux copies paralogues, a un rôle important dans la différenciation fonctionnelle, on ne s'attend à une conservation de la fonction que chez les orthologues présents en une seule copie (orthologues de type un-un) si le processus évolutif est arrivé à son terme. De plus, il faudra rester prudent si la lignée présente une duplication génomique complète, comme c'est le cas chez certains hémiascomycètes, où même les orthologues présents en une seule copie ont connu une période d'évolution comme membre d'une paire redondante (Studer, Robinson-Rechavi, Cell press 2009).

Qu'ils s'agissent de gènes orthologues ou paralogues, on retiendra que plus leur ancêtre commun est récent, plus la probabilité qu'ils aient conservés des fonctions équivalentes augmente.

PETITE PRÉCISION TERMINOLOGIQUE

Il est crucial de comprendre que l'orthologie et la paralogie, tout comme la notion plus générale d'homologie, sont des concepts permettant de définir une relation entre des gènes (le fait d'être issu d'un gène appartenant à un ancêtre commun) basée sur une histoire évolutive particulière (impliquant, respectivement, un événement de spéciation ou de duplication). L'histoire évolutive réelle des gènes étant inconnue, les relations d'orthologie et de paralogie ne peuvent qu'être inférées suite à l'étude de caractères observables. La plupart des études génomiques actuelles utilisent la similarité des séquences géniques et des protéines qu'elles encodent comme caractères observables pour inférer les relations d'orthologie ou de paralogie, soit de manière directe ou indirecte (par les analyses phylogénétiques). L'incapacité à distinguer la différence entre les caractères observables et les inférences qu'ils permettent, a conduit à des problèmes terminologiques. Ainsi il n'est pas rare de voir dans la littérature des termes tels que « homologie forte » ou « pourcentage d'homologie ». Ces termes sont dénués de sens. En effet, soit la relation d'homologie existe (les gènes proviennent d'un ancêtre commun) soit elle n'existe pas. En pratique l'homologie est inférée et donc basée sur la significativité d'une probabilité. Lorsque l'inférence de l'homologie est basée sur l'identité ou la similarité de séquences, la confiance accordée à cette inférence est proportionnelle au pourcentage d'identité ou de similarité observé entre les séquences d'intérêt.

II. MATÉRIEL

LES LEVURES HÉMIASCOMYCÈTES

Les levures sont des champignons unicellulaires appartenant au phylum des ascomycètes, un vaste embranchement de champignons comptant plus de 30 000 espèces. Les plus connues sont la levure bourgeonnante *Saccharomyces cerevisiae* ou levure du boulanger qui fait partie de la classe des hémiascomycètes, et la levure fissipare *Schizosaccharomyces pombe* qui fait partie de la classe des archiascomycètes. Ces deux levures modèles sont les premières pour lesquelles les séquences complètes des génomes ont été obtenues, respectivement en 1996 (Goffeau *et al.*, 1996) et 2002 (Wood *et al.*, 2002).

La classe des hémiascomycètes, c'est-à-dire des levures se reproduisant par bourgeonnement, récemment renommée subphylum des Saccharomycotina (Kurtzman *et al.*, 2011), comprend à elle seule plus de sept cent espèces. Elle a fait l'objet de divers projets de génomique comparative ces dernières années (Wolfe et Shields, 1997; Souciet *et al.*, 2000; Kellis *et al.*, 2003; Dujon *et al.*, 2004; Kellis *et al.*, 2004; Piskur et Langkjaer, 2004; Dujon, 2005; Wolfe, 2006; Jackson *et al.*, 2009; Souciet *et al.*, 2009). Le projet Génolevures en particulier a permis le séquençage complet de huit espèces de levures (Dujon *et al.*, 2004; Souciet *et al.*, 2009; Leh *et al.*, 2012) sélectionnées en fonction de leurs positions phylogénétiques, pour leur intérêt spécifique en tant qu'agents pathogènes de l'homme, ou pour leur intérêt industriel ou environnemental (Souciet *et al.*, 2000). De ce fait, la classe des hémiascomycètes possède beaucoup d'avantages pour une étude de génomique comparative.

Premièrement, elle présente un des plus grands ensembles de génomes eucaryotes séquencés actuellement composé de 12 génomes séquencés entièrement et une vingtaine de génomes séquencés partiellement.

Deuxièmement, elle comprend une espèce modèle, *S. cerevisiae*, annotée par des approches expérimentales et pouvant servir de référence à l'annotation fonctionnelle.

Troisièmement, les levures hémiascomycètes sont des organismes eucaryotes unicellulaires qui possèdent des génomes compacts et de petite taille. De par ces caractéristiques, elles forment un groupe idéal pour étudier (plus facilement) les mécanismes évolutifs chez les eucaryotes.

Quatrièmement, le phylum des hémiascomycètes qui est supposé s'être séparé des champignons filamenteux (Euscomycètes) il y a environ 300 à 400 millions d'années, couvre une large gamme évolutive du point de vue des réarrangements chromosomiques et du degré de divergence de séquences (Souciet *et al.*, 2000; Dujon *et al.*, 2004). Les valeurs moyennes d'identité de séquences entre orthologues sont de 65%, 61%, 51% et 48%, respectivement pour *Candida glabrata*, *Kluyveromyces lactis*, *Debaryomyces hansenii* et *Yarrowia lipolytica*, par comparaison avec *S. cerevisiae* (Dujon, 2006). Ce qui signifie que les protéines orthologues de *Candida glabrata* et *S. cerevisiae* sont aussi différentes que celles de l'humain par rapport au poisson, par exemple. Ceci facilite l'exploration des processus évolutifs à diverses distances phylogénétiques (Dujon, 2006). Différents mécanismes évolutifs identifiés à partir de l'analyse comparative détaillée de nombreux génomes de levures et d'expériences spécifiques réalisées chez *S. cerevisiae* ont leur équivalent dans les génomes d'organismes multicellulaires (Dujon, 2006).

De plus, il s'agit d'un des rares phylums présentant un cas de duplication génomique complète (whole-genome duplication, en anglais, voir Figure 5, p 21) (Wolfe et Shields, 1997; Dujon *et al.*, 2004; Kellis *et al.*, 2004), celle-ci ayant eu lieu avant la spéciation de *Candida glabrata* et de *S. cerevisiae*.

LE PROJET GÉNOLEVURES ET SES BASES DE DONNÉES

Génolevures est un projet de génomique comparative à grande échelle entre *Saccharomyces cerevisiae* et d'autres espèces de levures appartenant à différentes branches de la classe des hémiascomycètes.

Le projet Génolevures est mis en œuvre par un groupement de recherche du Centre National de la Recherche Scientifique composé de sept laboratoires français.

L'objectif principal du projet est de répondre à des questions fondamentales relatives à l'évolution moléculaire, portant par exemple sur l'identification des gènes spécifiques à une espèce ou à une classe, la distribution des gènes au sein des familles fonctionnelles, leur taux de divergence ou encore les mécanismes de recombinaison aléatoire des chromosomes.

La base de données Génolevures (<http://www.genolevures.org>) stocke et rend accessible les données et résultats obtenus par le Consortium Génolevures à la suite de plusieurs campagnes d'annotation. La base de données est divisée en une partie publique accessible à tous par Internet, et une partie privée où les membres du Consortium annotent les nouveaux génomes séquencés.

LES ESPÈCES

La base de données Génolevures publique reprend actuellement (9/02/2012) les séquences génomiques complètes de dix espèces de levures, à savoir : *Saccharomyces cerevisiae*, *Candida glabrata*, *Zygosaccharomyces rouxii*, *Kluyveromyces (Lachancea) thermotolerans*, *Saccharomyces (Lachancea) kluyveri*, *Kluyveromyces lactis*, *Ashbya (Eremothecium) gossypii*, *Debaryomyces hansenii*, *Pichia sorbitophila* et *Yarrowia lipolytica* (Figure 5). La base de données privée contient en plus la séquence génomique d'*Arxula adeninivorans*.

Ces espèces présentent une diversité de cycles de développement (avec et sans reproduction sexuée). La plupart sont haploïdes à l'exception de *Kluyveromyces thermotolerans*, *Saccharomyces kluyveri* et de *Pichia sorbitophila* (Figure 5). Pour chaque espèce, une souche de type haploïde a été séquencée sauf pour *Pichia sorbitophila* où une souche diploïde a été séquencée.

Ces espèces sont brièvement décrites ci-dessous :

Saccharomyces cerevisiae

Saccharomyces cerevisiae ou levure du boulanger est également utilisée pour la fabrication de yaourt, de vin et de bière.

S. cerevisiae peut produire l'énergie nécessaire à sa survie et à sa reproduction de deux manières différentes, en fonction du milieu ambiant :

- En aérobiose : l'énergie est produite par la respiration qui transforme du glucose en dioxyde de carbone et ATP ou par l'utilisation d'éthanol. Dans les deux cas, il y a consommation d'oxygène ;
- En anaérobiose : l'énergie est produite par fermentation alcoolique du glucose.

La respiration est utilisée en cuisine, tandis que la fermentation est privilégiée pour la production d'alcool tel que le vin ou la bière.

Elle peut se reproduire de manière asexuée par bourgeonnement ou de manière sexuée.

Candida glabrata

Candida glabrata est actuellement le second organisme pathogène le plus fréquemment isolé à partir d'infections fongiques superficielles et envahissantes chez l'homme, après *Candida albicans*. Malgré son nom, il est phylogénétiquement plus proches des *Saccharomyces* que des autres *Candida* (Dujon *et al.*, 2004). Il est habituellement isolé à partir de sites

d'infections ou d'excréments de mammifères. C'est une levure haploïde qui peut montrer une croissance en pseudo-hyphes (Csank et Haynes, 2000). Aucune reproduction sexuée n'a jamais été décrite et aucune cellule diploïde n'a été isolée, bien que des gènes homologues aux gènes responsables de la reproduction sexuée chez *S. cerevisiae* aient été identifiés.

Comparé à *S. cerevisiae*, avec laquelle il partage la duplication génomique complète, *C. glabrata* montre un taux significativement plus élevé de perte de gènes, aboutissant à la perte de certaines fonctions spécifiques.

Zygosaccharomyces rouxii

Zygosaccharomyces rouxii est une levure haploïde parmi les plus tolérantes aux concentrations élevées en sel (halotolérance) et en sucre, bien que ces caractéristiques dépendent de la souche (Pribylová *et al.*, 2007). Elle est employée dans les processus de fermentation (par exemple pour le vinaigre balsamique ou pour développer l'arôme de la sauce soja) mais est également un des principaux microorganismes responsable des dommages dans l'industrie alimentaire notamment aux jus de fruits, confitures, vinaigrettes (Jansen *et al.*, 2003).

Kluyveromyces thermotolerans

Kluyveromyces thermotolerans est généralement isolé à partir de fruits ou de la drosophile. Malgré son nom, sa capacité à se développer à hautes températures est variable, bien qu'une souche collectée à proximité de figues de Barbarie en décomposition dans les îles Cayman croisse vigoureusement à 37°C. Le génome mitochondrial de 23,5 kb de long a également été entièrement séquencé (Talla *et al.*, 2005).

Saccharomyces kluyveri

L'habitat de *Saccharomyces kluyveri* n'est pas encore entièrement caractérisé, mais l'espèce semble être largement répandue dans l'environnement (drosophile, sol, arbres).

Contrairement à *S. cerevisiae*, *S. kluyveri* utilise la fermentation et dégrade les sucres uniquement en absence d'oxygène. En raison de cette utilisation efficace du glucose, *S. kluyveri* devient progressivement un organisme modèle pour des applications industrielles telles que la production de protéines en conditions aérobies limitées en glucose (Møller *et al.*, 2004). De plus il est capable d'utiliser les pyrimidines et leurs produits de dégradation comme unique source d'azote (Gojkovic *et al.*, 1998).

Kluyveromyces thermotolerans et *Saccharomyces kluyveri* étant phylogénétiquement proches, il a été proposé dans les réassignées au genre *Lachancea* (Kurtzman, 2003).

Kluyveromyces lactis

L'habitat naturel de *Kluyveromyces lactis* est varié, mais beaucoup de souches ont été isolées à partir de produits dérivés du lait.

K. lactis est devenu un modèle attractif pour étudier les relations entre génétique et physiologie. En particulier, des études comparatives ont permis la compréhension des stratégies physiologiques remarquablement différentes qui sont reflétées par la dominance de la respiration sur la fermentation chez *K. lactis* contrairement à *S. cerevisiae* (Schaffrath et Breunig, 2000).

K. lactis est adapté à l'aérobiose et peut s'accommoder d'une large gamme de source de carbone. De plus la biogenèse de son système respiratoire n'est pas réprimée par le glucose (Breunig *et al.*, 2000). *K. lactis* peut croître à une haute densité cellulaire en conséquence de quoi, il a été utilisé avec succès dans des applications industrielles liées à la production de biomasse et/ou à l'expression à grande échelle de produits de gène particuliers (Schaffrath et Breunig, 2000).

Eremothecium gossypii

Eremothecium gossypii, précédemment appelé *Ashbya gossypii* (Kurtzman, 1995), a été décrit pour la première fois en 1926 par Ashby et Nowell comme un pathogène de plante causant la stigmato-mycose des graines de cotonnier ou d'agrumes subtropicaux. Il a été rapidement reconnu et utilisé dans l'industrie pour sa capacité à produire de grandes quantités de riboflavine, ou vitamine B2, responsable de sa couleur jaune (Wendland et Walther, 2005).

Le séquençage complet du génome d'*E. gossypii* a révélé un petit génome de 8,7 Mb répartis sur 7 chromosomes et contenant 4715 gènes codant pour des protéines (Souciet *et al.*, 2009).

Debaryomyces hansenii

Debaryomyces hansenii est une levure marine cryotolérante qui peut tolérer des niveaux de salinité jusqu'à 24%, alors que la croissance de *S. cerevisiae* est inhibée lorsque la salinité atteint 10%. Sa tolérance au froid et au sel est responsable de son rôle important dans divers processus agroalimentaires. En effet, *D. hansenii* est l'espèce la plus communément trouvée dans tous les types de fromages, contrairement à d'autres espèces de levures, dont la prédominance est dépendante du type de fromage (Seiler et Busse, 1990). *D. hansenii* est aussi commun dans les laiteries et dans la saumure (Seiler et Busse, 1990) parce qu'il peut se développer en présence de sel à basse température et métaboliser des acides lactiques et citriques. *D. hansenii*

présente également des activités protéolytiques et de lipolyse pendant l'affinage de fromages.

La plupart des souches de *D. hansenii* sont haploïdes et se reproduisent très rarement. D'un point de vue génomique, *D. hansenii* semble avoir le pourcentage de génome codant le plus élevé parmi les levures, avec 79.2 % de génome codant (6906 séquences codantes prédites sur un génome de 12,2Mb). *D. hansenii* est aussi la levure avec le génome le plus redondant (51.5 %). Les duplications de gène en tandem y sont 5 fois plus fréquentes que dans d'autres levures, tandis que peu de blocs dupliqués sont détectables (Dujon *et al.*, 2004). Les transporteurs d'allantoate sont les protéines les plus surreprésentées (Lépingle *et al.*, 2000).

Pichia sorbitophila

Pichia (Millerozyma) sorbitophila a été isolée en 1980 comme un contaminant d'une solution de sorbitol à 70 % (Rodrigues de Miranda *et al.*, 1980). Cette levure osmotolérante diffère significativement d'autres espèces de levures osmotolérantes ou halotolérantes dans sa capacité à résister à des concentrations extrêmement élevées en sucre ou en sel (NaCl à 3 à 4M). Malgré son osmotolérance, *P. sorbitophila* n'est pas classé comme halophilique (ni osmophilique) (Maresova et Sychrova, 2003).

Oliveira *et al.* (1996) ont croisés des clones provenant d'une ascospore entre eux et suggèrent dès lors que *P. sorbitophila* soit homothallique, c'est-à-dire que les gamètes des deux sexes sont produits par un seul gamétophyte.

P. sorbitophila est une levure diploïde provenant de l'hybridation de deux ancêtres distincts appartenant au groupe *Millerozyma*, l'un d'eux étant étroitement lié à *P. farinosa*. Les paires de chromosomes A/B et C/D sont en partie hétérozygotes et en partie homozygotes, G/H et K/L sont homozygotes et M/N et E/F/I/J sont hétérozygotes. Les quatre derniers chromosomes ne sont pas séparés en deux paires clairement distinctes car un changement d'identité est observé de E/I à E/F et de F/J à I/J. Cette rupture de synténie résulte d'un événement de translocation qui est très probablement apparu dans un des génomes parentaux avant l'événement d'hybridation (Leh *et al.*, 2012).

Les régions homozygotes résultent de la perte de « l'hétérozygotisation » qui a eu lieu dans le génome de *P. sorbitophila*. Les gènes situés dans ces régions proviennent d'un parent unique et sont représentées par deux copies alléliques identiques (Leh *et al.*, 2012).

Arxula adeninivorans

Arxula adeninivorans (précédemment connu comme *Trichosporon adeninivorans*) a été isolé à partir d'un sol et décrit pour la première fois en 1984 (Middelhoven *et al.*, 1984). C'est une levure haploïde non pathogène capable d'assimiler et de fermenter de nombreux composés comme unique source de carbone (notamment l'adénine et d'autres purines) ou d'azote et qui dégrade l'amidon de manière efficace. Il est thermo et halorésistant (il peut croître jusqu'à 48°C dans un milieu contenant plus de 20% de NaCl).

Jusqu'à 42°C sa croissance se fait sous forme de cellules bourgeonnantes. A des températures plus élevées il forme un mycélium. Ce dimorphisme conditionné est réversible. Les cultures sous forme de mycélium accumulent des concentrations en protéines deux fois plus élevées et ont une activité de glucoamylase de 2 à 5 fois plus élevées que les cellules bourgeonnantes. C'est pourquoi il est utilisé pour l'expression de protéines hétérologues et comme donneur de gènes pour construire des levures adaptées d'un point de vue biotechnologique. Par exemple son gène responsable de la glucoamylase a été exprimé chez *S. cerevisiae* de manière à le rendre capable d'assimiler et de fermenter l'amidon (Wartmann et Kunze, 2000).

Yarrowia lipolytica

Yarrowia lipolytica est régulièrement isolé à partir de différents produits alimentaires (fromages, saucisses...) ou à partir d'environnements naturels comme des gisements de pétrole. Son incapacité à vivre en condition anaérobie permet une élimination aisée des produits laitiers. La plupart des isolats sont haploïdes, mais un état diploïde stable existe à partir duquel on peut induire la formation d'asques comprenant quatre spores. Les cellules se différencient sous forme de levure, de pseudohyphes ou de vrais hyphes septés selon les conditions de croissance.

Y. lipolytica est physiologiquement et taxonomiquement très éloigné de *S. cerevisiae*. C'est une levure aérobie obligée, qui peut utiliser des hydrocarbures normaux et des graisses diverses comme sources de carbone ce qui a conduit à différentes applications industrielles, généralement reconnues comme sans-danger, s'étendant de la production d'une protéine cellulaire simple ou hétérologue (insérée par recombinaison génétique), à la synthèse d'acide citrique ou d'arômes (Barth et Gaillardin, 1997).

Parmi les levures séquencées entièrement, *Y. lipolytica* présente le plus grand génome (20Mb). Beaucoup de gènes ont une similarité plus grande avec les champignons filamenteux qu'avec *S. cerevisiae* ou n'ont aucun orthologue connu (Dujon *et al.*, 2004).

Quelques caractéristiques chiffrées des génomes de ces espèces sont présentées à la Figure 5 qui reprend la ploïdie, le nombre haploïde de chromosome, la taille du génome (en Mb) et le nombre de gènes prédits comme codant pour une protéine.

La majorité des études présentées dans cette thèse portent sur les neuf premiers génomes rendus accessibles sur Génolevures (soulignés par un trait rouge continu dans la Figure 5), soit tous à l'exception d'*Arxula adenivorans* et de *Pichia sorbitophila*.

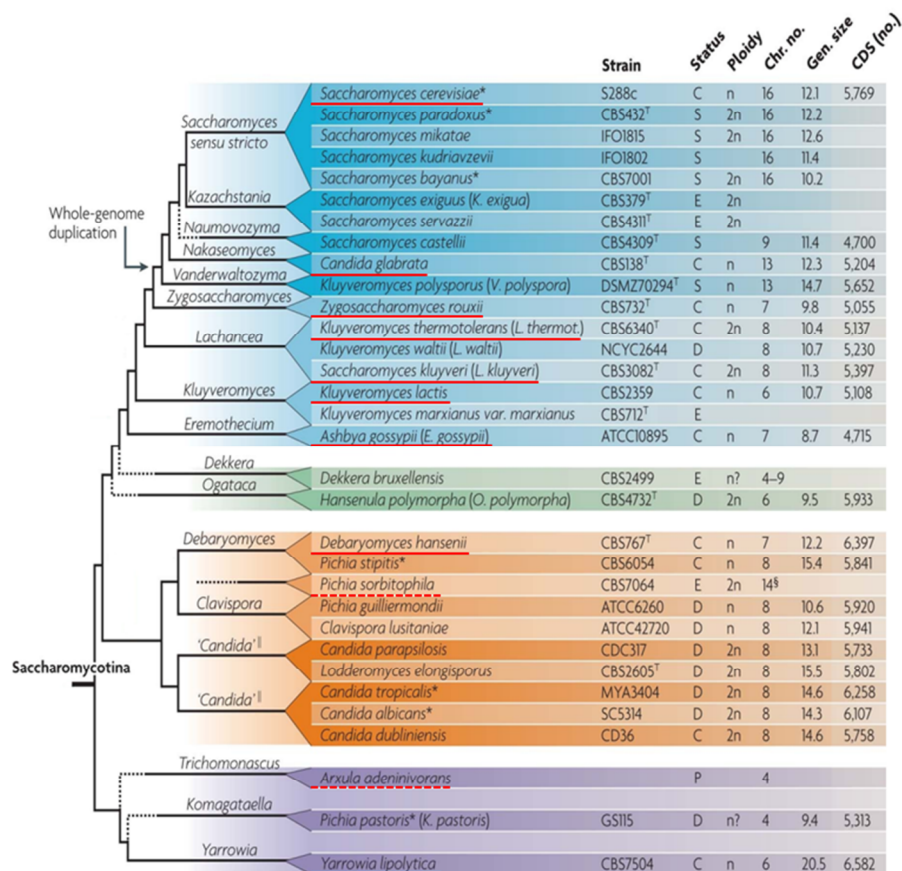


Figure 5. Vue d'ensemble des génomes de levures séquencés complètement ou partiellement (Dujon, 2010).

Les noms originaux des espèces au moment de la publication de la séquence sont indiqués (nouvelle dénomination entre parenthèses). Les triangles de couleur correspondent aux clades (ou genres). Les pointillés désignent les liens incertains ou ambigus. Les longueurs des branches sont arbitraires. Les onze espèces présentes dans les bases de données Génolevures sont soulignées en rouge. Parmi elles, les neuf espèces étudiées dans l'article présentant notre méthode d'identification des orthologues sont soulignées en traits continus. Les six colonnes situées à droite donnent respectivement la souche (strain), le statut de l'assemblage de la séquence (C, complet y compris le finissage ; D, ébauche d'assemblage ; E, étude exploratoire du génome ; P, séquençage en cours ; S, séquençage génomique complet avec une couverture de 3 à 7x assemblés en ~500 à 2000 contigs.), la ploïdie et le nombre haploïde de chromosomes ainsi que la taille du génome (en mégabases) et le nombre de gènes prédits comme codant pour une protéine (CDS). [§]Le génome nucléaire de *Pichia sorbitophila* (21,5 Mb) est composé de 7 paires de chromosomes suite à une hybridation et partiellement redevenu homozygote. Il compte un total de 11252 CDS (Leh *et al.*, 2012).

LE CHOIX DES DONNÉES GÉNOLEVURES

En plus des différents avantages des levures hémi-ascomycètes pour les études de génomique comparative, les données Génolevures présentent plusieurs atouts :

- une couverture ou profondeur de séquençage des génomes (c'est-à-dire le rapport entre la longueur de l'ensemble des séquences lues mises bout à bout et la longueur du génome cible) généralement élevée (> 6 X). permettant ainsi d'éviter les artefacts de perte de gènes observés avec des séquençages à faible couverture (Milinkovitch *et al.*, 2010). Comme nous le verrons ultérieurement, la profondeur du séquençage est importante du fait que la méthode développée dans le cadre de cette thèse est basée sur une étude du voisinage chromosomique.
- une annotation génomique cohérente des différentes espèces, les mêmes règles étant utilisées au sein du consortium pour l'annotation de l'ensemble des espèces.
- une nomenclature unique (voir ci-dessous) et un format de présentation identique pour les différentes espèces.

NOMENCLATURE DES ÉLÉMENTS CHROMOSOMIQUES

La base de données Génolevures utilise une nomenclature systématique pour les différents éléments chromosomiques (Durrens et Sherman, 2005). Cette nomenclature est constituée des éléments suivants : l'espèce, le projet, l'identifiant du chromosome, le numéro consécutif et le type (Figure 6).

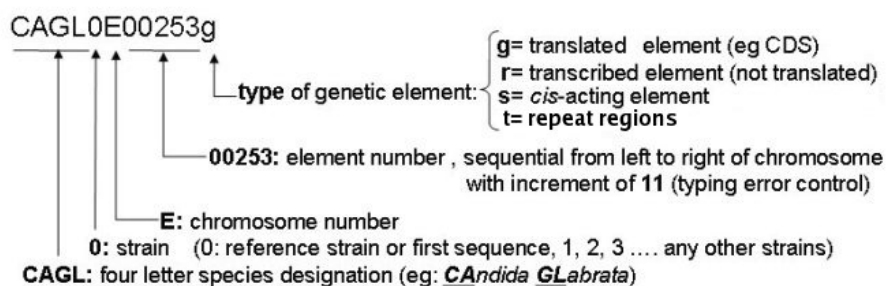


Figure 6. Illustration de la nomenclature utilisée par Génolevures.

Cette illustration est tirée du site Génolevures (<http://www.genolevures.org>).

L'espèce donne le nom de l'organisme séquencé, en utilisant un code à quatre lettres, deux pour le genre, deux pour l'espèce.

Le projet donne le numéro de la séquence de ce génome en commençant par 0. Ceci peut s'avérer utile pour différencier différents génomes ou souches de la même espèce, dans le cas d'un re-séquençage ou d'une ré-annotation.

L'identifiant du chromosome permet de différencier les chromosomes en utilisant une lettre écrite en majuscule.

Le numéro consécutif attribue un chiffre unique ordonné, augmentant de gauche à droite du brin séquencé, d'un télomère à l'autre. La numérotation initiale procède par incrément de 11 afin de laisser de la place pour d'éventuelles insertions.

Le type donne une indication optionnelle sur le type d'élément par une lettre minuscule : « g » pour un élément qui a ou pourrait avoir un produit de translation (protéine, pseudogène, ...); « r » pour un élément qui a ou pourrait avoir un produit de transcription uniquement (tRNA, rRNA, ...); « s » pour les éléments cis-actif n'ayant ni produit de translation, ni produit de transcription et « t » pour les régions répétées.

Les éléments sont numérotés indifféremment selon leur type. La nomenclature ne donne aucune indication par rapport au brin.

Dans la méthode présentée dans la partie méthodologique, seuls les éléments étant identifiés comme codant pour une protéine (CDS) sont utilisés pour l'étude du voisinage chromosomique, en effet, ils sont les seuls à être classés en famille sur la base de leur similarité, ce qui constitue un prérequis pour la méthode développée.

CLASSIFICATION DES PROTÉINES EN FAMILLES

Les protéomes des génomes complètement séquencés des bases de données Génolevures sont classés en familles d'après leur similarité dans des comparaisons de type tout-contre-tout. Quatre types d'alignements sont utilisés pour ces comparaisons : des alignements de Smith-Waterman (Smith et Waterman, 1981) avec et sans contrainte homéomorphique (recherche de la similarité sur la longueur complète de la séquence) et des alignements Blast (Altschul *et al.*, 1990) avec et sans contrainte homéomorphique. Chacun de ces quatre types d'alignements a servi de base à des clusterings réalisés sous différentes combinaisons de paramètres. Finalement, les familles sont constituées sur base d'un algorithme de consensus de clustering. La méthode originale est décrite dans (Nikolski et Sherman, 2007). Quatre types de familles sont définis :

1. Les familles robustes (GL3R.*)[†] sont celles qui restent identiques quels que soient les paramètres utilisés et sont les plus fiables. Certaines d'entre elles contiennent un « noyau » de protéines regroupées par H. Ferry et M. Nikolski d'après la littérature
2. Les familles consensus (GL3C.*) sont basées sur une combinaison de paramètres évalués par une procédure d'élection de Condorcet et, dans certains cas, par une vérification manuelle. Elles sont souvent constituées de plusieurs familles rassemblées.
3. Les familles (GL3M.*) ont une constitution très variable dépendante des paramètres utilisés. La plupart d'entre elles sont des familles notoirement compliquées comme les polyprotéines (produit d'expression d'un gène en une protéine, principalement non fonctionnelle qui est ensuite clivée en protéines plus petites capables d'assurer leur fonction) par exemple.
4. Les familles uniques (GL3U.*) qui correspondent à des singletons (une protéine par famille).

Les hémiascomycètes et plus particulièrement le sous-ensemble étudié par le projet Génolevures constitue donc un cadre pertinent d'un point de vue biologique et méthodologique pour l'établissement d'une méthode de recherche des orthologues.

[†] Le sigle « * » est remplacé par un identifiant de famille arbitraire composé de 4 chiffres.

RECHERCHE DES ORTHOLOGUES :
PARTIE MÉTHODOLOGIQUE

III. LES MÉTHODES ET BASES DE DONNÉES EXISTANTES

Un grand nombre de méthodes de prédiction et de bases de données d'orthologues ont été conçues ces dernières années.

Ces méthodes ou bases de données peuvent être divisées en trois grandes catégories selon le type d'informations prises en compte pour identifier les relations d'orthologie: celles basées sur les comparaisons de séquences, celles basées sur les arbres phylogénétiques et celles prenant en compte le voisinage des gènes (voir Table 1).

Pour chacune de ces trois catégories, nous présenterons ci-dessous quelques exemples, il en existe bien d'autres. Ces exemples ont été choisis pour leur notoriété, représentativité ou originalité, avec une préférence pour les bases de données incluant un génome de levure.

Table 1. Comparaison de quelques méthodes d'identification et bases de données d'orthologues.

	Méthode ou base de données	Détection de l'homologie	Nombre d'espèces et rang taxonomique	Etape manuelle	Site ou référence
PAR COMPARAISON DE SÉQUENCES					
Par paires d'espèces	Best Bidirectional Hit (BBH) ^{*,a,b,d}	BLASTP		Non	
	Reciprocal Smallest Distance (RSD) ^{*,b,d}	BLASTP		Non	(Wall <i>et al.</i> , 2003)
	Inparanoid ^{a,b,c,e}	BLASTP	99 eucaryotes + <i>E. coli</i>	Non	http://inparanoid.sbc.su.se
	RoundUp ^c	BLASTP	1807 espèces dont 226 eucaryotes	Non	http://roundup.hms.harvard.edu

*BBH et RSD ne sont pas à proprement parler des méthodes d'identifications des orthologues, mais elles sont à la base de certaines méthodes d'identifications d'orthologues et servent fréquemment de référence dans le cadre du benchmarking des méthodes d'identification des orthologues.

Méthodes évaluées dans l'étude de ^a (Hulsen *et al.*, 2006), ^b (Chen *et al.*, 2007), ^c (Altenhoff et Dessimoz, 2009), ^d (Salichos et Rokas, 2011), ^e (Boeckmann *et al.*, 2011) et ^f (Trachana *et al.*, 2011)

	Méthode ou base de données	Détection de l'homologie	Nombre d'espèces et rang taxonomique	Etape manuelle	Site ou référence
PAR COMPARAISON DE SÉQUENCES					
Multi-espèces	KOG ^{a,b} /COG ^c	BLASTP	7 eucaryotes / 66 unicellulaires	Oui	http://www.ncbi.nlm.nih.gov/COG/new
	MultiParanoid ^d	BLASTP	3 eucaryotes		http://multiparanoid.sbc.su.se
	OrthoMCL ^{a,b,c,d,f}	BLASTP	150 espèces (eucaryotes et procaryotes)		http://www.orthomcl.org
	HomoloGene ^c	BLASTP	20 eucaryotes		http://www.ncbi.nlm.nih.gov/homologene
	eggNOG (v. 3) ^{c,e,f}	FASTA	1113 espèces dont 121 eucaryotes		http://eggnoг.embl.de/version_3.0
	OMA ^{c,e,f}	Smith-Waterman	1109 espèces dont 124 eucaryotes		http://www.cbrg.ethz.ch/research/orthologous
	OrthoDB ^{e,f}	Smith-Waterman	166 espèces dont 73 champignons		http://cegg.unige.ch/orthodb5

Méthodes évaluées dans l'étude de ^a (Hulsen *et al.*, 2006), ^b (Chen *et al.*, 2007), ^c (Altenhoff et Dessimoz, 2009), ^d (Salichos et Rokas, 2011), ^e (Boeckmann *et al.*, 2011) et ^f (Trachana *et al.*, 2011)

Méthode ou base de données	Détection de l'homologie	Nombre d'espèces et rang taxonomique	Etape manuelle	Site ou référence
IMPLIQUANT LA CONSTRUCTION D'ARBRES PHYLOGÉNÉTIQUES				
phylomeDB	Smith-Waterman	808 espèces dont 60 champignons		http://phylomedb.org
PHANTHER ^e	BLASTP	48 espèces dont 5 champignons	Oui	http://www.pantherdb.org
TreeFam ^f	BLAST	79 espèces d'animaux	Oui	http://www.treefam.org
Ensembl Compara (64) ^{c,e}	BLASTP	61 chordés		http://www.ensembl.org
HOGENOM ^c	BLASTP	1470 espèces dont 25 champignons		pbil.univ-lyon1.fr/databases/hogenom
RIO ^b	Base de données Pfam			(Zmasek et Eddy, 2002)
Orthostrapper ^b	Familles d'homologues			(Storm et Sonnhammer, 2002)

Méthodes évaluées dans l'étude de ^a (Hulsen *et al.*, 2006), ^b (Chen *et al.*, 2007), ^c (Altenhoff et Dessimoz, 2009), ^d (Salichos et Rokas, 2011), ^e (Boeckmann *et al.*, 2011) et ^f (Trachana *et al.*, 2011)

Méthode ou base de données	Détection de l'homologie	Nombre d'espèces et rang taxonomique	Etape manuelle	Site ou référence
IMPLIQUANT L'ÉTUDE DU VOISINAGE DES GÈNES				
YGOB ^d	BLASTP	18 champignons	Oui	http://wolfe.gen.tcd.ie/ygob
SYNERGY	FASTA	9 champignons		http://www.broadinstitute.org/region/orthogroups
MultiMSOAR 2.0	BLASTP	7 mammifères		http://msoar.cs.ucr.edu/MultiMSOAR2.0

Méthodes évaluées dans l'étude de ^a (Hulsen *et al.*, 2006), ^b (Chen *et al.*, 2007), ^c (Altenhoff et Dessimoz, 2009), ^d (Salichos et Rokas, 2011), ^e (Boeckmann *et al.*, 2011) et ^f (Trachana *et al.*, 2011).

RECHERCHE DES ORTHOLOGUES PAR COMPARAISON DE SÉQUENCES

Les méthodes du premier type sont basées sur l'hypothèse que les séquences protéiques des gènes orthologues sont plus similaires entre elles qu'avec n'importe quelle autre séquence protéique. Elles considèrent donc que le fait que deux gènes soient orthologues implique que leurs séquences protéiques possèdent les meilleurs scores de similarité l'une pour l'autre. Réciproquement, elles considèrent les meilleurs scores bidirectionnels de similarité comme très probablement dû à une relation d'orthologie, suggérant ainsi une méthode très simple d'identification des orthologues.

Parmi ces méthodes, certaines s'appliquent par paires d'espèces. C'est notamment le cas des méthodes BBH et InParanoid brièvement décrites ci-dessous. D'autres méthodes travaillent sur des ensembles de plusieurs espèces et ont généralement mis au point des méthodes de regroupements des orthologues basées sur la théorie des graphes (KOG, eggNOG, OrthoDB, OrthoMCL, OMA).

MÉTHODES S'APPLIQUANT À UNE PAIRE D'ESPÈCES

Meilleur score bidirectionnel

La méthode du meilleur score bidirectionnel ou BBH, pour Best Bidirectional Hit (également nommée RBH pour Reciprocal Best Hit ou SymBeT pour Symmetrical Best hit) (Tatusov *et al.*, 1997), est la méthode la plus fréquemment utilisée pour déterminer les paires d'orthologues. Elle part du principe qu'une paire de protéines inter-espèce dans laquelle chaque protéine obtient le meilleur score de similarité avec l'autre protéine de la paire lorsqu'elle est testée contre l'ensemble des protéines de l'autre protéome constitue une paire d'orthologues. Cette définition opérationnelle de l'orthologie est généralement adéquate lorsque l'on compare deux génomes relativement proches mais atteint ses limites lorsque la distance évolutive est plus grande. En effet, par définition, la méthode du meilleur score bidirectionnel identifie uniquement les relations d'orthologie de type un-un. Or, lorsque la distance évolutive est plus grande, des duplications peuvent s'être produites dans l'une ou l'autre espèce après l'événement de spéciation. Dans ces cas, une relation d'orthologie de type un-plusieurs, voire plusieurs-plusieurs serait nécessaire pour décrire les relations d'orthologie. La méthode du meilleur score bidirectionnel ne pourra donc pas identifier l'ensemble des relations d'orthologie existantes.

Pour éviter cet écueil, des méthodes analysant l'ensemble des gènes se situant dans une plage de scores BLAST ont été mises au point. C'est notamment le cas de la méthode InParanoid, présentée ci-dessous.

InParanoid

Dans la méthode InParanoid (Remm *et al.*, 2001), tous les scores de similarité par paire sont calculés pour les différentes paires de protéomes : A contre A, B contre B, A contre B et B contre A via BLASTP. Ceux qui dépassent un certain seuil sont retenus pour autant que les régions alignées correspondent à une proportion suffisamment grande de la longueur totale de la séquence (Ostlund *et al.*, 2010). Ensuite, les BBH sont déterminés et marqués comme orthologues potentiels. Les paires intra-espèces qui ont un meilleur score que les paires d'orthologues sont également marquées comme orthologues additionnels. Ces 'inparalogues' se voient attribuer une valeur qui indique à quel point ils sont similaires à l'orthologue principal : une valeur de 100% est attribuée à l'orthologue principal et 0% à la séquence qui a le score de similarité minimum requis pour être considérée comme inparalogue d'un groupe donné. Finalement les groupes d'orthologues et inparalogues contenant certaines protéines en commun sont résolus de manière à ce que chaque protéine ne fasse plus partie que d'un seul groupe. Chaque groupe est caractérisé par une valeur de confiance basée sur un bootstrap (pourcentage des arbres dans lesquels deux séquences sont considérées comme orthologues). Le programme fait la distinction entre les inparalogues et les outparalogues sans réaliser d'arbres phylogénétiques pour chacun des groupes.

La version 7 de la base de données InParanoid comprend 99 espèces eucaryotes et *Escherichia coli* est utilisé comme outgroup.

MÉTHODES MULTI-ESPÈCES

Afin de pouvoir étendre l'identification des relations d'orthologie à plus de deux espèces, Tatusov et ses collègues (Tatusov *et al.*, 1997) ont introduit le concept de « groupe d'orthologues et d'inparalogues » défini dans l'introduction (p 8).

Les méthodes multi-espèces utilisent toutes soit BLAST (Altschul *et al.*, 1990), Smith-Waterman (Smith et Waterman, 1981) soit FASTA (Pearson, 1990) comme algorithme pour détecter la similarité de séquence. Les méthodes de regroupement des orthologues quant à elles sont variables.

euKaryotic Orthologous Groups (KOG)

La base de données KOG (Tatusov *et al.*, 2003) est la version spécifique aux eucaryotes de la base de données COG (Tatusov *et al.*, 2000) (Clusters of Orthologous Groups of proteins). En résumé, la procédure pour construire les COGs est constituée des étapes suivantes : (i) comparaison « tout contre tout » des séquences protéiques de plusieurs génomes par BLASTP. (ii) Détection et regroupement des inparalogues évidents, c'est-à-dire les protéines appartenant à un même génome et qui sont plus similaires entre elles qu'avec aucune autre protéine d'une autre espèce. (iii) Identification de trio de gènes appartenant à trois espèces différentes dans lesquels chaque gène présente un meilleur score bidirectionnel avec les deux autres gènes du trio ; les groupes d'inparalogues identifiés à l'étape deux étant traités comme une entité unique. Ces trios de gènes forment des « triangles » mutuellement consistants des meilleurs scores de similarité entre génomes. (iv) Fusion des triangles présentant un côté commun pour former les COG de manière à regrouper les orthologues appartenant à l'ensemble des espèces. Contrairement à la base de données COG, la procédure utilisée pour créer la base de données KOG contient une étape complémentaire de curation manuelle durant laquelle une attention particulière est accordée aux protéines multi-domaines. Les COG/KOG sont annotés au niveau de la fonction. La base de données KOG comprend 7 génomes eucaryotes, tandis que COG comprend 66 génomes d'unicellulaires.

eggNOG

La première étape de la procédure d'eggNOG est une recherche de similarité tout-contre-tout réalisée par FASTA (dans la version 3) (Powell *et al.*, 2012). Ensuite, un groupement des orthologues est réalisé selon une méthode similaire à celle utilisée pour la création des COG. La base de données offre une annotation fonctionnelle réalisée de manière automatisée, qui peut être vue comme une extension des bases de données COG/KOG, réalisée sans curation manuelle.

La base de données compte actuellement 1133 organismes dont 121 eucaryotes. Pour permettre une meilleure résolution des groupes d'orthologues dans les rangs taxonomiques fréquemment étudiés, la procédure a été appliquée à différents sous-ensemble d'organismes séparément. Par exemple, au sein des eucaryotes, 14 niveaux ont été créés, l'un comprenant notamment 37 espèces de champignons (fuNOGs).

OrthoMCL

L'algorithme de la méthode OrthoMCL (Li *et al.*, 2003) commence par un BLASTP de tout contre tout, après quoi les gènes des paires inter-espèces obtenant les meilleurs scores de similarité sont marqués comme orthologues putatifs et les gènes de paires intra-espèces ayant des scores plus élevés sont marqués comme paralogues récents. Une matrice de similarité est calculée suivie par un clustering de Markov (Enright *et al.*, 2002) pour déterminer les groupes d'orthologues. La version 5 comprend 150 espèces (eucaryotes et procaryotes).

Orthologous Matrix project (OMA)

L'algorithme d'OMA (Dessimoz *et al.*, 2005) crée une matrice à deux dimensions dans laquelle chaque ligne représente un groupe d'orthologues et chaque colonne représente une espèce. Cette matrice est créée en quatre étapes : (i) récupération des génomes et vérification de leur cohérence (comparaison de l'ADN et de la séquence en acides aminées, vérification de la présence des codons start et stop, suppression des fragments de moins de 50 acides aminés et des séquences dupliquées ayant une identité de plus de 99%) et comparaison avec les données d'autres bases de données ; (ii) alignement tout-contre-tout par la méthode de Smith-Waterman et identification des BBH ; (iii) vérification des BBH pour s'assurer qu'il s'agit d'orthologues et non de paralogues par comparaison avec une troisième espèce (si deux copies dupliquées subissent une perte différenciée au sein de deux espèces, la relation de paralogie peut être confondue avec une relation d'orthologie. La comparaison avec une troisième espèce possédant encore les deux copies dupliquées permet d'éviter cette erreur); (iv) les groupes d'orthologues sont formés à partir des cliques* de BBH vérifiés, par méthode graphique.

La onzième version d'OMA (Altenhoff *et al.*, 2011) est basée sur un algorithme qui compare les gènes en fonction des distances évolutives. La dernière matrice accessible sur le site (mars 2012) compte 1109 génomes, parmi lesquels 124 eucaryotes dont 23 champignons.

* En théorie des graphes, une clique est un ensemble de sommets deux-à-deux adjacents.

RECHERCHE DES ORTHOLOGUES BASÉE SUR LA PHYLOGÉNIE

Comme l'orthologie et la paralogie sont par définition liés à des événements évolutifs (spéciation et duplication, respectivement), une approche classique pour distinguer les orthologues inclut les analyses phylogénétiques et en particulier la réconciliation d'arbres phylogénétiques.

D'un point de vue pratique, la construction et l'analyse complète des arbres pour l'ensemble des gènes des génomes séquencés est un challenge important pour le génie logiciel et est coûteux en ressources informatiques. La plupart des études génomiques emploient dès lors des simplifications en utilisant au préalable une méthode basée sur la comparaison de séquences.

TreeFam

TreeFam (Li *et al.*, 2006; Ruan *et al.*, 2008) crée des groupes à l'aide d'un algorithme de regroupement hiérarchique des similarités BLAST dans une comparaison tout-contre-tout. Ensuite des arbres phylogénétiques sont construits pour les familles de gènes obtenues en utilisant différentes approches, notamment la méthode du maximum de vraisemblance (ML) et celle du neighbour-joining (NJ). Les orthologues et les paralogues sont inférés en utilisant un algorithme de réconciliation d'arbre appelé DLI (Duplication/Loss Inference). Les événements de duplication, de spéciation et de perte de gènes sont indiqués dans les arbres phylogénétiques par comparaison avec l'arbre des espèces. Dans certains cas, l'arbre est édité par les curateurs, lorsque certaines informations, comme la fonction des gènes par exemple, suggèrent que l'arbre créé automatiquement est incorrect. Les curateurs assignent un nom et un symbole à chaque famille TreeFam, ainsi qu'une brève description de la fonction des gènes de la famille basée sur la littérature. Bien que TreeFam identifie les orthologues, son objectif principal est de définir des familles composées d'un groupe de gènes qui descendent d'un seul gène dans le dernier ancêtre commun à tous les animaux (homologues), ou qui sont apparus pour la première fois chez les animaux et de fournir les arbres phylogénétiques de ces familles. La version 8.0 (10/02/2012) contient 79 espèces d'animaux.

PANTHER

L'objectif de PANTHER (Protein Analysis THrough Evolutionary Relationships) (Thomas *et al.*, 2003) est de classer les protéines par fonction moléculaire et par processus biologiques sur la base des termes de l'ontologie des gènes (GO, Gene Ontology) (Ashburner *et al.*, 2000). Pour définir l'appartenance d'une protéine à une famille, chaque séquence

protéique de la version 7 (Mi *et al.*, 2010) a été évaluée via les modèles cachés de Markov ou HMM (Krogh *et al.*, 1994) de la version 6.1 (Mi *et al.*, 2007) et assigné à la famille avec le plus haut score. Les familles de plus de 1000 séquences sont divisées manuellement en familles plus petites. Un alignement multiple est construit pour chaque famille avec le programme MAFFT (Kato et Toh, 2008). Ensuite, un arbre phylogénétique est estimé via l'algorithme GIGA (Gene tree Inference in the Genomic Age) (Thomas, 2010) qui utilise l'arbre des espèces et les familles de gènes présumées complètes pour inférer les arbres de gènes et localiser les événements de duplications par rapport aux événements de spéciation.

En ce qui concerne les orthologues, PHANTHER identifie les orthologues de type un-un, un-plusieurs et plusieurs-plusieurs. Dans les cas d'orthologues multiples, PHANTHER identifie la relation un-un qui a divergé le moins depuis l'événement de duplication en se basant sur l'arbre phylogénétique. Cette paire de gènes représente l'équivalent le plus proche entre les organismes, c'est-à-dire la paire de gènes pour laquelle la probabilité d'avoir conservé une fonction équivalente est la plus grande.

La version actuelle (version 7) contient l'ensemble des gènes codant pour des protéines appartenant à 48 organismes dont 5 champignons.

Ces deux méthodes utilisent la technique de réconciliation d'arbre (comparaison de la topologie de l'arbre des gènes et de l'arbre des espèces) pour inférer les événements de duplication et de spéciation. Ce type d'approche a des limitations. Les algorithmes de reconstruction d'arbres phylogénétiques produisent rarement des arbres entièrement fiables. Des ambiguïtés dans l'arbre des gènes ou dans l'arbre des espèces peuvent amener des erreurs dans l'inférence des événements de spéciation et de duplication (Kuzniar *et al.*, 2008) et donc dans l'inférence des gènes orthologues. De plus, ces méthodes ne peuvent être utilisées que lorsque l'arbre des espèces est connu.

PhylomeDB

La procédure utilisée pour créer les phylomes* de PhylomeDB (Huerta-Cepas *et al.*, 2008) a été publiée pour la première fois en 2007 (Huerta-Cepas *et al.*, 2007). Celle décrite ci-dessous correspond à la version 3 (Huerta-Cepas *et al.*, 2011b). La procédure commence par une recherche de similarité de type tout-contre-tout, utilisant l'algorithme de Smith-Waterman. Les séquences sont filtrées en fonction du score d'E-value† pour

* Un phylome est défini comme un ensemble complet des phylogénies des gènes d'un génome donné.

† Probabilité d'observer par hasard au sein des banques de séquences considérées un score d'alignement supérieur au score obtenu pour la paire de séquences considérée.

former des groupes de protéines considérées comme homologues. Ensuite les séquences des groupes d'homologues sont alignées par trois programmes différents. Ces alignements sont combinés et l'alignement obtenu est utilisé pour reconstruire des arbres phylogénétiques en utilisant la méthode du neighbour-joining (NJ). La topologie obtenue par NJ est ensuite évaluée par la méthode du maximum de vraisemblance (ML) en permettant une optimisation de la longueur des branches.

A partir de l'arbre obtenu, l'algorithme utilise le recouvrement des espèces présentes dans deux embranchements pour décider si le nœud qui les relie doit être considéré comme un événement de duplication ou de spéciation. Brièvement, l'algorithme analyse l'ensemble des nœuds qui connectent une protéine d'intérêt jusqu'à la racine de l'arbre et marque un nœud comme étant un événement de duplication si des gènes d'au moins une espèce se retrouvent dans chacun de ses deux nœuds fils. Par contre, si l'ensemble des espèces présentes dans un nœud fils diffère des espèces présentes dans l'autre nœud fils, le nœud est marqué comme étant un événement de spéciation.

La dernière mise à jour du site (août 2011) comprend 19 phylomes qui couvrent un total de 808 espèces. Parmi ces 19 phylomes, 4 fournissent les arbres phylogénétiques de l'ensemble des gènes de *Saccharomyces cerevisiae* avec différents groupes d'espèces, permettant ainsi une exploration à différents niveaux taxonomiques. L'un comprend 12 espèces appartenant au genre *Saccharomyces*, un autre comprend 21 espèces du subphylum des hémiascomycètes, les deux autres contiennent respectivement 12 et 60 espèces de champignons appartenant à différents groupes taxonomiques.

L'algorithme utilisé pour différencier les nœuds correspondant à des événements de duplication de ceux correspondant à des événements de spéciation peut amener des erreurs d'identification. Prenons le cas de nos neuf espèces de levure pour exemple : si suite à la duplication génomique complète chacune des deux espèces post-WGD, perd une copie différente, le nœud qui rejoint les deux copies restantes qui sont des outparalogues est marqué comme un événement de spéciation (puisque'il ne sépare pas de gènes appartenant à une même espèce) alors qu'il s'agit d'un événement de duplication.

RECHERCHE DES ORTHOLOGUES UTILISANT LE VOISINAGE DES GÈNES

L'utilisation de données complémentaires indépendantes, comme la conservation du voisinage chromosomique est hautement informative pour l'annotation correcte des orthologues. En effet, elle permet de s'assurer que la similarité de séquence n'est pas due à une évolution convergente et offre une nouvelle source d'information pour la différenciation des orthologues par rapport aux paralogues. On notera cependant que ce type de données, hautement informatives entre espèces étroitement liées, deviennent moins performantes pour des espèces séparées par de grandes distances évolutives.

Yeast Gene Order Browser

Le navigateur Yeast Gene Order Browser (YGOB) (Byrne et Wolfe, 2005) a été créé dans le but d'étudier différents aspects de l'évolution des génomes et plus particulièrement les conséquences de la duplication génomique complète qui s'est produite chez un ancêtre de *S. cerevisiae*. Il est constitué de deux éléments: un algorithme qui peut calculer le contexte synténique de chaque gène et l'évaluer et un navigateur visuel qui crée une image du contexte synténique autour d'un gène d'intérêt (Figure 7). Les deux structures clés du YGOB sont appelées « piliers » (pillars) et « pistes » (tracks). Le navigateur visuel est présenté comme une matrice dans laquelle chaque colonne ou pilier montre un ensemble de gènes homologues et chaque ligne ou piste représente des segments chromosomiques. Il y a une piste pour les espèces antérieures à l'événement de duplication génomique complète ayant eu lieu dans le subphylum des hémiascomycètes et deux pistes pour les espèces postérieures à cet événement. Un pilier a un emplacement prévu pour chaque gène qui peut être présent, c'est-à-dire un emplacement pour les espèces antérieures à la duplication génomique complète (WGD) et deux emplacements pour les espèces postérieures à cet événement. Chaque emplacement est soit vide, soit contient un gène qui est homologue aux autres gènes du pilier. Chaque pilier contient donc un groupe de gènes orthologues auxquels sont ajoutés les éventuels ohnologues présents parmi les espèces post-duplication génomique complète.

L'assignation des gènes aux piliers a été construite progressivement à partir de différentes sources : les annotations originales des gènes réalisées par les auteurs, des assignations automatiques basées sur des recherches de type BLASTP et plusieurs séries d'édition manuelle de l'ensemble complet de piliers réalisées au sein du laboratoire de K. Wolfe.

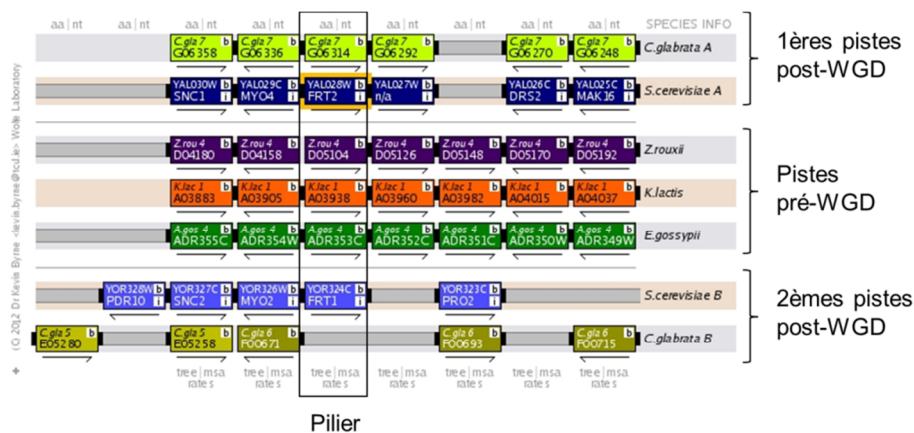


Figure 7. Illustration des piliers et pistes du navigateur visuel du YGOB.

Chaque piste représente des segments chromosomiques appartenant à une espèce. Il y a une piste par espèce, à l'exception des espèces post-WGD qui en ont deux. Chaque pilier comprend un ensemble de gènes orthologues ou ohnologues.

La version 6 associée à la publication décembre 2011 (Gordon *et al.*, 2011) comprend 18 génomes de levures : *Vanderwaltozyma polyspora*, le premier séquençage de *Saccharomyces castellii* et sa séquence complète renommée *Naumovozyma castellii*, *Candida glabrata*, *Saccharomyces bayanus*, *Saccharomyces cerevisiae*, *Zygosaccharomyces rouxii*, *Kluyveromyces lactis*, *Eremothecium gossypii*, *Lachancea kluyveri*, *Lachancea thermotolerans*, *Lachancea waltii*, *Tetrapisispora phaffii*, *Naumovozyma dairenensis*, *Torulaspora delbrueckii* ainsi qu'une première version de l'annotation (draft annotation) de *Tetrapisispora blattae*, *Kazachstania africana* and *Kazachstania naganishii*.

SYNERGY

SYNERGY (Wapinski *et al.*, 2007) est un algorithme qui utilise la similarité de séquence ainsi qu'un arbre phylogénétique des espèces pour reconstruire l'histoire évolutive des gènes dans un contexte multi-espèces et identifier les relations d'orthologie et de paralogie.

L'assignation des orthologues repose sur un pré-calcul des distances entre les gènes. La procédure commence par un alignement FASTA de tout-contre-tout. Les paires de gènes significativement similaires sont désignées sur base d'une combinaison de critères incluant l'E-value, le fait qu'une des protéines soit le meilleur score de l'autre et le pourcentage d'identité. Pour chaque paire de gènes ainsi identifiée, une distance évolutive est estimée. Celle-ci tient compte de la séquence en acides aminés (maximisation de la vraisemblance de substitutions de chaque position alignée dans l'alignement global des deux protéines) et de l'organisation chromosomique des génomes

(fraction des gènes voisins qui sont orthologues d'après une assignation préliminaire de gènes orthologues). Un graphe de similarité des gènes est construit à partir de ces données.

La reconstruction de l'arbre phylogénétique des gènes est réalisée étape par étape, des feuilles de l'arbre vers sa racine. A chaque nœud, SYNERGY commence par rassembler l'ensemble des gènes qui partagent une similarité de séquence significative. Il reconstruit ensuite un arbre phylogénétique pour chacun de ces groupes intermédiaires d'orthologues et utilise cet arbre pour partitionner les groupes de telle sorte que chaque groupe ne contienne que des gènes qui descendent d'un seul gène hypothétique dans l'espèce ancestrale correspondant à l'étape actuelle. Lorsque SYNERGY atteint la racine de l'arbre des espèces, une partition complète des groupes d'orthologues est réalisée ainsi qu'un arbre des gènes pour chacun de ces groupes.

L'algorithme a été appliqué à neuf génomes d'ascomycètes.

MultiMSOAR 2.0

MultiMSOAR 2.0 (Shi *et al.*, 2011) combine les informations de similarité de séquence et l'ordre des gènes pour identifier les orthologues de type un-un. Le système commence par construire des familles de gènes pour l'ensemble des génomes en utilisant la méthode de clustering MCL (Enright *et al.*, 2002) sur la base de la similarité de séquence déterminée via BLASTP. Ensuite MSOAR 2.0 (Shi *et al.*, 2010) est utilisé pour identifier les relations d'orthologie entre chaque paires de génomes. Ce programme tente de reconstruire l'histoire évolutive des gènes pour chaque paire de génomes en terme de réarrangements et d'événements de duplication et tente de minimiser la distance RD (nombre d'événements de réarrangements et de duplication nécessaires pour passer de l'environnement chromosomique d'un gène à l'environnement chromosomique de l'autre gène, voir Figure 8) par le principe de parcimonie.

Les paires d'orthologues sont ensuite combinées en prenant en compte la phylogénie des espèces pour former des ensembles d'orthologues contenant au maximum un gène par espèce. En plus d'identifier les groupes d'orthologues, MultiMSOAR 2.0 donne également des informations sur l'apparition, la duplication ou la perte de gènes dans l'évolution. Le programme a été utilisé sur 7 génomes de mammifères.

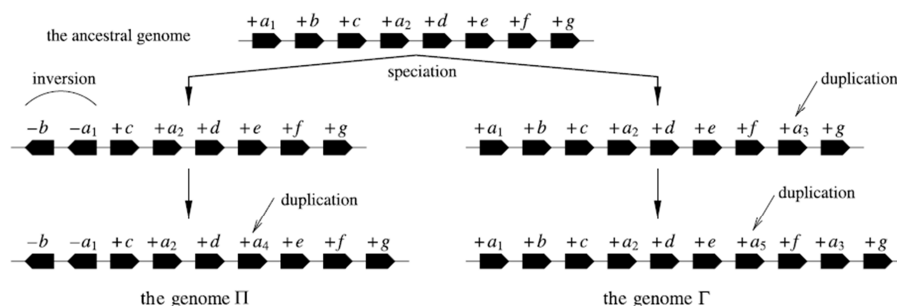


Figure 8. Histoire évolutive de deux génomes Π et Γ (Fu *et al.*, 2007).

Π a évolué à partir du génome ancestral par une inversion et une duplication génique, et Γ par deux duplications.

Le potentiel d'utilisation de ce type d'information pour identifier les orthologues est limité aux génomes dont l'événement de spéciation est relativement récent ou du moins pour lesquels les remaniements chromosomiques sont limités. En effet, il y a généralement peu de conservation de l'ordre des gènes au sein de génomes dont l'évolution a amené une divergence des gènes orthologues telle que leur pourcentage d'identité en acides aminés n'est plus que de 50% (Huynen et Bork, 1998).

Chez les levures, une étude de l'évolution de l'ordre de gènes réalisée entre *S. cerevisiae* et *S. bayanus* var. *uvvarum* montre que l'évolution est essentiellement due à des duplications géniques sur des chromosomes différents suivies par une perte différentielle des copies répétées, généralement liée à l'accumulation de nombreuses mutations (Fischer *et al.*, 2001).

COMPARAISON ET BENCHMARKING DES MÉTHODES

Les méthodes d'identification ou bases de données d'orthologues ont été développées pour rencontrer des besoins individuels. Elles analysent différents jeux de données, optimisent des critères différents et emploient des stratégies différentes de détermination de l'orthologie, ce qui rend leur comparaison difficile.

Dans une correspondance publiée suite au meeting « Quest for Orthologs » tenu à Hinxton au Royaume-Uni en 2009, (Gabaldón *et al.*, 2009) identifient au moins trois raisons qui rendent difficile l'évaluation des méthodes de prédiction des orthologues : premièrement, l'orthologie est définie en fonction de l'histoire des gènes qui est en grande partie inconnue, et ne peut donc être testée que de manière indirecte ; deuxièmement, l'arbitrage

optimal entre spécificité (1 - pourcentage de faux positifs) et sensibilité (1 - pourcentage de faux négatifs) dépend de l'objectif visé ; et troisièmement le manque de jeux de données standardisés pour tester les méthodes, ainsi que le manque d'un format de données standardisé sont des obstacles pratiques considérables pour comparer les méthodes.

ÉVALUATION DES MÉTHODES

En relation avec le premier point soulevé, à savoir que l'orthologie ne peut être testée que de manière indirecte, il n'y a actuellement pas de consensus sur les méthodes d'évaluations à utiliser. Les études évaluant les méthodes d'identification des orthologues (benchmarking) utilisent généralement plusieurs tests appartenant à un ou plusieurs des quatre types d'évaluation présentés ci-dessous.

Évaluation fonctionnelle

La recherche des orthologues étant fréquemment associée à un souhait d'annotation fonctionnelle des gènes, certaines études (Hulsen *et al.*, 2006; Altenhoff et Dessimoz, 2009; Fang *et al.*, 2010) proposent d'évaluer les groupes d'orthologues en se basant sur leur fonction biologique. Cette manière de procéder implique de faire l'hypothèse que les gènes orthologues ont un plus haut degré de conservation de la fonction que les gènes paralogues car dans le cas contraire, un groupe de gènes homologues aurait une aussi bonne évaluation qu'un groupe d'orthologues. De plus la tâche est ardue car (1) la fonction d'une protéine ne peut être véritablement connue que par des études biochimiques et/ou de leur structure. Or il est impossible d'étudier expérimentalement chaque gène de chaque espèce; (2) comme souligné par Altenhoff et Dessimoz (2009), il n'existe actuellement pas de définition universellement applicable et sans équivoque de « fonction conservée », c'est-à-dire que la similarité requise en terme de régulation, d'activité chimique, d'interactions avec d'autres gènes, etc. pour que deux gènes soient considérés comme orthologues varie souvent d'une étude à l'autre.

Néanmoins, ces études (Hulsen *et al.*, 2006; Altenhoff et Dessimoz, 2009; Fang *et al.*, 2010) proposent d'utiliser des données génomiques, telles que les profils d'expression génique, les interactions entre protéines, les termes de l'ontologie des gènes (GO terms) ou encore le nombre de la catégorie d'enzyme (EC numbers) pour évaluer les méthodes de prédictions des orthologues. Il est à noter que ce type d'évaluation peut être biaisé par une dépendance circulaire, en effet bon nombre de termes de l'ontologie des gènes, par exemple, sont attribués sur base d'alignements de séquences protéiques (Altenhoff et Dessimoz, 2009; Fang *et al.*, 2010).

La divergence fonctionnelle entre orthologues et paralogues (sub-/néo-fonctionnalisation des paralogues) ou l'altération de la fonction sur les longues distances évolutives suggèrent que ce type d'évaluation pourrait être biaisé en faveur des familles conservées ou ne possédant qu'une seule copie dans chaque espèce et serait moins approprié pour les grandes familles fortement diversifiées (Trachana *et al.*, 2011).

Évaluation par la phylogénie

Au sein d'espèces où la phylogénie est acceptée, les arbres phylogénétiques produits à partir des gènes identifiés comme orthologues par les différentes méthodes permettent d'évaluer ces méthodes en comparant la congruence* des arbres obtenus avec l'arbre des espèces, c'est-à-dire en comptant la fraction des embranchements corrects (Altenhoff et Dessimoz, 2009). Il est à noter que ce test permet de mettre en évidence la présence de faux positifs, mais ne donne aucune information relative aux faux négatifs.

Évaluation basée sur l'ordre des gènes

La conservation de l'arrangement des gènes sur les chromosomes est utilisée dans certaines études de comparaison de méthodes d'identification (Notebaart *et al.*, 2005; Hulsén *et al.*, 2006; van der Heijden *et al.*, 2007; Altenhoff et Dessimoz, 2009) pour valider l'assignation fonctionnelle des orthologues. Il est à noter que la conservation des gènes voisins est un indicateur fiable de la présence d'une relation d'homologie, mais n'est pas le seul fait des gènes orthologues puisque les paralogues issus d'une duplication segmentale, chromosomique ou génomique peuvent également présenter une conservation de leur voisinage chromosomique immédiat.

Évaluation par analyse des classes latentes

La comparaison des résultats de différentes alternatives méthodologiques contient des informations utiles. En effet, la concordance des groupes d'orthologues obtenus par différentes méthodes augmente la confiance en ces groupes (pour autant que les méthodes employées soient indépendantes), tandis que leur désaccord indique des erreurs possibles (soit des faux positifs, soit des faux négatifs). L'analyse des classes latentes ou LCA (pour Latent Class Analysis) est une technique statistique qui exploite les informations issues de la comparaison des résultats de différentes méthodes pour inférer la sensibilité et la spécificité de différentes méthodes en l'absence de « gold standard ». Elle a été utilisée pour comparer sept méthodes d'identification des orthologues (Chen *et al.*, 2007) basées sur la

* En phylogénie, on parle de congruence quand deux arbres ne se contredisent pas en termes de structure.

comparaison de séquences ou sur la phylogénie. Il en ressort que les méthodes d'identification des orthologues de cette étude basées sur la phylogénie (voir table 1^b) ainsi que la méthode du meilleur score bidirectionnel montrent de faibles taux de faux positifs mais des taux élevés de faux négatifs. Dans cette étude, les méthodes qui se classent le mieux sont InParanoid et OrthoMCL, qui ont toutes deux des taux de faux positifs (FP) et faux négatifs (FN) inférieurs à 20%.

Certains reprochent cependant à ce type d'évaluation que la cohérence de prédiction de différentes méthodes n'offre pas de mesure statistique de la véracité de ces prédictions, comme le permet une comparaison avec un gold standard, et n'a pas de signification biologique non plus (Hulsen *et al.*, 2006).

Il est à noter qu'actuellement aucun des articles évaluant les méthodes d'identification des orthologues n'a évalué les méthodes basées sur le voisinage des gènes. Une des études utilise YGOB (Salichos et Rokas, 2011), cependant la base de données du YGOB n'y est pas évaluée mais bien utilisée comme gold standard pour évaluer d'autres méthodes.

ELÉMENTS DÉTERMINANT LE CHOIX D'UNE MÉTHODE

Bien que les résultats de comparaison puissent différer selon les paramètres utilisés dans les différents algorithmes ou les méthodes d'évaluation utilisées, de manière générale, il ressort des études comparatives que le choix de la méthode ou base de données à utiliser pour identifier les orthologues est dépendant de quatre choses.

Premièrement, la méthode à utiliser dépend de la question à laquelle on tente de répondre en utilisant les relations d'orthologie (Kuzniar *et al.*, 2008). En effet, l'importance de la spécificité et de la sensibilité de la méthode sera différente selon que l'on cherche à identifier un équivalent fonctionnel impliqué dans un processus biologique particulier ou selon que l'on cherche à étudier les mécanismes fondamentaux à la base de l'évolution des génomes, par exemple. Suite à la comparaison de six méthodes d'identification des orthologues réalisée par Hulsen *et al.* (2006), les auteurs concluent que si l'objectif est d'avoir un ou plusieurs orthologues pour un grand nombre de protéines, une des méthodes qui permet des relations de type plusieurs-plusieurs (comme InParanoid par exemple) doit être utilisée. Par contre si la sélectivité (avoir le moins de faux positifs possible) est plus importante que la sensibilité (avoir le moins de faux négatifs), et qu'avoir un orthologue est suffisant, alors la méthode du meilleur score bidirectionnel donnera les meilleurs résultats (Salichos et Rokas, 2011).

Deuxièmement, la distance évolutive qui sépare les espèces à analyser. Selon le type d'évaluation choisi, le classement des différentes méthodes (et le choix qui en découle) peut également être influencé par la distance évolutive qui sépare les espèces à analyser. Par exemple, la conservation des caractéristiques fonctionnelles et de l'ordre des gènes est significativement plus faible lors d'une comparaison entre les orthologues de l'homme (*Homo sapiens*) et du vers (*Caenorhabditis elegans*) que lors d'une comparaison entre les orthologues de l'homme et de la souris (*Mus musculus*), qui sont des espèces plus proches (Hulsen *et al.*, 2006). Dans cette étude comparative, qui n'inclut aucune méthode basée sur le voisinage des gènes (Hulsen *et al.*, 2006), la méthode du BBH est la mieux évaluée sur la base de la conservation de l'ordre des gènes entre les espèces proches (Homme – souris) alors qu'elle est largement dépassée par InParanoïd lors d'une comparaison entre espèces plus éloignées (Homme – vers). Les méthodes basées sur le voisinage des gènes sont plus efficaces quand les distances évolutives entre espèces sont courtes.

Troisièmement, la qualité de l'annotation des génomes et la couverture en terme d'espèces. La qualité de l'annotation des génomes des espèces incluses dans une étude génomique ou phylogénétique affecte le résultat de cette étude (Milinkovitch *et al.*, 2010). Par exemple, dans une comparaison des groupes d'orthologues produits par deux versions d'annotation génomique différentes réalisée par Trachana *et al.* (2011), presque la moitié des erreurs résultent d'artéfacts d'annotation génomique (nouveaux gènes, gènes supprimés ou caractérisés comme pseudogènes). Cette étude montre également que le nombre de faux négatifs diminue lorsque la couverture en terme d'espèces augmente.

Quatrièmement, au-delà des aspects « théoriques » d'exactitude de prédiction, le choix de la méthode est également largement influencé par des aspects pratiques. Par exemple, bon nombre des méthodes décrites ci-dessus donnent accès aux bases de données d'orthologues qu'elles ont produites, mais pas à l'algorithme ou aux programmes informatiques qui ont permis de les produire (Alexeyenko *et al.*, 2006). Dès lors le choix peut être influencé par la disponibilité de l'algorithme, si l'on souhaite travailler sur ses propres données par exemple, ou par la disponibilité des données, par le panel de génomes analysés, la fréquence des mises à jour, voire la facilité d'utilisation de l'interface web, si on travaille avec les prédictions réalisées dans les bases de données.

IV. IDENTIFICATION DES ORTHOLOGUES PAR VOISINAGE ET SIMILARITÉ

GENÈSE ET OBJECTIFS DE LA MÉTHODE

Lors de mon arrivée au sein de l'équipe GENA, en 2005, Julie Diffels travaillait sur l'analyse des transporteurs transmembranaires chez les levures et avait entrepris de réaliser une étude de la synténie¹ à l'échelle locale des gènes en analysant manuellement le voisinage chromosomique de l'ensemble des gènes de quelques familles de transporteurs, notamment la famille du transporteur de carboxylate, Jen1 (Lodi *et al.*, 2007). Cette synténie à l'échelle locale était représentée de manière visuelle, en suivant le principe du navigateur YGOB.

La méthode IONS (Identification of Orthologs by Neighbourhood and Similarity) correspond à une version généralisée et automatisée de l'étude des voisinages chromosomiques réalisée manuellement par Julie Diffels.

APPLICATIONS SOUHAITÉES À L'ORIGINE DE LA CRÉATION DE LA MÉTHODE

L'idée de créer un algorithme permettant d'automatiser l'étude des voisinages chromosomiques provient de trois demandes particulières :

Premièrement, le consortium Génolevures, était en attente d'une méthode permettant l'identification des gènes orthologues qui partagent un voisinage chromosomique semblable, ou orthologues positionnels (c'est-à-dire les gènes orthologues dont la position au niveau du génome reflète le mieux la position génomique de leur ancêtre commun le plus récent), afin de les utiliser comme point d'ancrage pour une étude de la synténie à l'échelle du génome sur les différentes espèces actuellement séquencées et à venir.

Deuxièmement, au sein du laboratoire, nous étions intéressés par une méthode permettant d'automatiser l'étude des voisinages chromosomiques afin d'étudier plus aisément l'histoire évolutive des familles de transporteurs.

¹ Le terme de synténie est utilisé ici par souci de correspondance avec le terme utilisé dans l'article de (Lodi *et al.*, 2007), bien qu'il ne soit pas question de l'ordre des gènes (Voir définition de la synténie au chapitre suivant).

Troisièmement, un des objectifs de ma thèse était de réaliser une mise à jour de la base de données YETI (De Hertogh *et al.*, 2006) afin d'y inclure les nouvelles espèces séquencées par le consortium Génolevures et d'étudier l'évolution de la taille des familles de transporteurs. Cette mise à jour nécessitait un transfert de l'annotation fonctionnelle, d'où l'intérêt pour une méthode d'identification des homologues et orthologues.

Ces différents intérêts concoururent à nous pousser à développer notre propre méthode d'identification des orthologues. En effet, à l'époque, des méthodes d'identification des orthologues telles qu'InParanoid existaient, des méthodes d'identification de la synténie à l'échelle du génome également (ADHoRe (Vandepoele *et al.*, 2002) par exemple), mais aucune ne permettait de répondre à l'ensemble de nos exigences (voir cahier des charges, p 49). Un outil permettant d'obtenir le type de résultats présentés sur le navigateur YGOB aurait satisfait chacun, malheureusement cet outil n'existait pas, les bases de données d'orthologues et d'ohnologues du YGOB étant fortement basées sur une curation manuelle.

L'algorithme SYNERGY a été publié en 2007, à l'époque une première version d'IONS existait déjà (Seret, 2007), mais la méthode n'était pas encore publiée dans une revue internationale. SYNERGY utilise, entre autre, le critère de conservation du voisinage chromosomique et présente l'avantage par rapport à YGOB d'être automatisée. Cependant, il présente l'inconvénient d'utiliser des paramètres dépendant du jeu de données (et ne produit pas de résultats visuels), laissant ici un créneau ouvert pour notre méthode.

CHOIX DES CRITÈRES D'INFÉRENCE

La méthode IONS est basée sur deux critères : la similarité de séquence et la conservation du voisinage chromosomique.

Le critère de similarité de séquences est le plus fréquemment utilisé pour inférer l'homologie. Cependant, utilisé seul, il peut donner de fausses indications d'homologie lorsqu'il y a eu une évolution convergente et n'est donc pas suffisant.

L'ajout du critère de conservation du voisinage chromosomique permet de pallier à ce problème. Il a été choisi en fonction des applications souhaitées de la méthode.

DÉFINITION DES OBJECTIFS

Deux objectifs principaux ont été définis pour la méthode:

- Identifier les relations d'homologie entre les gènes codants.

- Identifier les groupes d'orthologues positionnels composés au maximum d'un gène par espèce.

Les groupes de gènes orthologues composés au maximum d'un gène par espèce sont de deux types.

Premièrement, les orthologues positionnels de type un-un pour lesquels aucun des orthologues du groupe ne possède un paralogue. Ce type de groupe d'orthologues a un intérêt particulier pour l'annotation fonctionnelle, car il présente le plus faible risque de modification de fonction.

Deuxièmement, les groupes d'orthologues positionnels dont certains membres possèdent une ou des copies paralogues. Ces orthologues sont pertinents pour les études synténiques.

Le concept d'orthologie positionnelle est important à cause des conséquences fonctionnelles du contexte génomique. Les gènes proches l'un de l'autre ont plus de chance d'interagir et l'expression génique peut être significativement affectée par la position au niveau du génome (Dewey, 2011).

CAHIER DES CHARGES

A côté de ces objectifs principaux, la méthode devait également répondre à différentes exigences :

1. Etre adaptée pour travailler sur plusieurs espèces en même temps, ce qui implique notamment que la méthode soit automatisée ;
2. Permettre l'identification des gènes orthologues entre espèces séparées par une distance évolutive relativement grande afin de pouvoir couvrir les espèces du subphylum des hémiascomycètes;
3. Assurer une cohérence des résultats lors de l'ajout de nouvelles espèces, ce qui implique de ne pas utiliser de paramètre directement dépendant des espèces analysées comme un arbre phylogénétique par exemple ;
4. Produire une sortie visuelle représentant les relations d'homologie entre les voisinages des groupes de gènes orthologues ou homologues identifiés ;
5. Produire un fichier de sortie dénombrant le nombre de gènes présents par espèce dans chacun des groupes d'homologues et d'orthologues formés.

LA MÉTHODE IONS

En dehors des objectifs et du cahier des charges et au vu des séquences génomiques à traiter, le choix a été fait de construire une méthode pour des séquences génomiques haploïdes.

La méthode d'identification des orthologues par voisinage et similarité ou IONS (pour « Identification of Orthologs by neighbourhood and Similarity ») est présentée dans le chapitre suivant sous la forme d'un article publié dans *Evolutionary Bioinformatics* en 2011, vol 7 : 123-133.

Mes contributions personnelles à cet article :

- *Développement du principe de base de la méthode en collaboration avec Julie Diffels*
- *Programmation en Perl de l'algorithme, avec l'aide ponctuelle de Julie Diffels et de Rémy Vandaele*
- *Test sur les hémiascomycètes et comparaison avec d'autres méthodes*
- *Rédaction*

**IONS: IDENTIFICATION OF ORTHOLOGS BY NEIGHBORHOOD
AND SIMILARITY**

-

**AN AUTOMATED METHOD TO IDENTIFY ORTHOLOGS IN
CHROMOSOMAL REGIONS OF COMMON EVOLUTIONARY
ANCESTRY AND ITS APPLICATION
TO HEMIASCOMYCETOUS YEASTS**

Marie-Line Seret and Philippe V. Baret

Université Catholique de Louvain, Earth and Life Institute (ELI), 1348
Louvain-la-Neuve, Belgium.

Corresponding author email: philippe.baret@uclouvain.be

Evolutionary Bioinformatics 2011:7 123-133

doi: 10.4137/EBO.S7465

This article is available from <http://www.la-press.com>.

© the author(s), publisher and licensee Libertas Academica Ltd.

This is an open access article. Unrestricted non-commercial use is permitted
provided the original work is properly cited.

ABSTRACT

Comparative sequence analysis is widely used to infer gene function and study genome evolution and requires proper ortholog identification across different genomes. We have developed a program for the Identification of Orthologs in one-to-one relationship by Neighborhood and Similarity (IONS) between closely related species. The algorithm combines two levels of evidence to determine co-ancestrality at the genome scale: sequence similarity and shared neighborhood. The method was initially designed to provide anchor points for syntenic blocks within the Génolevures project concerning nine hemiascomycetous yeasts (about 50,000 genes) and is applicable to different input databases. Comparison based on use of a Rand index shows that the results are highly consistent with the pillars of the Yeast

Gene Order Browser, a manually curated database. Compared with SYNERGY, another algorithm reporting homology relationships, our method's main advantages are its automation and the absence of dataset-dependent parameters, facilitating consistent integration of newly released genomes.

KEYWORDS

ortholog, synteny, shared neighborhood, hemiascomycete, yeast

INTRODUCTION

Given the increasing number of large-scale sequencing projects (see <http://www.ncbi.nlm.nih.gov/Genomes>), comparative genomic approaches are now widely used.¹⁻⁷ Indeed, comparison of genome sequences across species offers great potential for studying many aspects of their underlying biology, such as the prediction of gene function. Moreover, it provides insights into the processes of both genome and gene evolution. The reliable identification of orthologs in a one-to-one relationship is critical for many comparative genomics analysis, such as the construction of syntenic blocks,^{1,5,8} the reconstruction of accurate species or gene trees, or the automation of the functional annotation of genes.

Orthology is often interpreted as the functional equivalence of proteins across species, while in fact it defines only a particular relationship of homology in which two genes originating from a common single ancestral gene diverged following a speciation event.^{9,10} However, orthologs are more likely to have a functional similarity than inter-species paralogous genes.¹¹

In the process of identifying orthologs, the trend is to assume that if two sequences are significantly similar, they must be homologous; i.e., they must share a common origin.^{9,10} However, similarity may be a false indication of homology, for example, in cases of convergence and events of duplication and loss that tend to blur the tracing of co-ancestrality.¹ Therefore, the identification of orthologs among the set of homologs defined by similarity requires more specific analysis.

Most approaches for the identification of orthologs may be based on the following evidence: sequence similarity, reconciliation of genes and species phylogenies, and synteny conservation (see ¹¹⁻¹⁶). The implementation is either manual, semi-automatic (some parameters are defined *a priori* and differ from one dataset to the other), or automatic with constant parameters. To cope easily with the availability of new genomes, the challenge is to

develop a simple automated method in which parameters are not modified by the addition of new information.

We developed a program called IONS (Identification of Orthologs by Neighborhood and Similarity). This program relies on two types of evidence: sequence similarity at the protein level and the chromosomal neighborhood (see Algorithm). The method was initially developed for the Génolevures project, a large-scale comparative genomics project for *Saccharomyces cerevisiae* and other yeast species representative of the various branches of the hemiascomycetous class, which notably provides annotated sequence data and classifications of nine complete Hemiascomycete yeast genomes comprising a total of about 50,000 genes. In this context, it was used to identify subsets* of orthologs used as anchor points for the construction of syntenic blocks.^{1,17} In practice, blocks of conserved synteny are delineated as regions containing numerous orthologous genes that can be separated by a limited number of intervening genes (non-orthologous genes).¹⁸ This preliminary version of the method was also used to study the evolution of families of transporters.^{19,20} The purpose of the present paper is to give a full description of the final version of the method and to discuss its advantages in comparison with two other methods, the Yeast Gene Order Browser (YGOB) and SYNERGY.

Comparisons required application of the different methods to a common dataset, in our case the hemiascomycete phylum. The adjusted Rand index (ARI)²¹ was used as a quantitative indicator of the equivalence of partitioning for pairwise comparisons of methods.

MATERIAL AND METHODS

ALGORITHM

Input data and preparation

The input required for IONS consists of a database of genes belonging to a number of species encompassing a particular taxonomic class or phylum. This database must contain comprehensive information about the relative position of genes as well as assignment to a group of similar gene products or to any other group of putative homologs obtained by any other method (Additional file 1: Sample of input and output files).

* The terms "cluster" and "subset" are indifferently used to describe sets of query genes grouped by the method.

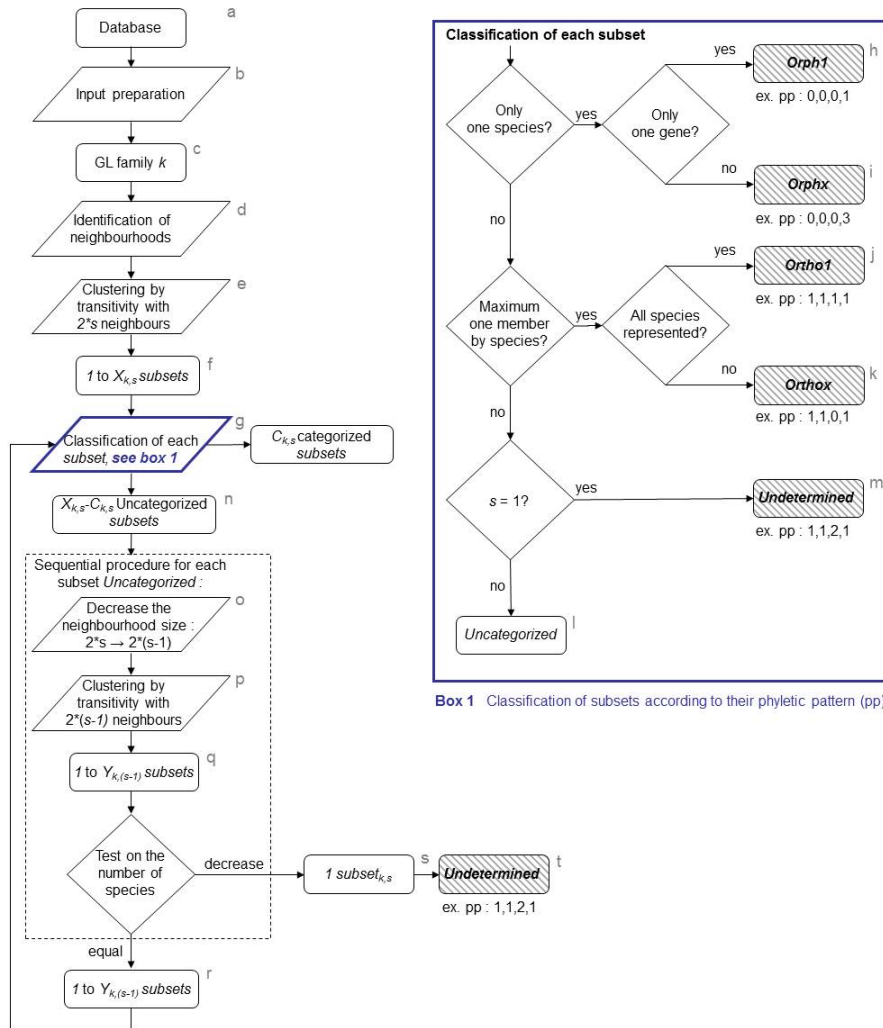


Figure 1. Flowchart of the IONS algorithm.

The algorithm is automatically applied to each GL family* k . The first clustering by transitivity, with s neighbors taken into account for each side of the query genes, produces 1 to X subsets with the genes belonging to the family k . These subsets $_{k,s}$ are classified into different categories: TSing, InPara, Ortho1, Orthox, and Undet. Subsets that do not correspond to one of these categories are uncategorized and enter the sequential procedure. In the sequential procedure, a test is made to ensure that the new subsets created with the narrow neighborhood (Y subsets $_{k,(s-1)}$) do not result in a reduction of the number of species. A comparison is made between the number of species represented in the subset with the wide neighborhood subset $_{k,s}$ and the maximum number of species represented in the 1 to Y subsets $_{k,(s-1)}$ obtained with the narrow neighborhoods. If the number of species

* GL family is used as an abbreviation of Génolevures family.

is equal, the 1 to Y subsets $_{k,(s-1)}$ obtained with the narrow neighborhood enter the classification. Otherwise, the subset with the wide neighborhood, subset $_{k,s}$, is validated and labeled as “Undet.” Striped frames indicate the end of the analysis for the genes belonging to the labeled subsets. Below these frames is an example of a possible phyletic pattern (pp).

For this study, the database (Fig. 1a) consisted of the genomes of nine species* covering the Hemiascomycetes class and was based on the assignment to Génolevures families (GL Family) defined by Nikolski *et al.*²² These families of similar genes were based on the calculation of pairwise similarities of sequences provided by BLAST and Smith-Waterman and a subsequent clustering. An algorithm was then applied to construct consensus families from competing clustering computations by an election method (see²² for details).

Figure 1 presents a flowchart of the algorithm we developed to infer orthology from the combination of similarity (family) and shared neighborhood, defined as the preserved co-localization of some genes on chromosomes of different species (independently of their order). First, the database was read to identify the list of genes belonging to each family (Fig. 1b). Then all families (Fig. 1c) were sequentially analyzed as described below.

Identification of neighborhoods

For a given family k having i members, the set of s neighbor genes $N_{i,k,n}$ with n comprised between $[-s,s]$ identified on each side of a “query gene” $G_{i,k}$ defines a neighborhood of size $2*s$ genes (Fig. 1d and additional file 2: Visual representation of main steps of the IONS algorithm leading to a subset of orthologs).

Assignment of genes to subsets through a clustering by transitivity

IONS proceeds by comparisons of the neighborhoods of the i query genes belonging to a particular family k . The translation products of the $2*s*i$ neighbor genes are tested pairwise: if a neighbor of a query gene belongs to the same family as a neighbor of another query gene (ie, if these genes are similar in sequence), the neighborhoods of the two query genes have one neighbor in common.

While analyzing a given family k , two query genes G_{ik} having at least one neighbor in common are assigned to the same cluster. It is noteworthy that, if

* The choice to limit our study to the nine species of the Génolevures database available at the time is mainly due to their high coverage of sequencing and coherent annotation (see chapter II p 22)

the neighborhoods of two query genes have at least one neighbor in common with the neighborhood of a third gene, but none with each other, the three query genes are assigned to the same cluster, through transitivity (Fig. 1e and Fig. 2).

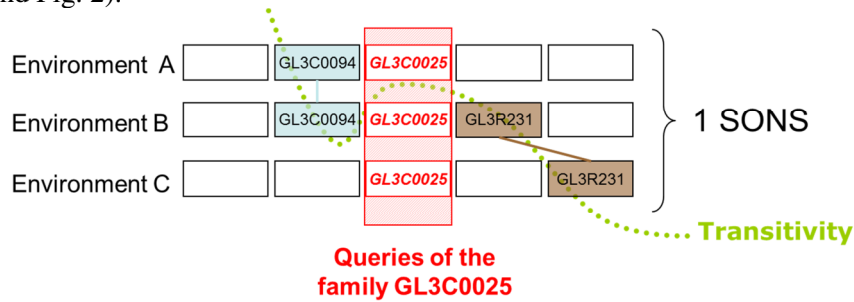


Figure 2. Clustering by transitivity.

If the environment* (A) possesses a homologous protein in (B) and if the environment (C) possesses a protein that is homologous to another protein of (B), then the environments (A, B, and C) are parts of the same subset of orthologs by neighborhood and similarity.

The program offers the opportunity to change the number of neighbors required to assign genes to the same subset of orthologs. In all cases, the output of the analysis is either a confirmation of the initial family on the basis of the neighborhood evidence or a splitting of the family into different clusters (Fig. 1f), which are sequentially numbered (e.g., GL3R1304_10010, GL3R1304_10020, etc.).

Classification of subsets (Fig. 1g and Box 1)

In a first step, the widest neighborhood size was used to calculate clusters, e.g., $s = 15$ if 15 neighbor genes were identified on each side of the query gene. The number of query genes in each species was called the phyletic pattern (pp) of a subset. According to our classification, subsets with genes in only one species were labeled as “TSing,” for technical singleton (Fig. 1h), if they comprised one gene only, and as “InPara” for in-paralogs (paralogs in a given lineage that all evolved by gene duplications that happened after the speciation event that separated the lineage under consideration from the other lineages²³) (Fig. 1i), if they comprised more than one gene. Subsets with a maximum of one gene in different species were called SONS (Subsets of Orthologs defined by Neighborhood and Similarity)¹⁹ and labeled as “Ortho1” (Fig. 1j) if one gene was present in each species (pp: 1,1,1,1) and “Orthox” (Fig. 1k) if at least one species did not have any gene assigned to this subset (pp: 1,1,1,0 e.g.,). Other subsets

* The term « environment » refers to the chromosomal environment of a gene composed of genes that flanked it.

comprising multiple genes from different species were considered as “Uncategorized” (Fig. 1l) and further assessed by a sequential procedure.

Sequential procedure

When a subset comprised genes from different species and if some species comprised more than one member (subsets labeled as “Uncategorized”), we progressively diminished the size of the neighborhood taken into account. Subsets in the “Uncategorized” category (Fig. 1n) were tested with a narrower neighborhood size, e.g., $s = 14$ for 14 genes each side (Fig. 1o). The clustering step (Fig. 1p) with the narrower neighborhood either confirmed the subset, if all neighborhoods still tied up the different homologs, or split it into a number of new subsets (Fig. 1q). Indeed, if a neighborhood was linked to the others by the 15th neighbor gene only, the subset was split when this part of the neighborhood was no longer taken into account.

If at least one of the new subsets comprised the same number of species as the initial subset in the wide neighborhood, the new subsets were validated (Fig. 1r) and a supplementary suffix was added (i.e., GL3R1304_10010_10, GL3R1304_10010_20). In turn, these new subsets were labeled as described in the previous section. “Uncategorized” subsets were recursively tested with a narrower neighborhood size, e.g., $s = 13$ (13 genes each side), and so on up to $s = 1$.

In the search of the orthologs in a one-to-one relationship, the algorithm was designed to privilege clusters of orthologs and in-paralogs rather than to split orthologs apart. Thus, if all new subsets comprised fewer species than the initial subset in the wide neighborhood, the procedure was stopped. In this case, genes were assigned to the subset defined in the previous step in terms of neighborhood width (Fig. 1s) and labeled as “Undet” for undetermined (Fig. 1t).

During the sequential procedure, if the neighborhood size (s) was equal to one neighbor, the procedure was halted and the subset considered as “Undet” (Fig. 1m). This “Undet” label means that the subset contained homologous genes for which the relationship (orthology or in-paralogy) could not be assessed using our neighborhood criteria.

A test case on Hemiascomycetes described in the next section illustrates the interest of the sequential procedure. With 15 neighbors, the IONS procedure assigned 33,258 of the 47,874 genes (see Fig. 3) to different types of subsets. Among these, 22,758 genes (47.53% of the total) formed subsets of orthologs (Ortho1 and Orthox). The sequential procedure was then applied to the remaining 14,616 genes. Additional file 3 shows the cumulative results

obtained after each step of this sequential procedure at the end of which 6,521 additional genes formed new subsets of orthologs, increasing the total percentage of genes classified into subsets of orthologs (Ortho1 and Orthox) to 61.16% of all 47,874 genes.

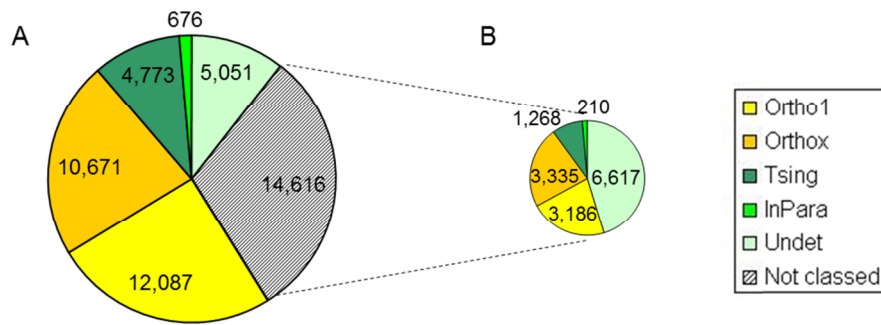


Figure 3. Improvement of the classification by the sequential procedure.

(A) Distribution of genes in the different types of subsets resulting from the analysis with 15 neighbors taken on each side of the query genes and (B) improvement of the classification thanks to the sequential procedure that allows classifying genes not classified in (A). **Abbreviations:** Ortho1, subset with one gene in each species; Orthox, subset with maximum one gene per species, all species are not represented; TSing, subset with one gene from one species only; InPara, subset with several genes belonging to the same species; Undet, subset with several species represented, at least one species represented by more than one gene.

Output

The IONS program produces two databases as well as a visual file and a neighborhoods file for each subset (Additional file 1: Sample of input and output files). The first database is ordered by gene and contains its family and the name and status of the subset to which it was assigned. The second database is organized by subsets. Each shows the family, the subset, the status of the subset, the number of genes of each species constituting this subset, and the total number of genes in the subset.

The neighborhoods file reports the presence-absence of the different families in the neighborhoods of the different queries of the subset. The visual file shows the names and families of the neighbors of genes belonging to a particular subset. The raw descriptions of neighborhood relationships available in these visual files may serve as support for ad hoc discussion of gene evolution in complex situations, like the emergence of ohnologs (duplicates arising from the Whole Genome Duplication,²⁴ abbreviated as WGD).

RESULTS

TEST CASE ON NINE HEMIASCOMYCETES

We applied the IONS method to resolve the homology relationships in the genomes of nine hemiascomycetous yeasts: *Saccharomyces cerevisiae* (SACE), *Candida glabrata* (CAGL), *Zygosaccharomyces rouxii* (ZYRO), *Saccharomyces (Lachancea) kluyveri* (SAKL), *Kluyveromyces (Lachancea) thermotolerans* (KLTH), *Kluyveromyces lactis* (KLLA), *Ashbya (Eremothecium) gossypii* (ERGO), *Debaryomyces hansenii* (DEHA), and *Yarrowia lipolytica* (YALI). These genomes add up to 47,874 proteins, which were classified into 7,927 families (see additional file 4: Results of the IONS method application to the 47,874 CDS of Génolevures), as well as 1,015 proteins being left aside because of ambiguous or complex affiliations¹ (Fig. 4).

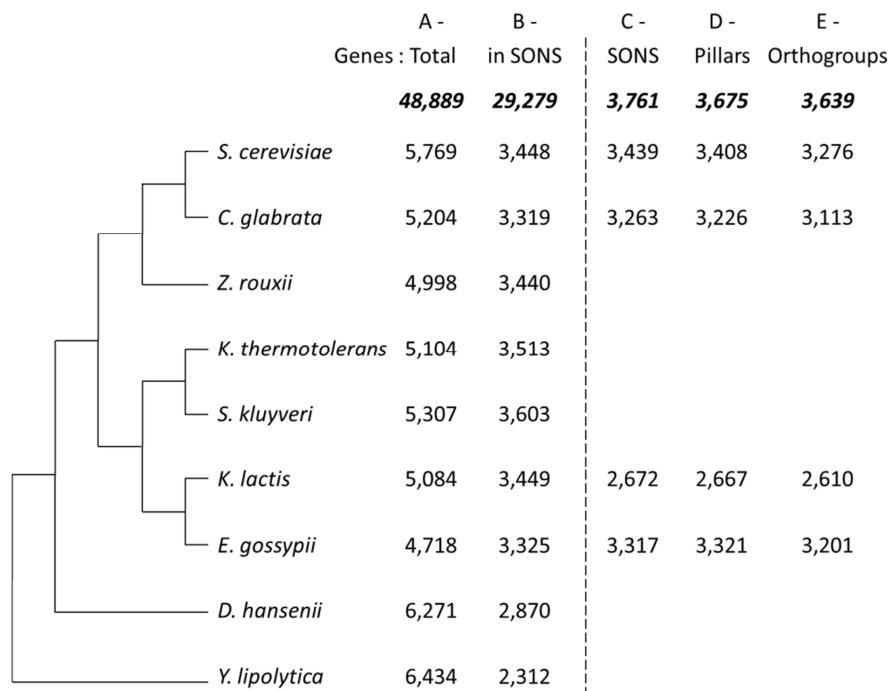


Figure 4. Cladogram of the nine hemiascomycetous yeasts.

The cladogram is based on the phylogenetic tree in Souciet *et al.*¹ The first two columns refer to the number of genes. **(A)** Repartition of the 48,889 genes in the different species. **(B)** Repartition in the different species of the 29,279 genes classified in the 4,107 subsets of orthologs (Ortho1 and Orthox) by the IONS method. The last three columns show the number of subsets obtained by the

different methods*. (C) Subsets of orthologs produced by the IONS method; (D) pillars of the YGOB; and (E) orthogroups produced by SYNERGY for the 12,691 genes involved in the comparison.

The results in Génolevures

Figure 3 shows the distribution of genes into the five categories of subsets (TSing, InPara, Ortho1, Orthox, and Undet). With a maximum of 15 neighbors taken into account, the neighborhood was informative, i.e., there was at least one neighbor in common with at least one other query gene, for 43,101 genes (90%). The method confirmed the co-ancestrality of orthologs in a one-to-one relationship of 1,309 families containing one gene in each species as well as identifying 388 new subsets of orthologs of this type (Table 1).

Table 1. Comparison of the 7,927 Génolevures families (similarity) to the subsets produced by the IONS method.

Families		IONS subsets			
		Identical to families	Not identical to families		
			Ortho1	Orthox	Others
One gene per species present in 9 species	1,689	1,309	–	397	413
	<i>15,201</i>	<i>11,781</i>		<i>3,007</i>	<i>413</i>
One gene per species present in fewer than 9 species	1,416	1,142	–	158	412
	<i>7,640</i>	<i>6,439</i>		<i>789</i>	<i>412</i>
Others	4,822*	3,800*	388	713	2,624
	<i>25,033</i>	<i>7,554</i>	<i>3,492</i>	<i>3,771</i>	<i>10,216</i>
Total	7,927	6,251	388	1,268	3,449
	<i>47,874</i>	<i>25,774</i>	<i>3,492</i>	<i>7,567</i>	<i>11,041</i>

Notes: The families that are fully confirmed by the neighborhood analysis are shown in the second column while columns three to five show the subdivision into different types of subsets (Ortho1, Orthox, and Others) of the families that could not be fully confirmed by the neighborhood analysis.

Numbers of subsets are in bold. Numbers of genes are in italics. *Among the 4,822 families in “Others,” there are 3,343 families containing only one gene in one species. The subset in IONS is identical.

Another 1,142 families of orthologs with a maximum of one member in each species were confirmed by shared neighborhood, while 1,268 new subsets of orthologs of this type were created (Table 1). In total, the number of genes classified into subsets of 2 to 9 orthologs (Table 2) increased from 47.71% (22,841 genes, Table 1), as inferred from the families based on similarity, to 61.16% (29,279 genes; see Tables 1 and 2 for details) with the IONS method.

* These last three columns are based on a common dataset composed of 12,691 genes for which we had no conflict with correspondence of names and that were classified into subsets of orthologs according to IONS.

Table 2. Types of subsets produced by the IONS method for the nine *Génolevures* species.

Type of subset	Number of subsets	Number of genes	Percentage of genes	Mean subset size*	Median subset size	Mean number of species*
Ortho1	1,697	15,273	31.9	9 ± 0.00	9	9 ± 0.00
Orthox	2,410	14,006	29.26	5.81 ± 2.32	7	5.81 ± 2.32
Subtotal	4,107	29,279	61.16			
TSing	6,041	6,041	12.6	1 ± 0.00	1	1 ± 0.00
InPara	272	886	1.85	3.26 ± 3.70	2	1 ± 0.00
Undet	936	11,668	24.37	12.47 ± 7.79	10	7.99 ± 1.78
Total	11,356	47,874	100			

Note: *Mean and standard deviation.

Interests

The subsets of orthologs (Ortho1 and Orthox) produced by the IONS method have already proven to be of particular interest to serve as anchor points for the construction of syntenic blocks¹ and to identify orthologs.¹⁷ Small differences between the results presented in ¹ and in this paper result from an improvement in the IONS method by the inclusion of the sequential procedure that progressively decreases the neighborhood by one neighbor at a time rather than the rough iteration made for 15, 10, 5, and 1 neighbors used in the previous version (see additional file 4 for the correspondence between the two classifications). This modification allows identification of more subsets of the Ortho1 type.

The method, especially the Undetermined subsets, is also useful as a starting point for manual dissection of a functional family. For example, SONS was used to suggest a model for the evolution of the hexose transporters and glucose sensors.²⁰ The IONS method was also used for analysis of the evolution of the ATP-binding cassette transporters conferring multiple drug resistance in hemiascomycetous yeasts¹⁹ and of the drug:H⁺ antiporter 1 family.²⁵

COMPARISON TO OTHER DATASETS OF ORTHOLOGS

The pillars of the YGOB and the orthogroups produced by SYNERGY

To assess the quality of subsets of orthologs (Ortho1 and Orthox) produced by the IONS method, we compared it to two different studies. The first comparison was done with a manually curated database that we assume to be the ‘gold standard’ for yeast genomes: the pillars of YGOB (v5, January 2011).²⁶ The second comparison concerned the orthogroups produced by SYNERGY,²⁷ an algorithm that reports orthology relationships using sequence similarity, synteny, and a given species phylogeny to reconstruct the underlying evolutionary history of genes. This method is automated but, in contrast to IONS, requires an *a priori* weighting of different parameters:

protein similarity (α), synteny similarity (β), and probability of duplication and losses when rooting a gene tree (γ). Other methods exist but either have not been applied to the Hemiascomycetes phylum or do not make use of the synteny evidence.

Both the methods (manual vs. automated, types of evidence) and results (subsets of different types of homologs) differ. Indeed, the pillars of the YGOB contain groups of orthologous genes that are allowed to contain one ohnolog in each post-WGD species. In contrast, the orthogroups produced by SYNERGY consist of sets of genes from extant species that are descended from a single gene in the species' last common ancestor,²⁷ which means that they contain orthologs as well as all in-paralogs produced since the most ancestral speciation event of the species studied.

Among the 34,709 genes of the six species studied in Wapinski *et al.*,²⁷ 17,210 (50%) were classified into groups of orthologs in a one to maximum one relationship (Additional file 5: Comparison of gene classification in different types of subsets by IONS, YGOB, and SYNERGY). In comparison, the manual curation of the YGOB database allowed classification of 42,046 (69%) of the 60,876 genes of the 11 species of the YGOB v5*. Of the 47,874 genes of the 9 species studied in our test case, our automatic method IONS classified 29,279 (61%) into groups of orthologs (Ortho1 and Orthox) that can serve as anchor points for syntenic studies.

The difference in percentages between YGOB and IONS can be mainly explained by the evolutionary distance between the species included in our study. Indeed, if we removed the two most ancestral species that are not present in the YGOB database (*D. hansenii* and *Y. lipolytica*), reducing our set of species to a subset of the species included in the YGOB database, the percentage of genes classified by IONS into groups of orthologs increased to 67%. Moreover, it has been shown that for short evolutionary distances, significant sequence divergence occurs before extensive rearrangements of chromosomes. At larger evolutionary distances, however, the number of chromosome rearrangements rises while protein-sequence divergence becomes limited by saturation and functional constraints (see Fig. 5 in ¹).

Comparison of a common dataset

A more precise comparison implies a focus on the same species and on the same subset of genes within these species. This comparison was restricted to

* The YGOBv5 contains the following 11 species: *Vanderwaltozyma polyspora*, *Saccharomyces castellii*, *C. glabrata*, *Saccharomyces bayanus*, *S. cereviae*, *Z. rouxii*, *K. lactis*, *E. gossypii*, *Lanchancea kluyveri*, *Lanchancea thermotolerans*, *Lanchancea waltii*.

genes belonging to four species* common to the three studies (Table 3) for which we had no conflict with correspondence of names and that were classified into subsets of orthologs according to IONS (Ortho1 and Orthox).

Table 3. Comparison of the classification of the 12,691 genes in subsets of orthologs by the IONS method, in the YGOB database, and by SYNERGY.

	IONS (SONS)	YGOB (Pillars)	SYNERGY (Orthogroups)
Number of groups	3,761	3,675	3,639
Mean number of genes per group	3.374	3.453	3.487
Standard deviation	0.899	0.764	1.274
Mean number of genes by species per group	1.000	1.005	1.040

The comparisons were done on 12,691 genes using the Rand index.²⁸ This index determines the similarity between two partitions as a function of positive and negative agreements based on the contingency table of the pairwise assignments of data items. The Rand index ranges from 0 to 1. The ARI²¹ introduces a statistical normalization to yield values close to zero for random partitions.²⁹ A value of 1 indicates a perfect identity between the partitions.

Table 4. Adjusted Rand index of the SONS to the YGOB pillars and SYNERGY orthogroups.

	SACE	CAGL	KLLA	ERGO
SACE		0.9768	0.9810	0.9822
CAGL	0.8954		0.9824	0.9824
KLLA	0.9042	0.9058		0.9961
ERGO	0.9019	0.9030	0.9134	

Notes: ARI with YGOB in upper diagonal and ARI with SYNERGY in lower diagonal. Analysis is restricted to Ortho1 and Orthox.

The Adjusted Rand Index, ARI, (Table 4) was very close to one (0.977–0.996) for the comparison with the pillars of the YGOB, indicating that the orthology assignments were almost exactly identical. The results of the IONS method differed a bit more from the SYNERGY orthogroups (ARI ranging from 0.895 to 0.913). The comparisons of the pillars of the YGOB to the orthogroups produced by SYNERGY also showed more divergent

* The YGOBv5 and IONS have 7 species in common but has the comparison was to be made with Synergy, the comparison was limited to the 4 species common to the three studies. A new version of YGOB (v6) has been released which includes seven new sequences: *Tetrapisipora phaffii*, *Naumovozyma dairenensis*, *Torulaspora delbrueckii* and the complete sequence of *Naumovozyma castellii* (which was previously draft sequenced and called *S. castellii*) as well as draft annotations of *Tetrapisipora blattae*, *Kazachstania africana* and *Kazachstania naganishii*. This new release does not increase the number of species in common with IONS and/or Synergy.

results (the ARI varied from 0.916 to 0.922, see additional file 6: ARI for the comparison between YGOB pillars and SYNERGY orthogroups). These small discrepancies with the SYNERGY orthogroups may be explained by the fact that the YGOB and IONS methods are essentially based on synteny, in contrast to SYNERGY for which synteny is only one of three parameters weighted *a priori* in an automatic assignment. Case studies¹⁹ tend to show that the IONS and YGOB methods are slightly more efficient in identifying orthology relationships in cumbersome contexts. The main advantage of the SYNERGY method vs. YGOB is the automation of the assignment.

Nevertheless, the SYNERGY method is based on three parameters (α , β , and γ) to be determined beforehand: weight protein similarity, synteny similarity, and probability of duplication and losses when rooting a gene tree. Moreover, the weighting of components is in part heuristic and will depend on the set of species considered. In this context, the IONS method may be a better option because it is both automated and based on single constant internal parameters: the width of the initial neighborhood and the synteny constraint (number of genes in common in the neighborhood).

A subset by subset comparison with YGOB is supplied in additional file 7: Discrepancies between the IONS subsets and the YGOB.

Implementation

The IONS program that finds orthologs subsets according to the method described above was written in Perl. The required input is a csv file (Additional file 1: Sample of input and output files) relating to genomes and that contains, in each line, the Coding DNA Sequences (CDS) name, the species abbreviation, the chromosome letter, the relative position, the family name, and the strand (this last information is optional). The program is available on the mini website: http://web.me.com/philogene/IONS-method/IONS_2011.html.

DISCUSSION

RELEVANCE OF THE METHOD

The IONS method subdivides precompiled sets of homologs (based on sequence similarity) using gene neighborhood in an iterative process that gradually decreases neighborhood size until a series of homologs with only one gene per species is obtained. The results are equivalent to those of the YGOB²⁶ or SYNERGY,²⁷ but the IONS method has several advantages:

1. The method is automated, in contrast to the YGOB method, which requires a time-consuming manual curation for each new genome.
2. There is no dataset-dependent parameter. While the parameters of SYNERGY must be redefined according to a new dataset, which can lead to contradictions between orthologs found in actual and subsequent results, the IONS method is applicable without reconfiguration. The addition of new species will not change the composition of extant groups of Ortho1 and Orthox; it will offer only the opportunity to complement them or to identify new groups.
3. The method is applicable to any predetermined families of homologs and versatile enough either to use with any existing package to define families of homologs or to use an existing database of families as an input*. Another option would have been to develop a full package integrating both the delineation of families and the search for orthologs. The limitation of this option is the impossibility of taking advantage of the new development in the definition of families and the difficulty of using pre-determined classifications such as eggNOG³⁰ on which our method was tested, giving results similar to the Génolevures dataset. Some standard methods of family determination are proposed on the website.
4. The algorithm is based on a conservative approach that favors the most stringent criteria and minimizes the number of false positives.

An originality of the method is to allow some flexibility in parameterization**, such as the neighborhood size and synteny constraint.

Neighborhood size

The initial number of neighbors considered on each side of the query gene was arbitrarily set to 15 based on current knowledge of the size of Hemiascomycetes syntenic blocks.¹⁸ This choice also seems to be suitable for novel yeast species because the distribution of mean syntenic blocks size ranges between 14 and 26 genes.¹

Note that this initial number of neighbors is not critical because an originality of the method is that the process is iterative. Evolutionary

* It is important to understand that contrarily to some methods like SYNERGY, the IONS method is based on pre-compiled families of homologous genes and attempts to delineate sets of orthologous genes among these. So the analysis of newly sequenced genomes requires a first step of compilation of families of homologous genes which can be done by several methods like the ones proposed on the website.

** The parameterization is independent of the set of species but the flexibility offered is interesting for particular analysis such as the detailed analysis of the evolution of a family of genes.

mechanisms are not the same in different parts of a genome, so we could not expect that a standard neighborhood size would be appropriate for all gene families. The sequential procedure is a way to circumvent this limitation. The size of the neighborhood used to fix a SONS may vary from subset to subset. In some cases, a subset is defined using 15 neighbors on both sides of the query gene; in other cases, the iterative process leads to the subdivision of the initial subset into smaller units using fewer neighbors (the criteria to stop the subdivision are described in the “Sequential procedure” section).

Synteny constraint

The fact that only one neighbor has to be in common to assign two query genes to the same cluster may seem a rather low requirement. The algorithm allows modification of this criterion (using more neighbors in common), which may slightly decrease the false-positive rate (7 genes of 12,691 in our comparison with YGOB as the gold standard). However, this increase strongly decreases the number of subsets of orthologs with one gene in each species because extensive chromosome rearrangements may occur for the most evolutionarily distant species*. In our test case on Hemiascomycetes, this number decreased from 1,697 identified subsets with a criterion of one neighbor to 948 with a requirement of two neighbors in common and to 428 with three neighbors. The IONS method also will benefit from the intensification of sequencing efforts because the method was designed to integrate new data quickly.

Evolutionary span

The rationale of the method is that the considered evolutionary span of the analyzed species is short enough to retain information on sequence similarity and neighborhood. Otherwise, new species are required. Indeed, at large evolutionary distances, while protein-sequence divergence becomes limited by saturation and functional constraints, extensive chromosome rearrangements may occur,¹ shuffling the traces of co-ancestrality. Any method of orthology detection must confront this limitation. The only solution is to diminish the evolutionary distance between genomes by filling the evolutionary gaps with newly sequenced genomes. The constant diminution of the cost of sequencing will certainly contribute to this objective and, as already mentioned, our method easily accommodates new species.

The IONS procedure reaches its maximum efficiency when families of homologs used as inputs are accurately and comprehensively calculated. If families are not accurate—for example, if a gene product is not present in a

* Experimental data, not shown.

family—the current version of our program will not be able to find an ortholog that was placed in a wrong family. Because the method is conservative, a lack of information will never lead to wrong results but will decrease the number of identified SONS.

Using high coverage genome sequences allows avoidance of the problem of false gene losses.³¹ High coverage also limits the probability of genome assembly errors that could lead to cases of false negatives in which orthology is not detected between two genes because of a lack of shared neighborhood. The probability of false positives resulting from assembly errors is close to zero because it would require a consistent misalignment of two regions in two different species.

The method was designed to yield a single final result, but the record of intermediate steps allows further analysis. For example, when the phylogenetic history is complicated by a whole genome duplication generating ohnologs, it is possible to easily identify the two SONS corresponding to the same set of ohnologs.

PERSPECTIVES

Species that are phylogenetically distant may present considerable sequence and synteny divergence, which makes it difficult to detect similarity at the nucleotide level and thus to classify gene products accurately into families. The addition of new species belonging to the same phylum will probably reduce sequence divergence, allowing a better classification of genes into families and improving results. The quality of the mapping, sequencing, and identification of the coding regions of these new species is crucial: comprehensive identification of genes and of their location relative to each other, as well as an accurate classification into families of homologs, is required to take advantage of the IONS method.

CONCLUSIONS

The identification of orthologs is a major issue in comparative genomics. The combination of both similarity and neighborhood evidence facilitates the identification of orthologs. The IONS method was developed using *Hemiascomycetes* genome sequences carried out by the Génolevures Consortium. The performance of IONS is comparable to that of more labor-intensive methods such as YGOB. The automatic nature of the procedure paves the way for easy application to new genomes.

ACKNOWLEDGEMENTS

We would like to thank the Génolevures Consortium, coordinated by Jean-Luc Souciet, for access to the database of protein sequences from *Zygosaccharomyces rouxii*, *Kluyveromyces thermotolerans*, *Saccharomyces kluyveri*, and *Eremothecium gossypii* as well as Julie Diffels for her help in the development of the IONS method. We also thank Thomas Rolland and André Goffeau for helpful discussions, Laurence Jassogne for help in language corrections and two anonymous reviewers for helpful comments. With the support of the Fonds Spécial de Recherche (FSR) de l'Université de Louvain.

DISCLOSURES

Author(s) have provided signed confirmations to the publisher of their compliance with all applicable legal and ethical obligations in respect to declaration of conflicts of interest, funding, authorship and contributorship, and compliance with ethical requirements in respect to treatment of human and animal test subjects. If this article contains identifiable human subject(s) author(s) were required to supply signed patient consent prior to publication. Author(s) have confirmed that the published article is unique and not under consideration nor published by any other publication and that they have consent to reproduce any copyrighted material. The peer reviewers declared no conflicts of interest.

REFERENCES

1. Souciet JL, Dujon B, Gaillardin C, *et al.* Comparative genomics of protoploid Saccharomycetaceae. *Genome Res.* 2009;1696–709.
2. Jackson AP, Gamble JA, Yeomans T, *et al.* Comparative genomics of the fungal pathogens *Candida dubliniensis* and *Candida albicans*. *Genome Res.* 2009;19:2231–44.
3. Wolfe KH. Comparative genomics and genome evolution in yeasts. *Phil Trans R Soc B.* 2006;361:403–12.
4. Dujon B. Hemiascomycetous yeasts at the forefront of comparative genomics. *Curr Opin Genet Dev.* 2005;15:614–20.
5. Dujon B, Sherman D, Fischer G, *et al.* Genome evolution in yeasts. *Nature.* 2004;430:35–44.

6. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*. 2003;423:241–54.
7. Rubin GM, Yandell MD, Wortman JR, *et al.* Comparative genomics of the eukaryotes. *Science*. 2000;287:2204–15.
8. Wapinski I, Pfeffer A, Friedman N, Regev A. Natural history and evolutionary principles of gene duplication in fungi. *Nature*. 2007;449:54–61.
9. Fitch WM. Distinguishing homologous from analogous proteins. *Syst Biol*. 1970;19:99–113.
10. Koonin EV. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet*. 2005;39:309–38.
11. Hulsen T, Huynen M, de Vlieg J, Groenen P. Benchmarking ortholog identification methods using functional genomics data. *Genome Biol*. 2006;7:R31.
12. Kuzniar A, van Ham RCHJ, Pongor S, Leunissen JAM. The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet*. 2008;24:539–51.
13. Salichos L, Rokas A. Evaluating ortholog prediction algorithms in a yeast model clade. *PLoS ONE*. 2011;6:e18755.
14. Alexeyenko A, Tamas I, Liu G, Sonnhammer ELL. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics*. 2006;22:E9–15.
15. Altenhoff AM, Dessimoz C. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol*. 2009;5:e1000262.
16. Chen F, Mackey AJ, Vermunt JK, Roos DS. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE*. 2007;2:e383.
17. Rolland T, Neuveglise C, Sacerdot C, Dujon B. Insertion of horizontally transferred genes within conserved syntenic regions of yeast genomes. *PLoS ONE*. 2009;4:e6515.
18. Fischer G, Rocha EPC, Brunet F, Vergassola M, Dujon B. Highly variable rates of genome rearrangements between hemiascomycetous yeast lineages. *PLoS Genet*. 2006;2:e32.

19. Seret ML, Diffels JF, Goffeau A, Baret PV. Combined phylogeny and neighborhood analysis of the evolution of the ABC transporters conferring multiple drug resistance in hemiascomycete yeasts. *BMC Genomics*. 2009;10:459.
20. Palma M, Seret ML, Baret PV. Combined phylogenetic and neighbourhood analysis of the hexose transporters and glucose sensors in yeasts. *FEMS Yeast Res*. 2009;9:526–34.
21. Hubert L, Arabie P. Comparing partitions. *J Classif*. 1985;2:193–218.
22. Nikolski M, Sherman DJ. Family relationships: Should consensus reign?— consensus clustering for protein families. *Bioinformatics*. 2007;23:e71–6.
23. Sonnhammer ELL, Koonin EV. Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet*. 2002;18:619–20.
24. Wolfe K. Robustness—it's not where you think it is. *Nature Genet*. 2000;25:3–4.
25. Dias PJ, Seret ML, Goffeau A, Correia IS, Baret PV. Evolution of the 12-Spanner Drug:H⁺ Antiporter DHA1 Family in Hemiascomycetous Yeasts. *OMICS*. 2010;14(6):701–10.
26. Byrne KP, Wolfe KH. The yeast gene order browser: Combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res*. 2005;15:1456–61.
27. Wapinski I, Pfeffer A, Friedman N, Regev A. Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics*. 2007;23:1549–58.
28. Rand WM. Objective criteria for the evaluation of clustering methods. *J Am Stat Asso*. 1971;66:846–50.
29. Handl J, Knowles J, Kell DB. Computational cluster validation in postgenomic data analysis. *Bioinformatics*. 2005;21:3201–12.
30. Muller J, Szklarczyk D, Julien P, *et al.* eggNOG v2.0: Extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res*. 2010;38:D190–5.
31. Milinkovitch MC, Helaers R, Depiereux E, Tzika AC, Gabaldón T. 2x genomes—depth does matter. *Genome Biol*. 2010;11(2):R16.

ADDITIONAL FILES

ADDITIONAL FILE 1- SAMPLE OF INPUT AND OUTPUT FILES

Input file

I. Sample of the coding sequences database

CDS name	Species	Chromosome	Position	GL family	Strand
CAGL0A00165g	CAGL	A	1	GL3R3564	-1
CAGL0A00187g	CAGL	A	2	GL3R0757	-1
CAGL0A00209g	CAGL	A	3	GL3R1034	1
CAGL0A00231g	CAGL	A	4	GL3C2428	1
CAGL0A00253g	CAGL	A	5	GL3C3460	-1

The input file is composed of six columns: the CDS name, the species abbreviated by a four letter code, the letter referring to the chromosome, the relative position of the CDS on the chromosome, the family in which the CDS has been classified and the strand on which it is encoded.

Output files

I. Sample of the output database

CDS name	GL family	Subset	Status	Number of neighbours
CAGL0F07139g	GL3C0034	10010_10_20_10	Ortho1	4
ZYRO0D09042g	GL3C0034	10010_10_20_20_20_10	Orthox	2
KLTH0F17974g	GL3C0034	10010_10_30	Undet	5
DEHA2D09020g	GL3C0533	10010_20	Orphx	9
YALI0A18425g	GL3R0548	10020	Orph1	15

II. Sample of the subsets database

Family	Subset	Status	SACE	CAGL	ZYRO	KLTH	SAKL	KLLA	ERGO	DEHA	YALI	Total
GL3C0034	10010_10_20_10	Ortho1	1	1	1	1	1	1	1	1	1	9
GL3C0034	10010_10_20_20_10_10	Orph1	1	0	0	0	0	0	0	0	0	1
GL3C0034	10010_10_20_20_10_20	Ortho1	1	1	1	1	1	1	1	1	1	9
GL3C0034	10010_10_20_20_20_10	Orthox	1	1	1	1	1	1	1	1	0	8
GL3C0034	10010_10_20_20_20_20	Orthox	0	1	1	1	1	1	1	0	0	6

The first database is ordered by gene and contains its family, the name and status of the subset to which it was assigned as well as the number of neighbors used to create this subset. The second database is organized by subsets. It shows the family, the subset, the status of the subset, its phyletic pattern and the total number of genes in the subset.

Output files

III. Sample of a neighborhoods file

	SACE0G03168g	CAGL0F07139g	ZYRO0E01606g	KLTH0H06204g	SAKL0A06710g	KLLA0B00561g	ERGO0A08822g
GL3C0127	1	0	0	1	0	0	0
GL3C1158	0	0	1	1	1	0	1
GL3C2473	1	1	1	1	1	1	1
GL3C2885	1	0	0	0	1	1	0
GL3C3840	1	1	1	0	1	0	0

IV. Sample of the GL families database

GL family	Content
GL3C0500	CAGL0F04235g; CAGL0G03729g; DEHA2E12914g; ERGO0D18788g; KLLA0E12343g; KLLA0F00396g; KLTH0D00726g; KLTH0E14498g; SACE0B01804g; SACE0O11880g; SAKL0B06776g; SAKL0D14762g; YALI0A10681g; ZYRO0C00528g; ZYRO0G13090g
GL3C0501	CAGL0F05071g; DEHA2A00462g; DEHA2E02024g; DEHA2F04488g; ERGO0G14652g; KLLA0C04686g; KLTH0G15114g; SACE0L13903g; SACE0O07810g; SAKL0A08008g; YALI0A07733g; YALI0B10406g; YALI0F22121g; ZYRO0D14454g; ZYRO0D14520g
GL3C0502	CAGL0F08261g; CAGL0I02486g; DEHA2D02112g; DEHA2G14058g; ERGO0E15048g; KLLA0A09185g; KLTH0B08514g; SACE0G11616g; SACE0H05346g; SACE0M10758g; SACE0O12320g; SACE0P00154g; SAKL0H02024g; YALI0F16819g; ZYRO0C11814g

The neighborhoods file reports the presence-absence of the different families in the neighborhoods of the different queries of the subset. The GL families' database lists all genes by family.

V. Example of visual file showing a subset of the GL3C0034 family

Subset of orthologs 10010_10_20_10

SPECIES	N -4	N -3	N -2	N -1	Query	N +1	N +2	N +3	N +4
SACE	SACE0G03278g GL3R1427	SACE0G03256g GL3C0127	SACE0G03234g GL3C2885	SACE0G03190g GL3C2473	SACE0G03168g GL3C0034	SACE0G03146g GL3C3840	SACE0G03124g GL3R1424	SACE0G03102g GL3R1423	SACE0G03080g GL3R1422
CAGL	CAGL0F07271g GL3C1425	CAGL0F07249g GL3R0805	CAGL0F07183g GL3C2473	CAGL0F07161g GL3U7853	CAGL0F07139g GL3C0034	CAGL0F07117g GL3C3840	CAGL0F07095g GL3R1424	CAGL0F07073g GL3R1423	CAGL0F07051g GL3R1422
ZYRO	ZYRO0E01496g GL3R0315	ZYRO0E01518g GL3C0632	ZYRO0E01540g GL3R3388	ZYRO0E01584g GL3C2473	ZYRO0E01606g GL3C0034	ZYRO0E01628g GL3C3840	ZYRO0E01650g GL3R1424	ZYRO0E01672g GL3R1423	ZYRO0E01694g GL3C1158
KLTH	KLTH0H06094g GL3R1426	KLTH0H06116g GL3R1427	KLTH0H06160g GL3C0127	KLTH0H06182g GL3C2473	KLTH0H06204g GL3C0034	KLTH0H06226g GL3R1424	KLTH0H06248g GL3R1423	KLTH0H06270g GL3C1158	KLTH0H06292g GL3R1422
SAKL	SAKL0A06556g GL3R0685	SAKL0A06578g GL3C2885	SAKL0A06666g GL3C4073	SAKL0A06688g GL3C2473	SAKL0A06710g GL3C0034	SAKL0A06732g GL3C3840	SAKL0A06754g GL3R1424	SAKL0A06776g GL3R1423	SAKL0A06798g GL3C1158
KLLA	KLLA0B00473g GL3R0685	KLLA0B00495g GL3C2885	KLLA0B00517g GL3C4073	KLLA0B00539g GL3C2473	KLLA0B00561g GL3C0034	KLLA0B00583g GL3C0255	KLLA0B00627g GL3C3095	KLLA0B00649g GL3R2835	KLLA0B00671g GL3C0003
ERGO	ERGO0A08954g GL3R1242	ERGO0A08910g GL3C0081	ERGO0A08866g GL3C4073	ERGO0A08844g GL3C2473	ERGO0A08822g GL3C0034	ERGO0A08800g GL3U5893	ERGO0A08778g GL3R1424	ERGO0A08756g GL3R1423	ERGO0A08734g GL3C1158
DEHA	DEHA2A09460g GL3R1423	DEHA2A09438g GL3R1424	DEHA2A09416g GL3C0318	DEHA2A09394g GL3C2473	DEHA2A09372g GL3C0034	DEHA2A09350g GL3C0019	DEHA2A09328g GL3R2152	DEHA2A09306g GL3R0818	DEHA2A09284g GL3C0264
YALI	YALI0B15722g GL3C4096	YALI0B15708g GL3C0213	YALI0B15686g GL3C0707	YALI0B15664g GL3C2473	YALI0B15642g GL3C0034	YALI0B15620g GL3U7390	YALI0B15598g GL3R0916	YALI0B15576g GL3R0839	YALI0B15554g GL3U7391

The visual file presents the neighborhoods of query genes of a particular subset. Query genes are written in red. Neighbors belonging to a same GL family are highlighted in the same color.

ADDITIONAL FILE 2 - VISUAL REPRESENTATION OF MAIN STEPS OF THE IONS ALGORITHM LEADING TO A SUBSET OF ORTHOLOGS

> Step c, choice of a particular family *k*, e.g. the family GL3C0025

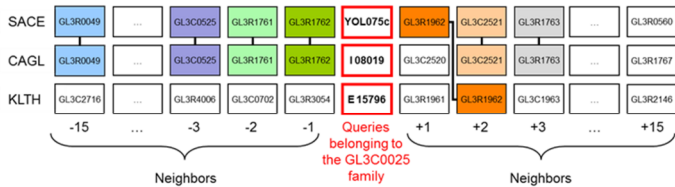
GL3C0025: YOL075c, CAGL0I08019g, ZYRO0B05588g, KLTH0E15796g, SAKL0C08074, KLLA0C04477g, ...

> Step d, identification of the neighbors of each query gene, e.g. YOL075c.

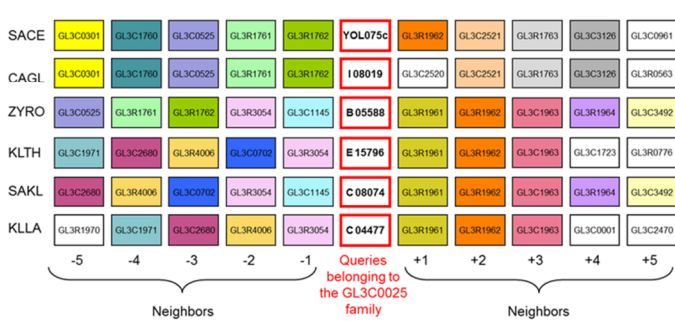
Here the neighbors are directly represented by the family to which they belong to



> Step e, clustering by transitivity

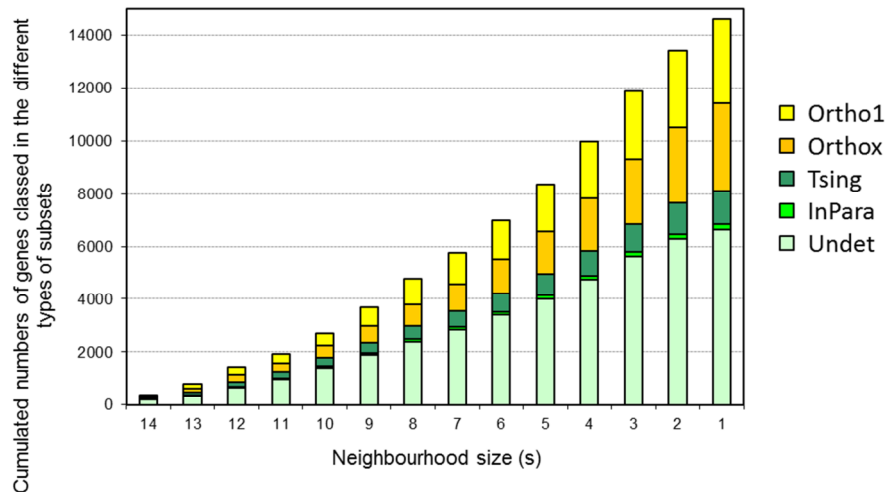


> Example of Subset of Orthologs by Neighborhood and Similarity



Neighbours belonging to a same GL family are highlighted in the same colour.

ADDITIONAL FILE 3 - CUMULATED NUMBERS OF GENES CLASSED BY THE SEQUENTIAL PROCEDURE.



The cumulated numbers of genes are given for each type of subset after each step of the sequential procedure, beginning with a neighbourhood size $s = 14$ till $s = 1$.

ADDITIONAL FILE 4 - RESULTS OF THE IONS METHOD APPLICATION TO THE 47,874 CDS OF GÉNOLEVURES

The file contains the CDS names with their GL family, subset, status and number of neighbours used to define the subset. The last column gives the label of the subset in (Souciet *et al.*, 2009). This classification was obtained with a previous version of the IONS method.

Due to its length this file was not reproduced here, you can download it from: <http://www.la-press.com/ions-identification-of-orthologs-by-neighborhood-and-similarityan-auto-article-a2814>

ADDITIONAL FILE 5 - ADJUSTED RAND INDEX FOR THE COMPARISON BETWEEN YGOB PILLARS AND SYNERGY ORTHOGROUPS

The comparison is done on the 12.691 investigated genes

	SACE	CAGL	KLLA
CAGL	0.9418		
KLLA	0.9536	0.9712	
ERGO	0.9569	0.9504	0.9362

ADDITIONAL FILE 6 – DISCREPANCIES BETWEEN THE IONS SUBSETS AND THE YGOB

Due to its length this file was not reproduced here, you can download it
from: <http://www.la-press.com/ions-identification-of-orthologs-by-neighborhood-and-similarityan-auto-article-a2814>

V. UTILISATION DES SONS COMME POINT D'ANCRAGE POUR LA SYNTÉNIE

Les sous-ensembles d'orthologues positionnels, dénommés SONS (pour « Subset of Orthologs by Neighbourhood and Similarity ») dans l'article précédent ont été utilisés pour construire les points d'ancrage d'une étude de la synténie au niveau génomique sur cinq espèces de levures : *Kluyveromyces (Lachancea) thermotolerans*, *Saccharomyces (Lachancea) kluyveri*, *Zygosaccharomyces rouxii*, *Kluyveromyces lactis* et *Ashbya (Eremothecium) gossypii*.

L'article décrivant ces résultats, intitulé « Comparative genomics of protoploid Saccharomycetaceae » (Souciet *et al.*, 2009) est repris en annexe 1.

DÉFINITION DE LA SYNTÉNIE

Dans cet article, un point d'ancrage de synténie est formé par un minimum de 5 gènes orthologues contigus appartenant à des SONS. Une fois ces points d'ancrage identifiés, on essaie d'étendre les zones de synténie entre les points d'ancrage proches.

En effet, contrairement à la définition généralement utilisée en génétique comparative du terme synténie (ou synténie conservée) qui fait référence à la fois à la conservation du contenu en gènes et à la conservation de leur ordre sur des segments de chromosomes entre deux espèces ou davantage, la définition utilisée ici est plus souple.

Suivant Fischer *et al.* (2006), la définition de synténie conservée utilisée ici permet la présence de gènes intercalés entre les paires synténiques d'orthologues de manière à retrouver un maximum de zones de conservation de l'ordre des gènes.

La définition de blocs de synténie décrits dans l'article du Consortium Génolevures (Souciet *et al.*, 2009) s'effectue donc en deux étapes :

1. Identification des points d'ancrage des blocs de synténie constitués de minimum cinq gènes orthologues physiquement adjacents.

2. Extension de la synténie entre les points d'ancrage proches si ceux-ci sont séparés par un maximum de dix gènes intercalés.

Ceci permet notamment de tenir compte de l'éventualité d'une perte de gènes dans une espèce par rapport à une autre espèce qui aurait conservé l'ensemble des gènes.

La figure supplémentaire n°5 de l'article de Souciet *et al.* (2009), reprise en annexe 1, montre une représentation schématique des orthologues composant un bloc de synténie. Les croix indiquent une perte de gène et les polygones seuls représentent des gènes intercalés spécifiques à une espèce.

INTÉRÊTS DE LA CONSERVATION DE LA SYNTÉNIE

Premièrement, la conservation de la synténie a un aspect prédictif pour l'assemblage des séquences et l'identification des gènes : si les gènes A, B et C appartiennent à une région reconnue comme synténique au sein d'un groupe d'espèces et que le gène A et le gène C ont été identifiés dans une région commune dans une espèce proche nouvellement séquencée, on peut s'attendre à trouver le gène B dans la même région.

Par exemple, dans le cadre de l'article de (Souciet *et al.*, 2009), la correspondance entre les centromères des chromosomes appartenant aux cinq espèces étudiées a pu être déduite de la conservation de la synténie des gènes flanquant et a permis de faire l'hypothèse que leur ancêtre commun possédait vraisemblablement huit chromosomes.

Deuxièmement, la détection de la conservation de la synténie présente un intérêt pour l'annotation fonctionnelle de deux manières : (1) elle peut permettre d'identifier les orthologues parmi les homologues, (2) elle peut permettre d'identifier des gènes susceptibles d'être fonctionnellement liés et/ou d'interagir, comme les opérons par exemple.

Par exemple, dans l'article de (Souciet *et al.*, 2009), la conservation de la synténie a permis de réaliser une étude de la conservation des paralogues dispersés. En considérant uniquement les paires de paralogues pour lesquelles l'orthologie des deux membres est clairement démontrée par la conservation de la synténie pour les cinq espèces (sous-estimation conservative), 114 cas ont pu être identifiés.

Troisièmement, la conservation de la synténie permet aussi de mieux comprendre l'évolution des génomes en identifiant les réarrangements qui les affectent tels que les transpositions, délétions, insertions, inversions, fusions et fissions.

Par exemple, les cinq espèces étudiées dans (Souciet *et al.*, 2009) n'ayant subi aucune duplication génomique, la conservation de la synténie a pu être étudiée afin de déterminer la présence éventuelle de duplications segmentales. A l'exception des éléments transposables connus et des régions subtélomériques, seuls quatre cas de duplication segmentale ont pu être identifiés au sein des cinq génomes.

D'autre part, en utilisant la conservation de la synténie comme indicatrice d'une origine ancestrale commune, 18 répétitions de gènes en tandem ont pu être identifiées comme étant antérieures à la divergence des espèces et conservées (malgré une variation dans le nombre de copies).

L'étude de la relation entre les taux de divergence de séquences et de réarrangements chromosomiques au cours de l'évolution des hémiascomycètes réalisée par Souciet *et al.* (2009), montre que cette relation est non linéaire (Fig. 5), ce qui indique qu'une divergence de séquence significative apparaît avant que les cartes chromosomiques ne soient réarrangées. Par contre, lorsque la distance évolutive augmente, les réarrangements chromosomiques augmentent alors que la divergence de séquence protéique devient limitée par saturation ainsi que par les contraintes fonctionnelles (Souciet *et al.*, 2009).

L'allure relativement lente des réarrangements a également permis d'inférer des événements ancestraux à partir des génomes contemporains d'après la conservation des blocs de synténie. En effet la probabilité de formation indépendante de ces blocs de synténie au sein de différentes lignées est négligeable par rapport à la probabilité qu'ils soient issus d'un ancêtre commun. Des génomes médians ont pu être calculés pour les cinq génomes protoploïdes ainsi que des arbres de réarrangements contenant des candidats pour les génomes ancestraux intermédiaires des différentes branches de l'arbre.

Enfin, (Souciet *et al.*, 2009) souligne que les blocs de synténie conservée contiennent fréquemment de petites insertions (gènes intercalés) ou délétions. Un grand nombre de ces gènes ont des homologues à des emplacements différents dans le même génome ou dans un autre génome et pourraient résulter d'un transfert d'un gène unique (peut-être par rétroposition) ou d'une duplication segmentale dispersée de petite taille. Certains gènes intercalés peuvent également correspondre à un transfert horizontal de gène, comme cela a été montré précédemment pour un gène chez *E. gossypii* (Hall *et al.*, 2005) ainsi que chez d'autres levures (Rolland *et al.*, 2009).

VI. DISCUSSION ET PERSPECTIVES DE LA PARTIE MÉTHODOLOGIQUE

COMPLÉMENTARITÉ DES CRITÈRES POUR L'INFÉRENCE DE L'HOMOLOGIE

La méthode proposée est basée sur deux critères : la similarité de séquences et la conservation du voisinage. La séquence nucléotidique d'un gène codant ancestral, et donc sa séquence en acides aminés est soumise à l'évolution, tout comme son environnement chromosomique. Cependant, l'évolution se produisant au niveau de la séquence d'un gène (par mutation, insertion ou délétion), et celle se produisant au niveau de sa région chromosomique (par duplication, inversion, transposition, translocation, fusion, fission ou délétion) sont différentes, car ces deux types d'objets sont soumis à des mécanismes évolutifs différents. Ces deux types d'objets permettent donc d'avoir des témoignages différents de l'appartenance des gènes existants dans les espèces actuelles à une histoire évolutive particulière, depuis les gènes ancestraux à partir desquels ils ont évolués.

Ces deux critères permettent d'inférer une relation d'homologie. En effet, une forte similarité entre deux séquences est généralement liée au fait que ces séquences proviennent au départ d'un ancêtre commun. L'ajout du critère de conservation du voisinage chromosomique permet d'exclure les cas d'évolution convergente et de renforcer la confiance à accorder à l'inférence.

FORCES DE LA MÉTHODE

En ce qui concerne les forces de la méthode, elles ont déjà été largement discutées dans l'article présenté précédemment, la plupart ne seront donc plus discutées ici. En voici un bref aperçu pour mémoire :

- Automatisation
- Paramètres non dépendants du jeu de données. L'ajout de nouvelles espèces phylogénétiquement situées entre les espèces déjà présentes ne change pas la composition des sous-ensembles de type Ortho1 et Orthox existants mais les complète.

- Applicable à n'importe quelles familles prédéterminées d'homologues afin de bénéficier des éventuelles améliorations dans ce domaine.
- Création de différents types d'outputs, notamment les fichiers visuels permettant l'analyse ultérieure de famille de gènes. L'utilisateur peut choisir de créer ces outputs ou non, ceci dans le but d'accélérer le temps d'exécution de l'algorithme.
- Flexibilité grâce à la possibilité de modifier la taille du voisinage pris en compte et la contrainte relative au nombre minimum de voisins homologues pour considérer que deux environnements sont issus d'un même environnement chromosomique ancestral. Ceci permet de choisir des critères stricts pour minimiser le nombre de faux positifs ou des critères plus lâches pour permettre l'identification des homologues et des orthologues dans des espèces fort distantes.

Ce dernier point sera discuté ci-dessous, afin de mettre en évidence ces implications pour l'identification des homologues et des orthologues.

POUR L'IDENTIFICATION DES HOMOLOGUES

Dans le cadre de l'étude réalisée sur les hémiascomycètes, nous avons choisi de débiter l'analyse avec un environnement chromosomique relativement large composé de 15 gènes disposés de part et d'autre de chacun des gènes d'intérêts. Nous avons également fait le choix de considérer que la présence d'un gène voisin homologue commun dans les voisinages chromosomique de deux gènes d'intérêts était suffisante pour établir que ces environnements chromosomiques ont évolué à partir d'un même environnement chromosomique ancestral. Ces choix, qui peuvent sembler laxistes ont été réalisés de manière intentionnelle afin de ne pas perdre la trace de l'environnement chromosomique ancestral dans les espèces les plus éloignées. Par exemple, dans la Figure 9b, les orthologues n'auraient pas été identifiés chez *Debaryomyces hansenii* et *Yarrowia lipolytica* si la contrainte du nombre de voisins homologues suffisants pour établir que ces environnements chromosomiques ont évolué à partir d'un même environnement chromosomique ancestral avait été supérieure à un.

Le risque lié à ces choix est de regrouper des gènes qui ne sont pas homologues (faux-positifs). Le critère d'un seul voisin homologue peut paraître trop peu discriminant suite à la taille du voisinage chromosomique considéré. En effet, la présence de quelques grandes familles dans la base de données amène le risque d'avoir deux fois la même famille comme voisin par hasard. Cependant, ce risque est relativement limité car la comparaison

des environnements chromosomiques n'est réalisée qu'entre les gènes classés dans une même famille d'après leur similarité. Cela signifie que pour obtenir ce type de faux positifs, il faudrait qu'une famille regroupe deux gènes fortement similaires suite à une évolution convergente et que ces deux gènes possèdent chacun, par pur hasard, un gène appartenant à une même famille dans leur environnement chromosomique.

Afin de rassurer les utilisateurs potentiels, la méthode a été conçue de manière à permettre à l'utilisateur de définir lui-même la « force » avec laquelle il souhaite que les voisinages soient liés pour considérer qu'ils sont issus d'un même environnement chromosomique ancestral. En effet, l'utilisateur peut jouer sur deux paramètres : (i) la largeur de l'environnement chromosomique à prendre en compte, en spécifiant le nombre de voisins à inclure de part et d'autre des gènes d'intérêt ; (ii) le nombre minimal de voisins homologues pour considérer que deux voisinages sont issus d'un même environnement chromosomique ancestral.

Le système hiérarchique de numérotation des sous-ensembles

Bien que l'objectif principal de la méthode ne soit pas d'identifier les gènes homologues, mais bien les orthologues, la numérotation hiérarchique adoptée pour les sous-ensembles de gènes créés au cours de la procédure progressive de la méthode IONS, permet facilement de retrouver l'ensemble des gènes homologues. Ainsi, tous les gènes d'une même famille dont la numérotation hiérarchique du sous-ensemble commence par le même nombre sont homologues. Par exemple, tous les gènes appartenant aux sous-ensembles dont la numérotation commence par 10010_... (10010_10, 10010_20, etc.) sont considérées comme homologues d'après les critères de similarité et de voisinage.

Par exemple, dans la Figure 9a, en ne tenant plus compte que d'un seul voisin de part et d'autre, la procédure progressive finit par séparer les ohnologues présents chez *Saccharomyces cerevisiae* et *Candida glabrata* et dont les environnements chromosomiques respectifs ont été modifiés différemment suite à l'évolution. Deux sous-ensembles d'orthologues numérotés 10010_10 et 10010_20 sont créés. Cependant, si on élargit l'environnement chromosomique (Figure 9b), on s'aperçoit que les gènes des sous-ensembles 10010_10 et 10010_20 sont tous homologues. Le choix de l'ohnologue rattaché au groupe d'orthologues des espèces antérieures à l'événement de duplication génomique complète est basé sur la proximité d'un voisin commun dans le voisinage chromosomique. Si chacun des ohnologues présente en position +1 ou -1 un voisin homologue avec un voisin également positionné en +1 ou -1 chez les espèces antérieures à l'événement de duplication, l'algorithme ne pourra pas séparer les

ohnologues et créera un sous-ensemble de type « Undet » composé d'orthologues et d'ohnologues.

POUR L'IDENTIFICATION DES ORTHOLOGUES

L'inférence d'une relation d'orthologie est plus difficile que l'inférence d'une relation d'homologie. La similarité de séquence ne permet pas directement de différencier les relations d'orthologie des relations de paralogie. Cependant, différentes études montrent que les meilleurs scores bidirectionnels entre séquences appartenant au génome de deux espèces différentes constituent généralement des paires de gènes orthologues.

Les relations d'orthologie peuvent être de type un-un, un-plusieurs ou plusieurs-plusieurs (cfr. introduction). La méthode IONS se concentre sur l'identification des groupes d'orthologues positionnels, qui contiennent au maximum un gène par espèce (un-un avec des absences possibles dans certaines espèces). De ce fait il existe des faux négatifs (relations d'orthologie non identifiées par la méthode).

Cependant, l'étude du voisinage chromosomique telle qu'elle a été implémentée dans l'algorithme de la méthode IONS, permet tout de même d'inférer des relations d'orthologie pour une majorité des gènes. Sur les neuf espèces d'hémiascomycètes analysées, 60% des gènes sont classés en groupes d'orthologues contenant au maximum un orthologue par espèce.

D'autre part, il existe également un risque de faux positif (relation d'orthologie inférée par la méthode alors qu'elle n'existe pas). Les faux positifs sont de deux types :

Premièrement, l'orthologie étant un type d'homologie, comme pour les homologues, si la contrainte synténique est d'un seul voisin homologue pour considérer que les environnements chromosomiques descendent d'un ancêtre commun, il existe un risque d'observer par hasard la présence de deux gènes appartenant à la même famille dans les voisinages de deux gènes non homologues.

Les possibilités de paramétrisation de l'algorithme (taille du voisinage à prendre en compte et nombre minimum de voisins homologues) ainsi que la procédure progressive (qui limite la taille du voisinage pris en compte lorsqu'une espèce possède plusieurs gènes dans le même sous-ensemble) permettent de rendre ce risque nul.

a. Subset of orthologs 10010_10

SPECIES	N -1	Query	N +1
SACE	SACE0G02200g GL3C0011	<i>SACE0G02222g GL3C0389</i>	SACE0G02244g GL3R2923
CAGL	CAGL0104136g GL3C0413	<i>CAGL0104180g GL3C0389</i>	CAGL0104202g GL3R2925
ZYRO	ZYRO0F04840g GL3C0522	<i>ZYRO0F04862g GL3C0389</i>	ZYRO0F04884g GL3R2925
KLTH	KLTH0G16962g GL3U6214	<i>KLTH0G16984g GL3C0389</i>	KLTH0G17006g GL3R2925
SAKL	SAKL0B03322g GL3R0184	<i>SAKL0B03344g GL3C0389</i>	SAKL0B03366g GL3R2925
KLLA	<i>KLLA0A03025g GL3R0184</i>	<i>KLLA0A03047g GL3C0389</i>	KLLA0A03069g Sing
ERGO	<i>ERGO0E01364g GL3R0184</i>	<i>ERGO0E01342g GL3C0389</i>	ERGO0E01320g GL3R2925
DEHA	DEHA2D05522g GL3C0472	<i>DEHA2D05500g GL3C0389</i>	DEHA2D05478g GL3R2925
YALI	YALI0B08184g GL3C1133	<i>YALI0B08206g GL3C0389</i>	YALI0B08228g GL3R2925

Subset of orthologs 10010_20

SPECIES	N -1	Query	N +1
SACE	<i>SACE0P06355g GL3C3145</i>	<i>SACE0P06402g GL3C0389</i>	SACE0P06424g GL3C3894
CAGL	<i>CAGL0L09295g GL3C3145</i>	<i>CAGL0L09339g GL3C0389</i>	CAGL0L09361g GL3U7756

b. Subset of orthologs 10010

SPECIES	N -5	N -4	N -3	N -2	N -1	Query	N +1	N +2	N +3	N +4	N +5
SACE	SACE0G02090g GL3C0001	SACE0G02112g GL3U6316	SACE0G02136g GL3R1717	SACE0G02178g GL3U6318	<i>SACE0G02200g GL3C0011</i>	<i>SACE0G02222g GL3C0389</i>	SACE0G02244g GL3R2925	SACE0G02266g Sing	SACE0G02288g GL3C3894	<i>SACE0G02310g GL3R1716</i>	SACE0G02332g GL3C0389
SACE	<i>SACE0P06377g GL3C0238</i>	<i>SACE0P06392g GL3R2213</i>	SACE0P06314g GL3U6471	SACE0P06336g GL3R0469	<i>SACE0P06358g GL3C3145</i>	<i>SACE0P06402g GL3C0389</i>	SACE0P06424g GL3C3894	SACE0P06446g GL3C0125	SACE0P06490g GL3C0040	<i>SACE0P06512g GL3C3788</i>	<i>SACE0P06534g GL3C3788</i>
CAGL	CAGL0104070g GL3R0843	CAGL0104092g GL3C0908	CAGL0104114g GL3C3316	CAGL0104125g GL3C4083	CAGL0104136g GL3C0413	<i>CAGL0104180g GL3C0389</i>	CAGL0104202g GL3R2925	CAGL0104224g Sing	CAGL0104246g GL3C3894	<i>CAGL0104268g GL3R1716</i>	CAGL0104290g GL3R1717
CAGL	<i>CAGL0L09207g GL3C0238</i>	<i>CAGL0L09229g GL3R2213</i>	CAGL0L09251g GL3U7757	<i>CAGL0L09273g GL3R0469</i>	<i>CAGL0L09295g GL3C3145</i>	<i>CAGL0L09339g GL3C0389</i>	CAGL0L09361g GL3U7756	CAGL0L09383g GL3C3678	CAGL0L09449g GL3C2214	CAGL0L09471g GL3R2215	CAGL0L09493g GL3R0891
ZYRO	ZYRO0F04752g GL3C1014	ZYRO0F04774g GL3R2213	ZYRO0F04796g GL3C0238	ZYRO0F04818g GL3U6858	ZYRO0F04840g GL3C0522	<i>ZYRO0F04862g GL3C0389</i>	ZYRO0F04884g GL3R2925	ZYRO0F04906g Sing	ZYRO0F04928g GL3C3678	<i>ZYRO0F04950g GL3R1716</i>	ZYRO0F04994g GL3R0432
KLTH	KLTH0G16830g GL3U6213	<i>KLTH0G16852g GL3C0389</i>	<i>KLTH0G16874g GL3R0201</i>	<i>KLTH0G16896g GL3R0184</i>	KLTH0G16962g GL3U6214	<i>KLTH0G16984g GL3C0389</i>	KLTH0G17006g GL3R2925	KLTH0G17028g Sing	KLTH0G17050g GL3C3678	<i>KLTH0G17072g GL3R1716</i>	KLTH0G17094g GL3C0008
SAKL	SAKL0B03190g GL3C4075	<i>SAKL0B03212g GL3R0469</i>	<i>SAKL0B03234g GL3C3145</i>	<i>SAKL0B03256g GL3R1261</i>	<i>SAKL0B03278g GL3R0184</i>	<i>SAKL0B03300g GL3C0389</i>	<i>SAKL0B03322g GL3R0184</i>	SAKL0B03344g GL3R2925	SAKL0B03366g GL3R2925	SAKL0B03388g Sing	<i>SAKL0B03410g GL3C4695</i>
KLLA	<i>KLLA0A02937g GL3R3383</i>	<i>KLLA0A02959g GL3C3171</i>	<i>KLLA0A02981g GL3C3145</i>	<i>KLLA0A02993g GL3R1261</i>	<i>KLLA0A03015g GL3R0184</i>	<i>KLLA0A03037g GL3C0389</i>	<i>KLLA0A03059g GL3C0389</i>	KLLA0A03081g Sing	<i>KLLA0A03103g GL3C3894</i>	<i>KLLA0A03125g GL3R1717</i>	<i>KLLA0A03147g GL3C0011</i>
ERGO	<i>ERGO0E01474g GL3R3383</i>	<i>ERGO0E01496g GL3C3171</i>	<i>ERGO0E01518g GL3C3145</i>	<i>ERGO0E01540g GL3R1261</i>	<i>ERGO0E01562g GL3R0184</i>	<i>ERGO0E01584g GL3C0389</i>	<i>ERGO0E01606g GL3R2925</i>	ERGO0E01628g Sing	<i>ERGO0E01650g GL3C3678</i>	<i>ERGO0E01672g GL3R1716</i>	<i>ERGO0E01694g GL3R1717</i>
DEHA	DEHA2D05610g GL3C0166	DEHA2D05588g GL3R1354	DEHA2D05566g GL3R2401	DEHA2D05544g GL3C0243	DEHA2D05522g GL3C0472	<i>DEHA2D05500g GL3C0389</i>	DEHA2D05478g GL3R2925	DEHA2D05456g Sing	DEHA2D05434g GL3C0208	DEHA2D05412g GL3C0021	DEHA2D05390g GL3C0245
YALI	YALI0B08074g GL3R2081	YALI0B08096g GL3R1256	YALI0B08118g GL3C3683	YALI0B08140g GL3R1769	YALI0B08162g GL3C1133	<i>YALI0B08184g GL3C0389</i>	YALI0B08206g GL3R2925	YALI0B08228g Sing	YALI0B08250g Sing	YALI0B08272g GL3C0578	YALI0B08294g GL3R1973

Figure 9. Illustration de la numérotation hiérarchique des sous-ensembles de gènes homologues.

En a, avec un voisin les ohnologues sont séparés. En b, avec 5 voisins, ils font partie d'un seul ensemble d'homologues.

Deuxièmement, dans le cas présent, le risque de faux positif principal est de confondre un paralogue avec un orthologue. En effet, lors de certains types de duplications (segmentale, chromosomique ou génomique), l'environnement direct du gène est également dupliqué.

Dans le cas d'une duplication segmentale, une augmentation de la taille du voisinage et de la contrainte synténique peut permettre d'identifier l'orthologue positionnel.

Dans le cas d'une duplication chromosomique ou génomique, les deux copies sont positionnelles mais elles entraînent la formation de deux lignées distinctes dans les espèces post-duplication. Si une lignée a maintenu une plus grande proportion des gènes voisins adjacents au gène ancestral d'intérêt, la méthode pourra créer un groupe d'orthologues contenant les gènes de cette lignée et ceux des espèces pré-duplication. Dans d'autres cas l'algorithme ne sera pas en mesure de différencier ces lignées et créera un groupe d'homologues rassemblant l'ensemble des gènes (orthologues et ohnologues). Si les environnements chromosomiques ont subi de fortes modifications après les événements de spéciations, il peut arriver que la méthode rassemble deux gènes appartenant chacun à une des deux lignées dans un groupe d'orthologues, créant ainsi un faux positif. Ce risque est accru lors de pertes différentielles d'une des copies dans les espèces postérieures à l'événement de duplication (Figure 10).

Notons que la comparaison avec les données du YGOB suggère que le taux d'erreur de type faux positifs est faible. La comparaison montre 8 cas de divergence sur 12691 : 7 où l'assignation à un même groupe d'orthologues est due à la présence d'un seul voisin homologue et 1 où l'assignation au groupe d'orthologues est liée à des gènes relativement éloigné du gène d'intérêt.

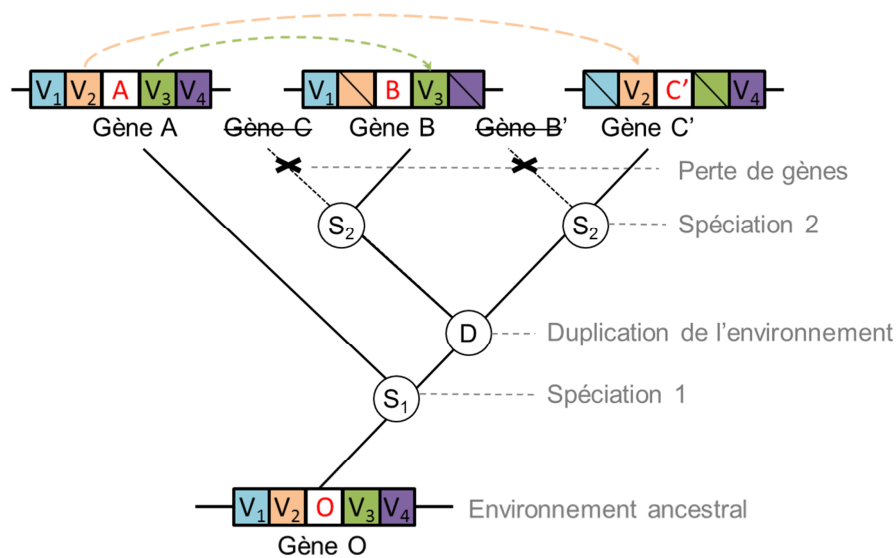


Figure 10. Illustration d'un regroupement incorrect par transitivité.

Par transitivité avec l'environnement chromosomique du gène A, les deux gènes outparalogues (B et C') sont considérés comme orthologues suite à la perte différentielle des gènes B' et C.

LIMITES DE LA MÉTHODE

La méthode présente quatre types de limitations. Les trois premières sont directement dépendantes des choix réalisés lors du développement de la méthode. La quatrième a été constatée à posteriori suite à son utilisation et est dépendante du pourcentage et des types de duplications présentes dans le jeu de données.

Premièrement, pour atteindre l'efficacité maximale de la méthode, les familles de gènes homologues utilisées comme input doivent être complètes et correctes.

Deuxièmement, l'algorithme de recherche des orthologues positionnels a été construit pour des séquences génomiques haploïdes.

Troisièmement, les espèces utilisées doivent être suffisamment proches d'un point de vue évolutif pour que la trace de l'environnement chromosomique ancestral subsiste au sein de leurs génomes. L'évaluation de cette distance évolutive critique n'est pas aisée. En effet, il n'existe pas de calibrage spécifique de l'horloge moléculaire de levures. En dehors des modifications

génomiques elles-mêmes, aucun renseignement indépendant, comme des fossiles, n'est disponible (Rolland et Dujon, 2011).

Afin d'avoir un ordre de grandeur, et à défaut d'autre information la divergence de séquences sera utilisée ici comme approximation de la distance évolutive bien que nous soyons conscients que cette horloge est surtout appropriée aux durées évolutives courtes et perde en précision avec l'augmentation de la distance évolutive (Rolland et Dujon, 2011).

L'algorithme a été lancé séparément sur quatre paires d'espèces, dont les pourcentages d'identité moyenne entre orthologues sont connus et relativement différent (Dujon, 2006). Les résultats exprimés en pourcentage de gènes de *S. cerevisiae* classés comme Ortho1 sont présentés à la Figure 11.

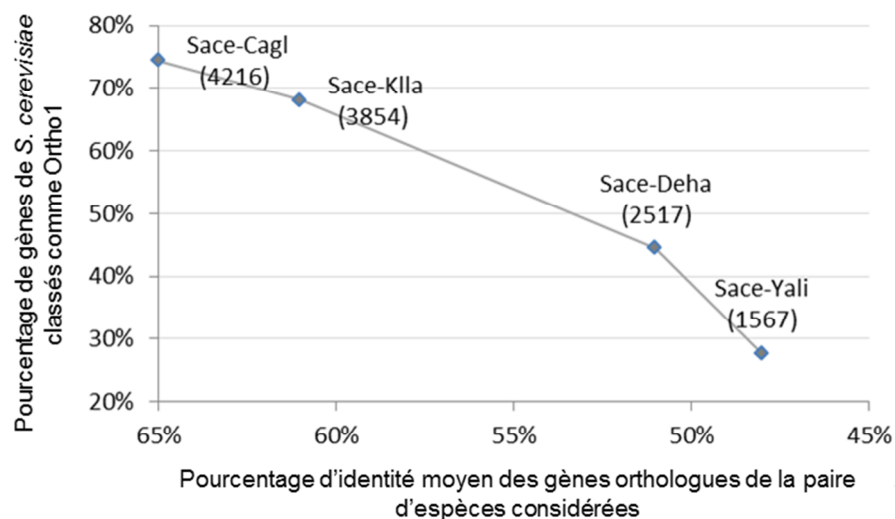


Figure 11 Pourcentage des gènes de *S. cerevisiae* classés comme Ortho1 en fonction du pourcentage d'identité moyen entre orthologues de la paire d'espèces considérées.

Les chiffres repris entre parenthèses correspondent au nombre de gènes classés comme Ortho1 dans une espèce. Ces chiffres ont été divisés par le nombre total de gènes de *S. cerevisiae* analysés (5662) puis exprimés en pourcent.

Pour la paire d'espèces *S. cerevisiae* et *C. glabrata*, dont la valeur moyenne d'identité de séquences entre orthologues est de 65% (Dujon, 2006), l'algorithme identifie 4216 paires d'orthologues positionnels, soit un orthologue pour 74% des gènes de *S. cerevisiae* analysés ce qu'on peut considérer comme un maximum vu la proximité de ces espèces. Pour la paire *S. cerevisiae* - *Y. lipolytica*, dont la valeur moyenne d'identité de séquences entre orthologues est de 48% (Dujon, 2006), l'algorithme n'identifie plus

que 1567 paires d'orthologues, soit un orthologue pour 28% des gènes de *S. cerevisiae* analysés. Sur base de ces données, on peut considérer qu'il n'y a pas de véritable distance évolutive critique pour l'utilisation de la méthode. En effet, la diminution du nombre d'Ortho1 identifiés est directement proportionnelle à la diminution du pourcentage d'identité moyen entre orthologues. L'utilisateur est donc libre de choisir sa limite en fonction de ses objectifs. Par exemple, si un utilisateur souhaite pouvoir identifier un orthologue pour 50% des gènes analysés, il devra se limiter à des espèces dont la valeur moyenne d'identité de séquences orthologues est supérieure ou égale à 55% environ.

Quatrièmement, dans le cadre de l'analyse des neuf génomes réalisée dans l'article précédent et sur base des paramètres définis (voisinage initial de 15 gènes de part et d'autres et contrainte d'un gène synténique), le pourcentage de groupes du type « Undet » est relativement élevé (24,37% soit 11668 gènes, cfr. Table 2, p 61).

Ces groupes « Undet » sont dus à la présence dans une espèce (au minimum) d'un gène dupliqué soit (1) au sein du même environnement chromosomique par duplication en tandem soit (2) dans un second environnement chromosomique identique au premier (au moment de la duplication), allant de quelques gènes à un chromosome entier, ce qui est le cas respectivement pour les duplications segmentales et chromosomiques ou génomiques.

En ce qui concerne les gènes dupliqués au sein du même environnement chromosomique par duplication en tandem, une étude réalisée sur la même base de données de neuf espèces que celle de l'article présentant la méthode IONS a permis de mettre en évidence la présence de 469 groupes de gènes en tandem comprenant au total 999 séquences codantes (Despons *et al.*, 2010). Ces gènes en tandems sont répartis dans 350 des 936 groupes « Undet » et pourraient donc expliquer jusqu'à 37% (350/936) des classifications de type « Undet » en fonction de la constitution du reste du groupe (il est possible que la présence de gènes en tandem ne soit pas la seule cause de la classification du groupe dans le type « Undet »).

Ces tandems sont généralement des duplications récentes apparues au sein d'une espèce. De plus les gènes en tandems peuvent faire l'objet de conversions géniques. Dans ces cas de figure, déterminer quelle copie est l'orthologue positionnel des gènes présents dans les autres espèces n'a pas de sens.

Pour les duplications segmentales, l'inférence de l'orthologue positionnel est théoriquement possible bien qu'elle dépende du rapport entre la taille du segment dupliqué et la taille du bloc de synténie incluant la copie ancestrale du gène.

Enfin dans le cas de duplication chromosomique ou génomique, la notion d'orthologue positionnel n'a pas de sens puisque les deux copies sont « positionnelles ».

PERSPECTIVES

Les limites de la méthode permettent de définir différentes pistes d'amélioration de celle-ci.

Premièrement, afin de s'assurer de ne manquer aucun orthologue suite à une erreur dans la constitution des familles, on pourrait imaginer une étape supplémentaire dans l'algorithme qui permettrait de s'assurer qu'il n'y a, au sein des génomes étudiés, aucun autre environnement chromosomique qui présenterait une synténie importante (contrainte synténique à définir) avec les environnements des gènes classés dans le groupe d'orthologues.

Deuxièmement, en ce qui concerne l'élargissement aux séquences polyploïdes, dans l'état actuel de l'algorithme, les espèces polyploïdes peuvent être facilement analysées en utilisant un seul haplotype pour chaque jeu de chromosomes homologues. Par contre, l'analyse d'un génome tel que le génome de *P. sorbitophila*, une levure diploïde provenant de l'hybridation de deux ancêtres distincts et présentant deux paires de chromosomes hétérozygote avec un événement de translocation* provoquant un changement d'identité du chromosome E avec le chromosome I puis avec le chromosome F et du chromosome J avec le chromosome F puis le chromosome I, est plus compliquée.

Dans ce cas précis, l'analyse de l'identité entre les gènes des chromosomes ayant permis de mettre le problème en évidence, l'algorithme IONS a été utilisé une première fois pour identifier la synténie des chromosomes de *P. sorbitophila* avec ceux de *D. hansenii*, espèce la plus proche d'un point de vue phylogénétique (parmi les génomes repris dans la base de données Génolevures). Les fichiers visuels produits ont permis aux experts de délimiter précisément la zone de translocation. Le génome de *P. sorbitophila* a ensuite été analysé par la méthode IONS via la création manuelle par un expert de deux « demi-génome » haploïde comprenant chacun l'équivalent d'un haplotype pour chaque paire de chromosomes. Sans cette séparation artificielle en deux « demi-génome » haploïdes identifiés par un code d'espèce différent (ex : PIS1 et PIS2), l'utilisation de l'algorithme demanderait au minimum une modification de l'étape de classification qui attribue un statut aux différents sous-ensembles créés (voir Figure 1 box 1 de

* La translocation de chromosomes est un échange de matériel chromosomique entre des chromosomes non homologues.

l'article, p 54) afin de permettre aux groupes d'orthologues (Ortho1 et Orthox) d'avoir deux copies alléliques pour la ou les espèces diploïdes.

Troisièmement, concernant la distance évolutive entre espèces, on notera simplement que l'ajout d'espèces phylogénétiquement situées entre deux espèces éloignées peut permettre d'améliorer l'identification des orthologues entre ces espèces. Par exemple, l'ajout des sept espèces de Génolevures comprises entre *S. cerevisiae* et *Y. lipolytica* sur l'arbre phylogénétique permet, via la transitivité entre voisinages, d'augmenter légèrement le nombre d'orthologues identifiés entre ces deux espèces de 1567 à 1697 (soit de 28% à 30%)

Quatrièmement, concernant la proportion relativement élevée de groupes du type « Undet » dus à la présence des duplications en tandem, segmentales, chromosomiques ou génomiques, différents types d'amélioration peuvent être apportées selon les cas.

Pour les gènes en tandem, une première amélioration à apporter à l'algorithme serait de les identifier, ce qui pourrait être réalisé facilement sur base de la classification en familles et de la position relative des gènes sur le chromosome.

Si le tandem de gènes dupliqués est conservé dans plus d'une espèce, la similarité peut éventuellement permettre d'inférer les orthologues positionnels, comme cela a été réalisé manuellement pour les orthologues de SNQ2 à la figure 3 de l'article « Combined phylogeny and neighborhood analysis of the evolution of the ABC transporters conferring multiple drug resistance in hemiascomycete yeasts » (Seret *et al.*, 2009) (cfr. p 103).

Dans les autres cas, les gènes en tandem étant fréquemment dûs à une duplication récente et sujets aux conversions géniques, il pourrait être utile de considérer le groupe de gènes en tandem comme une seule unité lors de la classification des groupes par la méthode IONS. Dès lors, les groupes de gènes dont la classification de type « Undet » est due exclusivement à la présence de gènes en tandem pourraient être classés dans un nouveau type de groupe d'orthologues (par exemple on pourrait créer une classe OrthoTan, pour les groupes d'orthologues comprenant des gènes en tandem).

Cette modification de l'algorithme IONS pourrait être réalisée très aisément pour les duplications en tandems dont les gènes sont côte à côte. Elle est un petit plus complexe pour les gènes en tandems séparés par des gènes intervenants.

En ce qui concerne les duplications dans un second environnement chromosomique identique au premier (segmentales, chromosomiques, génomiques), on peut distinguer trois cas selon la taille de la duplication.

Pour les duplications segmentales de taille inférieure à la taille du bloc de synténie incluant la copie ancestrale du gène, la mobilisation de l'information recueillie dans le calcul du nombre de voisins en synténie présenté dans les fichiers de voisinage (cfr. p 73) pourrait permettre de différencier la copie qui a été dupliquée, pour autant que la synténie soit suffisamment conservée et, dans le cas de notre méthode, que la taille du segment dupliqué soit inférieure à la taille du voisinage considéré pour la classification.

Pour les duplications segmentales dont la taille est supérieure à la taille du bloc de synténie incluant la copie ancestrale du gène, la distinction entre la copie ancestrale et la copie dupliquée est impossible via la synténie.

Enfin dans le cas de duplication chromosomique ou génomique, on a vu que la notion d'orthologue positionnel n'a pas de sens puisque les deux copies sont « positionnelles ». Cependant, l'utilisation d'un système de calcul du nombre de voisins en synténie devrait néanmoins permettre de déterminer l'environnement chromosomique qui a été le moins modifié par rapport aux espèces antérieures à cette duplication génomique.

De plus, pour toutes les duplications créant un second environnement chromosomique identique au premier (segmentales, chromosomiques, génomiques) et dont les gènes dupliqués sont conservés dans plus d'une espèce, l'utilisation du calcul du nombre de voisins en synténie devrait permettre d'identifier les différentes lignées d'orthologues positionnels initiées à partir des gènes dupliqués. Il pourrait également être utilisé pour quantifier la robustesse des inférences des relations d'orthologie.

**RECHERCHE DES ORTHOLOGUES :
APPLICATION AUX TRANSPORTEURS
CHEZ LES LEVURES**

VII. LES PROTÉINES DE TRANSPORT

Les cellules échangent des composés avec leur environnement (métabolites, ions, acides aminés, nucléotides, ...). Pour réaliser ces échanges, les composés passent à travers des membranes biologiques telles que la membrane plasmique ou les membranes d'organites comme le noyau, le réticulum endoplasmique ou les mitochondries.

Les membranes biologiques sont composées d'une bicouche de lipides. Celle-ci est assez perméable à quelques petites molécules non chargées, comme l'oxygène par exemple, mais son caractère fortement hydrophobe retarde la diffusion des composés hydrophiles, polaires, ou ioniques. L'importation/exportation de ces composés, nécessaires à la survie d'une cellule, se fait via les protéines de transport ou transporteurs.

RÔLES DES TRANSPORTEURS

Les systèmes de transport servent la cellule de multiples manières.

Par exemple, ils permettent l'entrée de toutes les substances nutritives essentielles dans le cytoplasme puis dans les organites, permettant la métabolisation des sources exogènes de carbone, d'azote, de soufre et de phosphore. Ils fournissent le moyen de réguler les concentrations des métabolites, en catalysant l'excrétion des produits terminaux des voies métaboliques en provenance des organites.

Ils prennent également en charge l'influx et l'efflux des espèces ioniques qui doivent être maintenues à des concentrations qui diffèrent drastiquement de celles du milieu externe. Le maintien de conditions favorables à la vie exige un potentiel de membrane, des gradients de concentration d'ions et des concentrations cytoplasmiques appropriées pour tous les minéraux qui participent comme cofacteurs aux processus métaboliques.

Les transporteurs participent aussi au transport des protéines, des glucides complexes et des lipides au sein et à l'extérieur de la cellule. Ces macromolécules ont une variété de rôles biologiquement importants dans la protection (contre un stress, une toxicité environnementale, la prédation), la communication intra et inter-espèces ou encore dans la pathogenèse.

Ils exportent de manière active les substances toxiques et les fongicides vers le milieu extracellulaire et permettent ainsi aux organismes vivants de conduire des « guerres biologiques » en sécrétant, par exemple, des antibiotiques, des agents antiviraux, des agents antimycosiques ou des

toxines qui peuvent conférer un avantage sélectif aux organismes qui les produisent dans le but de favoriser leur survie. Par ailleurs, beaucoup de ces toxines sont elles-mêmes des protéines-canaux ou des peptides qui ont une fonction perturbatrice du transport cellulaire (Saier, 2000).

TYPES DE TRANSPORTEURS

Pratiquement tous les processus de transport transmembranaire sont réalisés via des protéines transmembranaires intégrales, c'est-à-dire des protéines qui traversent la bicouche lipidique de part en part. Celles-ci sont parfois associées à d'autres protéines et forment des complexes.

Les protéines de transport sont réparties en deux grandes classes, les transporteurs, également appelés perméases ou pompes, et les canaux*. Les transporteurs lient un soluté spécifique puis changent leur conformation pour lui permettre de passer de l'autre côté de la membrane. Les canaux, quant à eux, interagissent beaucoup moins avec les solutés à transporter. Ils forment des pores qui s'étendent à travers la bicouche lipidique de la membrane. Lorsqu'ils sont ouverts, ces pores permettent à des solutés spécifiques (généralement des ions inorganiques de taille et de charge appropriée) de passer à travers eux et ainsi de traverser la membrane. Le transport via les canaux est beaucoup plus rapide que via les transporteurs. Bien que l'eau puisse diffuser à travers la bicouche lipidique des membranes, toutes les cellules contiennent des canaux spécifiques, appelés canaux à eau ou aquaporines, qui augmentent fortement la perméabilité des membranes à l'eau (Alberts *et al.*, 2008).

Les systèmes de transport transmembranaire sont de deux types : le transport passif, qui ne nécessite pas d'apport en énergie, et le transport actif qui en nécessite.

L'ensemble des canaux ainsi que certains transporteurs ne permettent aux solutés de traverser la membrane que de manière passive, c'est-à-dire en suivant leur gradient électrochimique, qui combine le gradient de concentration et le gradient électrique. Le transport passif peut être vu comme une diffusion accélérée : les protéines de transport permettent simplement le passage de molécules à travers la membrane à une vitesse beaucoup plus rapide que par la diffusion normale en leur fournissant un passage de taille appropriée, amenant ainsi plus rapidement à un équilibre

* Dans les autres chapitres de cette thèse, le terme « transporteur » (ou « transporter » en anglais) est utilisé de manière générale pour l'ensemble des protéines de transport y compris les canaux.

des concentrations. Les transporteurs passifs ne produisent pas de gradient de concentration.

Le transport actif est réalisé par des transporteurs et implique le déplacement d'un soluté à travers une membrane contre son gradient de concentration, ce qui exige un apport en énergie. Le soluté est transporté à travers la membrane via une modification de conformation réversible du transporteur qui expose successivement le site de liaison du soluté d'un côté de la membrane puis de l'autre. L'apport en énergie provient généralement de l'hydrolyse d'ATP. Les transporteurs sont alors appelés des pompes. Cependant, il peut également provenir d'autres sources, comme la dissipation du gradient de concentration d'un autre soluté par exemple.

On parlera de transport primaire si une source d'énergie primaire, comme une réaction chimique (l'hydrolyse de l'ATP par exemple), l'absorption de lumière ou un flux d'électron, est couplée au processus. On parlera de transport secondaire si une source d'énergie secondaire, c'est-à-dire un gradient électrochimique d'ions, générée aux dépens d'une source primaire d'énergie est couplée au processus.

Les systèmes de transport actif peuvent être classés en 3 types : les uniporteurs, les symporteurs et les antiporteurs, selon le nombre de solutés transportés et la direction respective de transport. Les uniporteurs catalysent le transport d'un seul soluté. Les symporteurs catalysent le transport de deux solutés, ou plus, se déplaçant dans la même direction et les antiporteurs catalysent l'échange de deux ou plusieurs solutés dans des directions opposées.

Chez les levures, la plupart des systèmes de transport actif secondaire sont liés au gradient de protons (H^+) plutôt qu'au gradient d'ions sodium (Na^+). Certains d'entre eux sont impliqués dans le transport de glucides (maltose) par exemple, tandis que d'autres, tels que les protéines de la famille DHA1 (drug : H^+ antiporter-1), confèrent une résistance à de multiples fongicides.

Parmi les transporteurs actifs primaires qui hydrolysent l'ATP en ADP et phosphate, également appelées ATPases, on retrouve notamment la superfamille des transporteurs ABC (ATP-binding cassettes) qui constitue la plus grande famille de protéines de transport et dont il sera question au chapitre suivant.

SYSTÈME DE CLASSIFICATION

En juin 2001, l'International Union of Biochemistry and Molecular Biology (IUBMB) a formellement adopté le système TC (Transport Classification) de classification des transporteurs (Busch et Saier, 2002).

La base de données associée à ce système, appelée TCDB (Transport Classification database) est une base de données représentative. Une première base de données de type exhaustif, appelée YETI (Yeast Transport Information), a été réalisée en 2006 (De Hertogh *et al.*, 2006) de manière semi-automatique sur cinq espèces de levures.

L'actualisation de cette base de données de transporteurs et son extension aux cinq nouvelles espèces d'hémiascomycètes séquencées par Génolevures a été réalisée. Elle est présentée au chapitre IX sous forme d'un article soumis dans FEMS Yeast Research.

VIII. ETUDE DE L'ÉVOLUTION D'UNE FAMILLE DE TRANSPORTEURS

Avant son utilisation à l'échelle de génomes complets, une version antérieure de la méthode IONS a préalablement été testée sur trois familles de transporteurs afin de valider la méthode et de s'assurer qu'elle produisait les résultats nécessaires aux études de l'évolution de familles de gènes.

L'étude de l'évolution des transporteurs ABC conférant une résistance multiple aux fongicides est présentée ci-dessous sous forme d'un article publié dans *BMC Genomics* en 2009, vol 10 : 459.

Les études des deux autres familles ont fait l'objet d'une collaboration avec d'autres universités.

L'étude des transporteurs d'hexoses a été réalisée en collaboration avec Margarida Palma du Centro de Recursos Microbiológicos de la New University of Lisbon et a fait l'objet d'une publication dans *FEMS Yeast Research* en 2009 intitulée : « Combined phylogenetic and neighbourhood analysis of the hexose transporters and glucose sensors in yeasts ». Elle est reprise en annexe 2 et consultable en ligne gratuitement à l'adresse suivante : <http://onlinelibrary.wiley.com/doi/10.1111/j.1567-1364.2009.00511.x/full>

L'étude de la famille DHA1 réalisée en collaboration avec Paulo Jorge Dias et Isabel Sá Correia de l'Instituto Superior Técnico de la Technical University of Lisbon, a fait l'objet d'une publication dans *OMICS A Journal of Integrative Biology* en 2010 intitulée : « Evolution of the 12-Spanner Drug:H⁺ Antiporter DHA1 Family in Hemiascomycetous Yeasts ». Elle est reprise en annexe 3.

Ces trois études permettent de faire le lien entre la partie méthodologique et la partie appliquée aux transporteurs. Elles ont notamment permis de mettre en évidence l'utilité d'une paramétrisation de la méthode. La version actuelle de la méthode IONS a été créée suite au travail réalisé sur ces trois familles. Les améliorations apportées visent à diminuer le temps de calcul, augmenter le nombre de sous-ensembles d'orthologues positionnels définis via l'introduction de la procédure progressive ainsi qu'à permettre à l'utilisateur de choisir la taille de l'environnement chromosomique à prendre en compte et la contrainte synténique qu'il souhaite. De plus, deux types de fichiers de résultats ont été ajoutés. Premièrement, la base de données des sous-ensembles qui donne le profil phylétique des sous-ensembles formés, c'est-à-dire le nombre de gènes inclus dans le sous-ensemble pour chaque espèce.

Deuxièmement, le fichier de voisinage qui permet de quantifier le nombre de voisins communs entre les voisinages de chaque paire de gènes appartenant à un même sous-ensemble.

Mes contributions personnelles à cet article :

- *Développement du principe de base de la méthode d'identification du voisinage, en collaboration avec Julie Diffels, et programmation*
- *Curation manuelle des sous-ensembles d'homologues et d'orthologues et préparation des figures*
- *Rédaction de l'article avec Julie Diffels et révision par Philippe Baret et André Goffeau*
- *Les tables ont été réalisées par Julie Diffels, la recherche de motifs par André Goffeau et l'analyse phylogénétique par Philippe Baret*

**COMBINED PHYLOGENY AND NEIGHBORHOOD ANALYSIS OF THE
EVOLUTION OF THE ABC TRANSPORTERS CONFERRING MULTIPLE
DRUG RESISTANCE IN HEMIASCOMYCETE YEASTS**

Marie-Line Seret¹, Julie F Diffels¹, André Goffeau² and Philippe V Baret*¹

Address: ¹Unité de Génétique (GENA), Université Catholique de Louvain, Louvain-la-Neuve, Belgium and ²Unité de Biochimie Physiologique

(FYSA), Institut des Sciences de la Vie (ISV), Université Catholique de Louvain, Louvain-la-Neuve, Belgium

* Corresponding author email: philippe.baret@uclouvain.be

BMC Genomics 2009, 10:459

doi:10.1186/1471-2164-10-459

This article is available from: <http://www.biomedcentral.com/1471-2164/10/459>

© 2009 Seret et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ABSTRACT

BACKGROUND

Pleiotropic Drug Resistant transporters (PDR) are members of the ATP-Binding Cassette (ABC) subfamily which export antifungals and other xenobiotics in fungi and plants. This subfamily of transmembrane transporters has nine known members in *Saccharomyces cerevisiae*. We have analyzed the complex evolution of the pleiotropic drug resistance proteins (Pdrp) subfamily where gene duplications and deletions occur independently in individual genomes. This study was carried out on 62 Pdrp from nine hemiascomycetous species, seven of which span 6 of the 14 clades of the *Saccharomyces* complex while the two others species, *Debaryomyces hansenii* and *Yarrowia lipolytica*, are further apart from an evolutive point of view.

RESULTS

Combined phylogenetic and neighborhood analyses enabled us to identify five Pdrp clusters in the *Saccharomyces complex*. Three of them comprise orthologs of the Pdrp *sensu stricto*, Pdr5p, Pdr10p, Pdr12p, Pdr15p, Snq2p and YNR070wp. The evolutive pathway of the orthologs of *Snq2* and *YNR070w* is particularly complex due to a tandem gene array in *Eremothecium gossypii*, *Kluyveromyces lactis* and *Saccharomyces (Lachancea) kluyveri*. This pathway and different cases of duplications and deletions were clarified by using a neighbourhood analysis based on synteny. For the two distant species, *Yarrowia lipolytica* and *Debaryomyces hansenii*, no neighborhood evidence is available for these clusters and many homologs of *Pdr5* and *Pdr15* are phylogenetically assigned to species-based clusters. Two other clusters comprise the orthologs of the *sensu lato* Pdrp, Aus1p/Pdr11p and YOL075cp respectively. The evolutionary pathway of these clusters is simpler. Nevertheless, orthologs of these genes are missing in some species.

CONCLUSION

Numerous duplications were traced among the Hemiascomycetous Pdrp studied. The role of the Whole Genome Duplication (WGD) is sorted out and our analyses confirm the common ancestrality of Pdr5p and Pdr15p. A tandem gene array is observed in *Eremothecium gossypii*. One of the copies is the ortholog of *Snq2* while the other one is lost in the post-WGD species. The neighborhood analysis provides an efficient method to trace the history of genes and disentangle the orthology and paralogy relationships.

BACKGROUND

The phenotype for pleiotropic drug resistance (PDR) in *Saccharomyces cerevisiae* (SACE) was discovered in 1973 by RANK and BECH-HANSEN [1], who reported that single gene mutations were responsible for resistance to multiple drugs of different chemical structures and different targets. The genetic and molecular mapping revealed that these mutations are located in the transcription factors Pdr1p and Pdr3p [2-4] and increased the expression of a series of target genes [5] including those encoding the ABC efflux pumps Pdr5p [6-8], Snq2p [9], Pdr10p and Pdr15p [10,11] or Pdr11p [10].

Analysis of the full genome sequence of *Saccharomyces cerevisiae* [12] revealed that these transporters are members of a large phylogenetic subfamily of ABC transporters named Pdrp (for pleiotropic drug resistance proteins), where nuclear binding folds (NBF) alternate with domains comprising six predicted transmembrane spans (TMS) to form a NBF-TMS-

NBF-TMS pattern [13]. NBF includes a distinctive series of amino acid sequence motifs, including Walker A, Walker B and ABC signatures involved in ATP binding and hydrolysis.

The inventory of all full-sized *Saccharomyces cerevisiae* ABC transporters reveals six Pdrp *sensu stricto* namely: Pdr5p, Pdr10p, Pdr12p, Pdr15p, Snq2p and YNR070wp* [13], characterized by three criteria: (1) The alternation NBF-TMS-NBF-TMS, which is different from the classical TMS-NBD-TMS-NBD topology reported for the mammalian multidrug resistance MDR and MRP drug efflux pumps [14,15]; (2) The presence of a cysteine residue (C) instead of the lysine residue (K) in N-terminal Walker A motifs as well as the specific NVEQ motif in the C-terminal ABC signature [6,13]; (3) The efflux of multiple drugs [16-18].

The structural determinants for the broad substrate specificity of Pdrp pumps have not yet been identified even though dozens of publications have reported hundreds of amphiphilic substrates for Pdr5p and Snq2p. These two pumps share many substrates even though some drugs are preferentially handled by either one or the other [17,19]. In contrast, Pdr12p exhibits clear substrate specificity for short chain weak organic acids [20].

Fully sequenced Ascomycota and Basidiomycota were all examined and found to contain several Pdr proteins [21] but their evolutionary lineage is unclear. A previous phylogenetic classification of Pdrp from five species spanning the full Hemiascomycetes phylum identified seven subclusters [22]. GBELSKA *et al.* [23] showed that in contrast to the other ABC subfamilies which remain as single copy in each species during evolution, a series of gene duplications occurs independently in the PDR subfamily in five hemiascomycetous species. GAUR *et al.* [24] compared the NBF domains of *S. cerevisiae* and *Candida albicans* Pdrp. They concluded that *C. albicans* contains a cluster of four proteins homologs to Pdr5p and a single protein homolog to Snq2p.

In addition to the Pdrp clusters *sensu stricto*, the *S. cerevisiae* genome contains two other ABC phylogenetic clusters exhibiting the alternation of NBF-TMS-NBF-TMS domains [13]. One cluster comprising Aus1p and Pdr11p are sterol influx pumps [25] while another cluster comprising YOL075cp is of undetermined function. As the members of these two phylogenetic clusters show neither drug efflux properties nor the K/C substitution in the Walker A1 motif, we classified them as Pdrp *sensu lato*.

* YNR070wp is also known has Pdr18 (Rockwell *et al.*, 2009). A recent study suggests that Pdr18 plays a role in plasma membrane sterol incorporation, this physiological trait contributing to a multidrug resistant phenotype (Cabrito *et al.*, 2011).

The aim of the present paper is to trace the evolution of the Pdrp subfamily in nine species belonging to the *Hemiascomycetes* phylum, believed to have separated from the filamentous fungi 300 to 400 million years ago [26,27]. Seven of these species span the 14 clades constituting the *Saccharomyces complex* [28]. Among those, three novel genome sequences from the *Lachancea* or *Zygosaccharomyces* clades were analyzed: *Saccharomyces (Lachancea) kluyveri* (SAKL), *Zygosaccharomyces rouxii* (ZYRO) and *Kluyveromyces (Lachancea) thermotolerans* (KLTH) [29]. Two hemiascomycetous sequences of older lineage; *Debaryomyces hansenii* (DEHA) and *Yarrowia lipolytica* (YALI) were also explored. An exhaustive database of all the chromosomal neighbors of the PDR genes from all nine hemiascomycetous species was developed (SERET *et al.*, in preparation) and allowed us to carry out combined phylogenetic and neighborhood analysis. Gene neighbourhood analysis has been pioneered by BYRNE and WOLFE [30] for tracing the ancestral gene blocks produced by the Whole Genome Duplication (WGD). We will illustrate the extent and the limits of the information provided by analysis of gene neighborhood data for deciphering the evolution of the large Pdrp subfamily in addition to classical phylogenetic analysis based only on amino acid sequence similarity.

RESULTS

The Pdrp transporters are a well characterized family of genes implied in drug resistance. All genes share typical features, such as Walker A and B and an ABC signature, and are very similar in terms of sequence. In the Génolevures database, the family GL3C0025 which contains all SACE Pdrp comprises 62 members (Table 1) of which 6 are considered as pseudogenes or fragments (Additional file 1). The phylogenetic tree in Figure 1, based exclusively on full-length proteins to improve multiple alignments, shows that the 56 full-Pdr proteins of the GL3C0025 family (Additional file 2) branch into five clusters labeled A to E. Each cluster is named according to its *S. cerevisiae* member(s): A, the cluster Pdr12p (8 members); B, the cluster Snq2p/YNR070wp (11 members); C, the cluster Pdr5p/ 10p/15p (25 members) which ramifies into 4 subclusters labeled C1, C2, C3, C4; D, the cluster YOL075cp (7 members); E, the cluster Aus1p/Pdr11p (5 members).

Neighborhood analysis (Additional file 3) was carried out to delineate Subsets of Orthologs by Neighborhood and Similarity (SONS). Seven SONS labeled "a" to "g" (Figure 2) provide information on the evolution of 40 proteins displaying informative neighborhood among the 62 PDR from the GL3C0025 family.

Table 1: The members of the GL03C0025 family comprising all SACE Pdrp.

Species	Acronym	Clade ¹	PDR sensu stricto	Homologs to YOL75c to AUS1		Full size ABC transporters	Number of frag- ments
<i>Saccharomyces cerevisiae</i>	SACE	I	6	I	2	9	-
<i>Candida glabrata</i>	CAGL	4	4	I	I	6	-
<i>Zygosaccharomyces rouxii</i>	ZYRO	7	8	I	-	9	-
<i>Kluyveromyces thermotolerans</i>	KLTH	10	2	I	-	3	2
<i>Saccharomyces kluyveri</i>	SAKL	10	7	I	I	9	I
<i>Kluyveromyces lactis</i>	KLLA	11	4	I	-	5	-
<i>Eremothecium gossypii</i>	ERGO	12	3	-	I	3	-
<i>Debaryomyces hansenii</i>	DEHA	-	5	-	-	6	3
<i>Yarrowia lipolytica</i>	YALI	-	5	I	-	6	-
Total			44	7	5	56	6

¹ according to KURTZMAN[28]

For most genes, the phylogenetic clusters are confirmed by the neighborhood analysis. But in some cases, divergences are observed. In the next sections, each phylogenetic cluster of genes is discussed in an evolutionary perspective integrating the evidence from the neighborhood analysis.

CLUSTER A (PDR12P)

This cluster comprises 8 members in SACE, CAGL (*Candida Glabrata*), ZYRO, SAKL, KLLA (*Kluyveromyces lactis*) and YALI (Figure 1). No full-Pdrp orthologs are detected in KLTH, even though the neighborhood is well maintained on chromosome E, ERGO (*Eremothecium gossypii*) and DEHA. The neighborhood analysis (Figure 2 SONS a) shows a clear connection (9 neighbors) between *SAKL0H10670g* and the PDR belonging to the other species, except YALI. A second PDR copy is present in SAKL on chromosome C. *SAKL0C05654g* only shares one neighbor, belonging to the GL3R2315 family. The very high amino acid identity (98%, 17 substitutions) between the two paralogs suggests a very recent duplication. A tandem gene array occurred in ZYRO where the two copies diverged a little more (87% amino acid identity). None of these duplications (on different chromosomes in SAKL and within the same chromosome in ZYRO) were maintained in other species. Pdr12p and Snq2p are in closely related clusters as they share a high similarity rate. Despite this proximity of sequence, neighborhood evidence shows that they evolved separately, at least since the speciation time of KLLA.

CLUSTER B (SNQ2P/YNR070WP)

The phylogenetic Snq2p cluster comprises 11 members that are represented in all the examined species except YALI. A shared neighborhood is present from ERGO to SACE (Figure 2 SONS b). It comprises a single ortholog of Snq2p in each species except in SACE where YNR070wp is a paralog* of Snq2p. Moreover, neighborhood analysis shows an apparent tandem gene array in ERGO chromosome B. However, the two copies are different (38% amino acid identity). ERGO0B8140p has 63% identity to Snq2p and only 41% identity to Pdr5p (Figure 3). In contrast, ERGO0B8162p has a higher amino acid identity to Pdr5p (53%) than to Snq2p (36%). A similar heterologous tandem in a similar neighborhood is observed for KLLA and SAKL with, however, the insertion of an increasing number of genes between the two members of the repeat. ZYRO also presents two genes in this neighbourhood but they are no longer in tandem gene array. It is

* YNR070w is not the ortholog of Snq2p. The duplicated copy of the chromosomal environment of Snq2p originated from the WGD can still be identified in *C. glabrata* and *S. cerevisiae* but the ortholog copy of Snq2p lacks (data not shown).

ortholog copies of *Pdr5/15* from ZYRO are inserted in a neighborhood that can be traced back in KLLA, SAKL and KLTH but none of these chromosomal environments contain a *Pdr5/15* homolog. However, in each of these species a *Pdr5/15*-like copy is present in a neighborhood different from that of the ZYRO triplicate. The KLTH, SAKL and KLLA copies show a strong phylogenetic link with Pdr5p/15p: nearly 67% amino acid identity for KLTH0G19448p, nearly 70% amino acid identity for SAKL0C08008p and nearly 65% amino acid identity for KLLA0F21692p. *KLLA0F21692g* and *SAKL0C08008g* share a common neighborhood (see Figure 2, SONS f) which is not shared with *KLTH0G19448g* (see Figure 2, SONS g). The ancestor of the ZYRO copies was duplicated by the WGD and produced the ohnologs *Pdr5* and *Pdr15* [30]. In fact, *Pdr15*, which is located on chromosome D, possesses two common neighbors with ZYRO while *Pdr5*, located on chromosome F, shares five common neighbors with ZYRO. Only the *Pdr15* copy was retained in CAGL in this neighborhood. A second PDR, *CAGL0M01760g*, with a high amino acid identity with Pdr5p/15p is present in CAGL in another neighborhood.

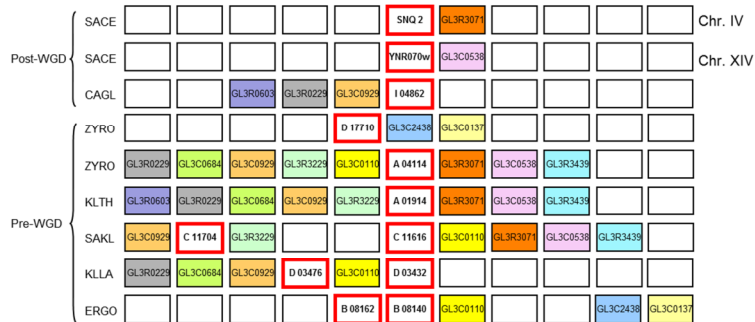
Pdr10p shows high amino acid identity to Pdr5p/15p (65% and 62% respectively). Segment alignment reveals that Pdr10p can be differentiated from other SACE Pdrp by a 22 amino acid insert (SDAAIMGNDKTVAKEHYSSPSS), located at the intersection of the two half PDR regions (NBF-TMS-insert-NBF-TMS). This marker means that Pdr10p can be detected in *S. cerevisiae*, *Saccharomyces bayanus*, *Saccharomyces paradoxus* and *Saccharomyces mikatae* but no ortholog was found in CAGL and in the pre-WGD species (data not shown, inferred from Broad Institute [31]). The event that gave rise to Pdr10p is thus post-WGD.

In addition to the major subcluster C1 (Figure 1) comprising Pdr5/10/15p, cluster C contains three additional phylogenetic subclusters. Two of these subclusters show a spectacular species-based clustering. Indeed, very similar proteins tend to cluster with proteins belonging to the same species rather than with orthologs in other species. One of the subclusters (C3) comprises five DEHA members. The neighborhood analysis suggests a putative link between *DEHA2F16478g* and the orthologs of *YOL075c*. The other subcluster (C4) comprises four YALI members. Neighborhood analysis of this subcluster is not conclusive but phenotype analysis reveals that one of them (YALI0C20265p) is an azole efflux transporter with properties similar to those of Pdr5p/Snq2p [32] while another (YALI0E14729p) has the unique property to export C16 hydrocarbon residues [32].

▪ SONS a - PDR 12



▪ SONS b - SNQ2 / YNR070w



▪ SONS c – PDR5 / PDR15

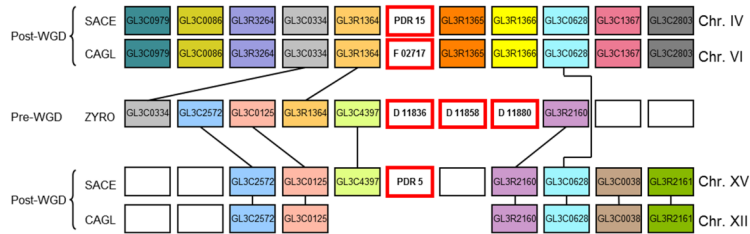


Figure 2. Manually* curated Subsets of Orthologs by Neighborhood and Similarity of Pdrp.

(Part 1). SONS a: Orthologs of Pdr12 (cluster A). SONS b: Orthologs of Snq2 and YNR070w (cluster B). SONS c: Orthologs of Pdr5 and Pdr15 (cluster C).

* As some subsets were of the type “Undet” with several members in several species, they were curated manually to distinct between the different chromosomal environments.

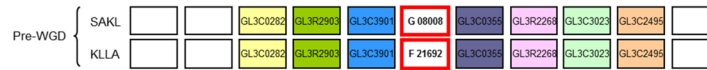
▪ SONS d – YOL075c



▪ SONS e – AUS1



▪ SONS f



▪ SONS g

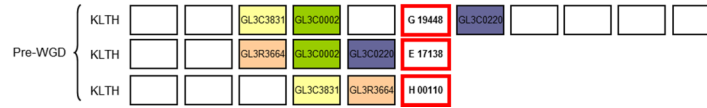


Figure 2. Manually curated Subsets of Orthologs by Neighborhood and Similarity of Pdrp.

(Part 2). SONS d: Orthologs of *YOL075c* (cluster D). SONS e: Orthologs of *Aus1* (cluster E). SONS f: Subset of Orthologs by Neighborhood and Similarity consisting of *SAKLOG08008g* and *KLLA0F21692g* (cluster C). SONS g: Subset of Orthologs by Neighborhood and Similarity made up of *KLTHOG19448g* (cluster C) and the two telomeric fragments *KLTHOE17138g* and *KLTHOH00110g*. Each box framed in red represents a PDR. Adjacent boxes represent the PDR neighbors. Homologous neighbors, based on Génolevures families, are highlighted in the same color. These SONS were calculated with 15 neighbors on each side, but the representation was truncated to 5 neighbors for the sake of clarity.

The last subcluster (C2) contains five members: *SAKL0C06996g*, *SAKL0C11704g*, *KLLA0D03476g* as well as *ERGO0B08162g* and *ERGO0G05126g* which are very similar (73% amino acid identity). Three of them, *ERGO0B08162g*, *KLLA0D03476g* and *SAKL0C11704g* are in tandem gene array. Interestingly, they belong to the C cluster (Pdr5p/10p/15p) when

using phylogenetic criteria but share some neighbors with the B cluster (Snq2p). This is also the case of *ZYRO0D17710g* of the subcluster C1.

CLUSTER D (YOL075CP)

This cluster comprises a total of 7 members; one per species except in DEHA and ERGO. A spectacular shared neighborhood presence is observed for the orthologs of *YOL075c* (see Figure 2, SONS d). Eight common neighbors are, for instance, identified in the post-WGD species SACE and CAGL, four common neighbors are identified in the pre-WGD species and four common neighbors link ZYRO, one of the pre-WGD species to the post-WGD species. The YALI homologue (YALI0D25828p) has no conclusive neighborhood but it exhibits high amino acid identity (41%) to the SACE YOL075cp. This suggests that all hemiascomycetous orthologs of *YOL075c* originate from an ancestor existent prior to the speciation event of YALI.

The neighborhood analysis links a tandem gene array made up of *DEHA2F16478g* and the fragment *DEHA2F16500g* to cluster D. However, the *DEHA2F16478g* is classed in subcluster C3 by phylogeny and does not display the specific motifs of YOL075cp orthologs.

CLUSTER E (AUS1P/PDR11P)

This cluster comprises members of four species only: SACE, CAGL, SAKL and ERGO. Using neighborhood analysis, an ortholog of Aus1p was only identified in CAGL. There is no evidence that the two SACE paralogs, Aus1p on chromosome O and Pdr11p on chromosome I, issue from the WGD as there is no shared neighborhood with pre-WGD species. In SACE, both Aus1p and Pdr11p are essential for anaerobic growth as they contribute to the uptake of ergosterol under conditions which prevent the aerobic synthesis of sterols [25]. The sterol transporter function was recently demonstrated for the CAGL ortholog of Aus1p [33]. The two characteristic Pdrp motifs, N-terminal Walker A and the C-terminal ABC signature, remain, but only partially (Additional file 1).

FRAGMENTS

Six fragments of 72 to 1099 amino acids belonging to the GL3C0025 protein family have been identified. Despite their subtelomeric location, that we define as the last 30 kb of each chromosome following FAIRHEAD and DUJON [34], the two fragments *KLTH0E17138g* and *KLTH0H00110g* share the neighborhood of *KLTH0G19448g*, the ortholog of *Pdr5/15*, which may suggest duplications followed by pseudogenization. Because phylogenetic

analysis of fragments may be error-prone, their evolutionary significance will not be analyzed.

DISCUSSION

Orthologous PDR transporters with typical NBF-TMS-NBF-TMS topology are found in all higher fungi such as *S. cerevisiae* [6,13], the hemiascomycetous human pathogen *C. albicans* [35], *Euascomycetes* [36], *Archiascomycetes* [37,38], *Basidiomycetes* [39] and all plants [40] including *Arabidopsis thaliana* [41] and *Nicotiana plumbaginifolia* [42]. In contrast, the PDR topology has not been detected in protozoa [43] or chordates [44] where multiple drug resistance results from the overexpression of MDR/MRP efflux transporters of opposite TMS-NBF-TMS-NBF topology.

Our analysis of 56 full-sized Pdrp from 9 different Hemiascomycetes shows that the Génolevures subfamily GL3C0025 [29] is subdivided into five phylogenetic clusters labeled A to E (Figure 1). Each of these clusters can be characterized by its SACE members originally identified by DECOTTIGNIES and GOFFEAU [13]. Within the phylogenetic *Saccharomyces* complex defined by KURTZMAN [28], a total of 40 Pdrp homologs sharing significant chromosomal neighborhood were classified in seven SONS, labeled "a" to "g" (Figure 2), in which phylogenetic and neighborhood data are combined.

It now appears that the Pdrp *sensu stricto* phenotype can be allocated only to cluster A (Pdr12p), to cluster B (Snq2p/YNR070wp) and to cluster C (Pdr5p/10p/15p), given that all three share not only the NBF-TMS-NBF-TMS topology but also the two additional Pdrp traits: efflux drug pumping and typical Walker A1 and ABC Signature 1 motifs.

On the other hand, no experimental information is available concerning the unknown substrates of the members of cluster D (YOL075cp), while the SACE members of cluster E (Aus1p/Pdr11p) are clearly sterol influx pumps [25]. Moreover, as neither cluster D nor cluster E contains members with the typical Pdrp motifs, they may be considered to be Pdrp *sensu lato*.

Our combined phylogenetic and neighborhood analyses support the evolutionary pattern illustrated in Figure 3 for the Pdrp *sensu stricto* of the *Saccharomyces* complex [28].

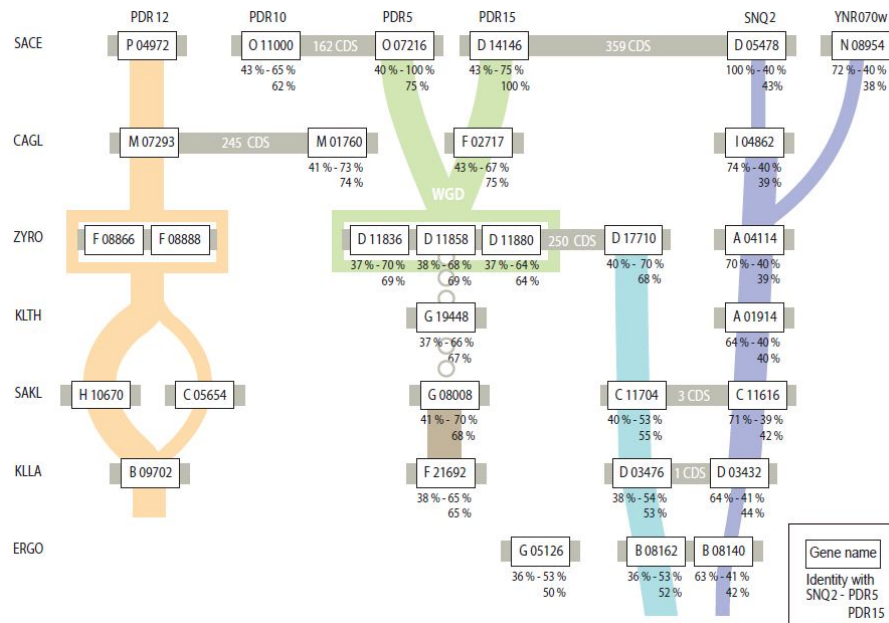


Figure 3. Relationships between the PDR *sensu stricto* based on the neighborhood analysis.

Each box represents a PDR gene. Grey lines join PDR genes located on the same chromosome. When non-adjacent, the number of coding sequences between two PDR genes is noted in white. Note that *SAKLOC05654g*, *SAKLOC11616g* and *SAKLOC11704g* are on the same chromosome. Shared neighborhoods are represented by connectors whose colors match those of Figure 1. Two colors were used for SONS b in order to distinguish the lineage of *ZYROOD17710g* and the one of *Snq2* and *YNR070w*. A large connector links PDR genes sharing more than one common neighbor while a thin connector links PDR genes sharing one neighbor only. Dotted connectors link genes whose relationship is based on phylogenetic evidence only. For the story leading to *Snq2* and *Pdr5/15*, the percentage of identity with their homologs in the other species is given. The label WGD stands for Whole Genome Duplication. Our representation does not imply that the present ERGO species is the ancestor of all the other species we analyzed.

Pdr12p is a *Pdr sensu stricto*, according to our three criteria. Its origin remains questionable* as no *Pdr12p* neighbors are shared either with *Pdr5p/15p* or *Snq2p*. In SACE, the function of *Pdr12p*, which effluxes food preservatives such as short chain weak acids [20], is different from that of *Snq2p* and *Pdr5p* which share the function to efflux a series of antifungal azoles and other hydrophobic substrates [17]. Nevertheless, the SACE *Pdr12p* sequence is closer to SACE *Snq2p* (46%) than to SACE *Pdr5p*

* We just meant that for now, we do not know how the different *Pdr sensu stricto* are related, that is how and when they begin to diverge.

(38%). This may suggest a common origin of both Pdr12p and Snq2p clusters. The sequence analysis of new yeast species, phylogenetically situated between DEHA and ERGO, should make it possible to disentangle the exact evolution lineage.

The unlinked ancestor of Pdr5p/15p ohnologs of KLTH jumped and triplicated on ZYRO chromosome D. An ancestor of the three copies, already in this neighborhood, was duplicated through WGD and produced Pdr5p and Pdr15p in all *Saccharomyces*.

The phylogenetic tree of the Pdr5p/15p clusters shows several subclusters (C3 and C4) which aggregates members belonging to specific species such as DEHA or YALI. These are cases where species-based clustering is suspected of hindering a functional inference based on sequence similarity only. For example, the subcluster C4 contains four YALI members belonging to four different chromosomal fragments in unshared environments. Despite being inside the same subcluster, these genes seem to have very different functions. One of them, *YALI0E14729g*, exhibits the function of alkane efflux, unique as yet, while another, *YALI0C20265g*, seems to control the classic azole resistance function of the SACE Pdr5p/15p pumps.

The evolutionary patterns of the Pdrp *sensu lato* are simpler and totally independent from that of the Pdr *sensu stricto*. The members of Pdr *sensu lato* Aus1p/Pdr11p cluster are present in four different species only: ERGO, SAKL, CAGL and SACE. These genes may be lost in other species. Indeed, sterol influx pumps are essential only under anaerobiosis conditions. However, no clear relation between the presence of homologs of Aus1p or Pdr11p sterol-influx pumps and facultative anaerobic growth could be detected among the species composing the *Saccharomyces* complex [45].

The members of the Pdr *sensu lato* YOL075c cluster are of unknown function and belong to species of both ancestral clade (*YALI0D25828g*) and recent clades. No homolog of YOL075c is found in DEHA and ERGO. Extensive neighbourhood preservation is observed for all *Kluyveromyces* and *Saccharomyces* species. Even though the substrates of this cluster of putative transporters are unknown, its continued presence in many pre- and post-WGD species from the *Saccharomyces* complex indicates that these gene products exert an important physiological function. This function and the subcellular localization of the transporter have not yet been determined. Curiously all members of the YOL075c SONS have at least 100 less amino acid residues (1247 to 1328) than all other members of the GL3C0025 family. Moreover, the cysteine of SGCT motif of the N-Walker A is replaced by the "classic" lysine residue and the NVEQ motif of the C-Walker Signature is replaced by SGGE (Additional file 1).

In three cases only, we observed a divergence between the neighborhood and phylogenetic analyses. *ERGO0B08162g*, *KLLA0D03476g* and *SAKL0C11704g* are allocated to the C cluster (Pdr5p/15p) by phylogeny and to the SONS b (Snq2p/YNR070wp) by neighborhood. This divergence can be explained by the fact that they belong to a tandem gene array in which a copy has been subjected to neofunctionalization. In another case, *ZYRO0D17710g* (subcluster C1) is allocated to SONS b (Snq2p/YNR070wp) by neighborhood. Its high amino acid similarity with Pdr5p (70%) and significant difference of sequence with *SAKL0C11704g* and *ZYRO0A04114g* suggest an evolution based on gene conversion* with one of the member of the *ZYRO0D11836g*, *ZYRO0D11858g* and *ZYRO0D11880g* tandem gene array. Finally, the divergence observed for *DEHA2F16478g* (subcluster C3 and SONS d) may be linked to a species-based clustering of DEHA genes within the subcluster C3.

CONCLUSION

In order to assess the evolution of genes using a neighbourhood analysis, we considered four more species in addition to those examined in the study by Gbelska *et al.* [23]. Our classification is consistent for the 24 genes in the five species we share with Gbelska *et al.* Moreover, the additional species enabled us to follow more accurately the evolution of the chromosomal environments of the Pdrp and to distinguish between paralogous and orthologous status, and, in so doing, clarify the fate of some ERGO and KLLA genes. We found that some orthology relationships are confused by SYNERGY [23,46] (see Additional file 4). Our analysis converged with the Yeast Gene Order Browser [30], except for the orthology relationship of the two members of the KLLA tandem gene array which are inverted in the YGOB[†].

We demonstrate that, combined with phylogenetic analysis, neighborhood analysis allows us to trace back the detailed evolutionary history of most members of the phylogenetic subfamilies of Pdrp *sensu stricto* from SACE, CAGL, ZYRO, KLTH, SAKL, KLLA and ERGO. These species span 6 out of the 14 clades from the phylogenetic *Saccharomyces* complex [28] that may have separated from the other yeasts around 150 million years ago. We may therefore conclude that Pdr5p and Pdr15p issued from a single ancestral gene present before the WGD. In contrast, both Pdr10p (a Pdr5p paralog) and

* Gene conversion is a process by which DNA sequence information is transferred from one DNA helix (which remains unchanged) to another DNA helix, whose sequence is altered.

[†] This error of the YGOB has been corrected since version v3 associated to the publication by (Gordon *et al.*, 2009).

YNR070wp (a Snq2p paralog) originated from independent post-WGD duplication events.

We show that for a large subfamily such as Pdrp, whose members share high amino acid sequence similarity, neighborhood analysis can be used to trace the duplicates from pre-WGD or post-WGD and to describe in detail the tandem or repeat duplications and the gains, loss or transposition of gene copies.

METHODS

PDRP

Protein sequences from *Saccharomyces cerevisiae* (SACE), *Candida glabrata* (CAGL), *Zygosaccharomyces rouxii* (ZYRO), *Kluyveromyces thermotolerans* (KLTH), *Saccharomyces kluyveri* (SAKL), *Kluyveromyces lactis* (KLLA), *Eremothecium gossypii* (ERGO), *Debaryomyces hansenii* (DEHA) and *Yarrowia lipolytica* (YALI) were retrieved from the Génolevures database in July 2007. A bioinformatics analysis based on BLAST and Smith-Waterman alignments, clustering and Condorcet election, was applied to the Génolevures data set by NIKOLSKI and SHERMAN [47] in order to define consensus families. All SACE Pdrp were found to belong to a single family, the GLC2004, recently relabeled GL3C0025. This family comprises 62 members (Table 1). Six peptides of 72 to 1099 amino acid residues were considered as fragments, leaving 56 full-sized members which were further analyzed. Genes are labelled according to the Génolevures nomenclature [48].

PRINCIPLE OF ASSIGNATION TO SONS

Within the Génolevures GLC3C0025 family, the Subsets of Orthologs by Neighborhood and Similarity (SONS) were identified as schematized in Additional files 2 and 3. Fifteen neighbor genes were retrieved on each side of the 62 PDR (called "query gene"). Each of these 30 genes is assigned to a Génolevures family. As family assignment is based on similarity criteria, neighbor genes belonging to the same GL family were considered orthologs by similarity. Two PDR query genes, sharing at least one neighbour being classified in the same GL family, were regarded as belonging to the same SONS. Indeed, in this case, the orthology based on similarity evidences is confirmed by the neighborhood information. Out of the 62 Pdrp, 40 show informative neighborhoods. Due to the complexity of the PDR evolutionary history, the SONS were manually curated to remove some weak links between chromosomal environments. The curated output of the SONS

analysis is presented in Figure 2. For the sake of clarity, the representation is limited to five neighbor genes on each side of the query PDR genes. In fact, no conclusive neighbourhood could be identified in the DEHA or YALI species. Therefore all neighbor analyses of Figure 2 excluded DEHA and YALI and were restricted to the seven species we examined from the *Saccharomyces* complex defined by KURTZMAN [28], except one DEHA gene in SONS d.

MOTIFS

The Pdrp sequences were aligned with ClustalW [49] on <http://www.ebi.ac.uk/clustalw/> and visualized with Jalview [50]. The N-Walker A and C-terminal ABC signature motifs were identified based on alignments with the motifs defined by DECOTTIGNIES and GOFFEAU [13] for the Pdrp of *S. cerevisiae*. Identity was calculated after removing the filter for low complexity.

PHYLOGENETIC TREE

Multiple alignment was calculated by MUSCLE [51]. The phylogenetic tree in Figure 1 was constructed using PROTDIST and NEIGHBOR in the PHYLIP suite [52]. The use of Maximum likelihood (PROTML) or parsimony approach gave identical results. Only NEIGHBOR JOINING results are, therefore, presented. Representation of the phylogenetic tree made use of the Dendroscope application [53].

AUTHORS' CONTRIBUTIONS

M-LS and JD designed and developed a computer program to identify SONS. M-LS did the manual curation of SONS and the preparation of figures. JD prepared the tables. AG identified the PDR motifs and PB did the phylogenetic study. M-LS and JD drafted manuscripts with revisions from AG and PB. All authors read and approved the final manuscript edited by M-LS.

ACKNOWLEDGEMENTS

We would like to thank the Génolevures consortium, coordinated by Jean-Luc SOUCIET, for helpful discussions and access to the database of protein sequences from *Zygosaccharomyces rouxii*, *Kluyveromyces thermotolerans*, *Saccharomyces kluyveri* and *Eremothecium gossypii*. We acknowledge the constructive comments of two anonymous reviewers.

REFERENCES

1. Rank GH, Bech-Hansen NT: Single nuclear gene inherited cross resistance and collateral sensitivity to 17 inhibitors of mitochondrial function in *S. cerevisiae*. *Molecular and General Genetics MGG* 1973, 126:93-102.
2. Carvajal E, vandenHazel HB, CybularzKolaczowska A, Balzi E, Goffeau A: Molecular and phenotypic characterization of yeast PDR1 mutants that show hyperactive transcription of various ABC multidrug transporter genes. *Molecular & General Genetics* 1997, 256:406-415.
3. Delaveau T, Delahodde A, Carvajal E, Subik J, Jacq C: Pdr3, A New Yeast Regulatory Gene, Is Homologous to Pdr1 and Controls the Multidrug-Resistance Phenomenon. *Molecular & General Genetics* 1994, 244:501-511.
4. Subik J, Ulaszewski S, Goffeau A: Genetic-Mapping of Nuclear Mucidin Resistance Mutations in *Saccharomyces cerevisiae*. *Current Genetics* 1986, 10:665-670.
5. DeRisi J, Hazel B van den, Marc P, Balzi E, Brown P, Jacq C, Goffeau A: Genome microarray analysis of transcriptional activation in multidrug resistance yeast mutants. *FEBS Letters* 2000, 470:156-160.
6. Balzi E, Wang M, Leterme S, Vandyck L, Goffeau A: Pdr5, A Novel Yeast Multidrug-Resistance Conferring Transporter Controlled by the Transcription Regulator Pdr1. *Journal of Biological Chemistry* 1994, 269:2206-2214.
7. Bissinger PH, Kuchler K: Molecular-Cloning and Expression of the *Saccharomyces cerevisiae* Sts1 Gene-Product - A Yeast ABC Transporter Conferring Mycotoxin Resistance. *Journal of Biological Chemistry* 1994, 269:4180-4186.
8. Meyers S, Schauer W, Balzi E, Wagner M, Goffeau A, Golin J: Interaction of the Yeast Pleiotropic Drug-Resistance Gene-Pdr1 and Gene-Pdr5. *Current Genetics* 1992, 21:431-436.
9. Servos J, Haase E, Brendel M: Gene Snq2 of *Saccharomyces cerevisiae*, Which Confers Resistance to 4-Nitroquinoline-NOxide and Other Chemicals, Encodes A 169-Kda Protein Homologous to Atp-Dependent Permeases. *Molecular & General Genetics* 1993, 236:214-218.
10. Decottignies A, Grant AM, Nichols JW, de Wet H, McIntosh DB, Goffeau A: ATPase and multidrug transport activities of the overexpressed yeast ABC protein Yor1p. *Journal of Biological Chemistry* 1998, 273:12612-12622.

11. Wolfger H, Mahe Y, ParleMcDermott A, Delahodde A, Kuchler K: The yeast ATP binding cassette (ABC) protein genes PDR10 and PDR15 are novel targets for the Pdr1 and Pdr3 transcriptional regulators. *FEBS Letters* 1997, 418:269-274.
12. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG: Life with 6000 genes. *Science* 1996, 274:563-567.
13. Decottignies A, Goffeau A: Complete inventory of the yeast ABC proteins. *Nature Genetics* 1997, 15:137-145.
14. Chen CJ, Chin JE, Ueda K, Clark DP, Pastan I, Gottesman MM, Roninson IB: Internal Duplication and Homology with Bacterial Transport Proteins in the Mdr1 (P-Glycoprotein) Gene from Multidrug-Resistant Human-Cells. *Cell* 1986, 47:381-389.
15. Gros P, Croop J, Housman D: Mammalian Multidrug Resistance Gene - Complete Cdna Sequence Indicates Strong Homology to Bacterial Transport Proteins. *Cell* 1986, 47:371-380.
16. Kolaczowski M, vanderRest M, CybularzKolaczowska A, Soumillion JP, Konings WN, Goffeau A: Drugs, ionophoric peptides, and steroids as substrates of the yeast multidrug transporter Pdr5p. *Journal of Biological Chemistry* 1996, 271:31543-31548.
17. Kolaczowski M, Kolaczowska A, Luczynski J, Witek S, Goffeau A: In vivo characterization of the drug resistance profile of the major ABC transporters and other components of the yeast pleiotropic drug resistance network. *Microbial Drug Resistance- Mechanisms Epidemiology and Disease* 1998, 4:143-158.
18. Leonard PJ, Rathod PK, Golin J: Loss of Function Mutation in the Yeast Multiple-Drug Resistance Gene Pdr5 Causes A Reduction in Chloramphenicol Efflux. *Antimicrobial Agents and Chemotherapy* 1994, 38:2492-2494.
19. Golin J, Ambudkar SV, May L: The yeast Pdr5p multidrug transporter: How does it recognize so many substrates? *Biochemical and Biophysical Research Communications* 2007, 356:1-5.
20. Piper P, Mahe Y, Thompson S, Pandjaitan R, Holyoak C, Egner R, Muhlbauer M, Coote P, Kuchler K: The Pdr12 ABC transporter is required for the development of weak organic acid resistance in yeast. *Embo Journal* 1998, 17:4257-4265.

21. Cannon RD, Lamping E, Holmes AR, Niimi K, Baret PV, Keniya MV, Tanabe K, Niimi M, Goffeau A, Monk BC: Efflux-Mediated Antifungal Drug Resistance. *Clin Microbiol Rev* 2009, 22:291-321.
22. De Hertogh B, Hancy F, Goffeau A, Baret PV: Emergence of species-specific transporters during evolution of the Hemiascomycete phylum. *Genetics* 2006, 172:771-781.
23. Gbelska Y, Krijger JJ, Breunig KD: Evolution of gene families: the multidrug resistance transporter genes in five related yeast species. *FEMS Yeast Research* 2006, 6:345-355.
24. Gaur M, Choudhury D, Prasad R: Complete inventory of ABC proteins in human pathogenic yeast, *Candida albicans*. *Journal of Molecular Microbiology and Biotechnology* 2005, 9:3-15.
25. Wilcox LJ, Balderes DA, Wharton B, Tinkelenberg AH, Rao G, Sturley SL: Transcriptional profiling identifies two members of the ATP-binding cassette transporter superfamily required for sterol uptake in yeast. *Journal of Biological Chemistry* 2002, 277:32466-32472.
26. Berbee ML, Taylor JW: Fungal molecular evolution: Gene trees and geologic time. *Mycota VII: Systematics and Evolution, Pt B* 2001, 7:229-245.
27. Kurtzman CP: Phylogenetic circumscription of *Saccharomyces*, *Kluyveromyces* and other members of the Saccharomycetaceae, and the proposal of the new genera *Lachancea*, *Nakaseomyces*, *Naumovia*, *Vanderwaltozyma* and *Zygorulasporea*. *FEMS Yeast Research* 2003, 4:233-245.
28. Kurtzman CP, Robnett CJ: Phylogenetic relationships among yeasts of the '*Saccharomyces* complex' determined from multigene sequence analyses. *FEMS Yeast Research* 2003, 3:417-432.
29. Souciet JL, Dujon B, Gaillardin C, Johnston M, Baret PV, Cliften P, Sherman DJ, Weissenbach J, Westhof E, Wincker P, Jubin C, Poulain J, Barbe V, Segurence BG, Artiguenave F, Anthouard V, Vacherie B, Val ME, Fulton RS, Minx P, Wilson R, Durrens P, Jean G, Marck C, Martin T, Nikolski M, Rolland T, Seret ML, Casaregola S, Despons L, Fairhead C, Fischer G, Lafontaine I, Leh V, Lemaire M, de Montigny J, Neugeglise C, Thierry A, Blanc-Lenfle I, Bleykasten C, Diffels J, Fritsch E, Frangeul L, Goeffon A, Jauniaux N, Kachouri-Lafond R, Payen C, Potier S, Pribylova L, Ozanne C, Richard GF, Sacerdot C, Straub ML, Talla E: Comparative genomics of protoploid Saccharomycetaceae. *Genome Res* 2009 in press.

30. Byrne KP, Wolfe KH: The Yeast Gene Order Browser: Combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res* 2005, 15:1456-1461.
31. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES: Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 2003, 423:241-254.
32. Thevenieau F, Le Dall MT, Nthangeni B, Mauersberger S, Marchal R, Nicaud JM: Characterization of *Yarrowia lipolytica* mutants affected in hydrophobic substrate utilization. *Fungal Genetics and Biology* 2007, 44:531-542.
33. Nakayama H, Tanabe K, Bard M, Hodgson W, Wu S, Takemori D, Aoyama T, Kumaraswami NS, Metzler L, Takano Y, Chibana H, Niimi M: The *Candida glabrata* putative sterol transporter gene CgAUS1 protects cells against azoles in the presence of serum. *Journal of Antimicrobial Chemotherapy* 2007, 60:1264-1272.
34. Fairhead C, Dujon B: Structure of *Kluyveromyces lactis* subtelomeres: duplications and gene content. *FEMS Yeast Research* 2006, 6:428-441.
35. Prasad R, Dewergifosse P, Goffeau A, Balzi E: Molecular-Cloning and Characterization of A Novel Gene of *Candida albicans*, Cdr1, Conferring Multiple Resistance to Drugs and Antifungals. *Current Genetics* 1995, 27:320-329.
36. DelSorbo G, Andrade AC, VanNistelrooy JGM, Vankan JAL, Balzi E, DeWaard MA: Multidrug resistance in *Aspergillus nidulans* involves novel ATP-binding cassette transporters. *Molecular & General Genetics* 1997, 254:417-426.
37. Iwaki T, Giga-Hama Y, Takegawa K: A survey of all 11 ABC transporters in fission yeast: two novel ABC transporters are required for red pigment accumulation in a *Schizosaccharomyces pombe* adenine biosynthetic mutant. *Microbiology-Sgm* 2006, 152:2309-2321.
38. Turi TG, Rose JK: Characterization of A Novel *Schizosaccharomyces pombe* Multidrug-Resistance Transporter Conferring Brefeldin-A Resistance. *Biochemical and Biophysical Research Communications* 1995, 213:410-418.
39. Nakaune R, Adachi K, Nawata O, Tomiyama M, Akutsu K, Hibi T: A novel ATP-binding cassette transporter involved in multidrug resistance in the phytopathogenic fungus *Penicillium digitatum*. *Applied and Environmental Microbiology* 1998, 64:3983-3988.

40. Verrier PJ, Bird D, Buria B, Dassa E, Forestier C, Geisler M, Klein M, Kolukisaoglu U, Lee Y, Martinoia E, Murphy A, Rea PA, Samuels L, Schulz B, Spalding EJ, Yazaki K, Theodoulou FL: Plant ABC proteins - a unified nomenclature and updated inventory. *Trends in Plant Science* 2008, 13:151-159.
41. Sanchez-Fernandez R, Davies TGE, Coleman JOD, Rea PA: The *Arabidopsis thaliana* ABC protein superfamily, a complete inventory. *Journal of Biological Chemistry* 2001, 276:30231-30244.
42. Crouzet J, Trombik T, Frayssé AS, Boutry M: Organization and function of the plant pleiotropic drug resistance ABC transporter family. *FEBS Letters* 2006, 580:1123-1130.
43. Ullman B: Multidrug-Resistance and P-Glycoproteins in Parasitic Protozoa. *Journal of Bioenergetics and Biomembranes* 1995, 27:77-84.
44. Dean M, Hamon Y, Chimini G: The human ATP-binding cassette (ABC) transporter superfamily. *Journal of Lipid Research* 2001, 42:1007-1017.
45. Merico A, Sulo P, Piskur J, Compagno C: Fermentative lifestyle in yeasts belonging to the *Saccharomyces* complex. *Febs Journal* 2007, 274:976-989.
46. Wapinski I, Pfeffer A, Friedman N, Regev A: Automatic genomewide reconstruction of phylogenetic gene trees. *Bioinformatics* 2007, 23:I549-I558.
47. Nikolski M, Sherman DJ: Family relationships: should consensus reign?-consensus clustering for protein families. *Bioinformatics* 2007, 23:e71-e76.
48. Durrens P, Sherman DJ: A systematic nomenclature of chromosomal elements for hemiascomycete yeasts. *Yeast* 2005, 22:337-342.
49. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD: Multiple sequence alignment with the Clustal series of programs. *Nucl Acids Res* 2003, 31:3497-3500.
50. Clamp M, Cuff J, Searle SM, Barton GJ: The Jalview Java alignment editor. *Bioinformatics* 2004, 20:426-427.
51. Edgar RC: MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 2004, 5:1-19.
52. Felsenstein J: PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* 1989, 5:164-166.

53. Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R: Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics* 2007, 8:460.
54. Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, de Montigny J, Marck C, Neuveglise C, Talla E, Goffard N, Frangeul L, Aigle M, Anthouard V, Babour A, Barbe V, Barnay S, Blanchin S, Beckerich JM, Beyne E, Bleykasten C, Boisrame A, Boyer J, Cattolico L, Confanioleri F, de Daruvar A, Despons L, Fabre E, Fairhead C, Ferry-Dumazet H, Groppi A, Hantraye F, Hennequin C, Jauniaux N, Joyet P, Kachouri R, Kerrest A, Koszul R, Lemaire M, Lesur I, Ma L, Muller H, Nicaud JM, Nikolski M, Oztas S, Ozier-Kalogeropoulos O, Pellenz S, Potier S, Richard GF, Straub ML, Suleau A, Swennen D, Tekaia F, Wesolowski-Louvel M, Westhof E, Wirth B, Zeniou-Meyer M, Zivanovic I, Bolotin-Fukuhara M, Thierry A, Bouchier C, Caudron B, Scarpelli C, Gaillardin C, Weissenbach J, Wincker P, Souciet JL: Genome evolution in yeasts. *Nature* 2004, 430:35-44.
55. Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, Steiner S, Mohr C, Pohlmann R, Luedi P, Choi S, Wing RA, Flavier A, Gaffney TD, Philippsen P: The *Ashbya gossypii* Genome as a Tool for Mapping the Ancient *Saccharomyces cerevisiae* Genome. *Science* 2004, 304:304-307.
56. Gattiker A, Rischatsch R, Demougin P, Voegeli S, Dietrich FS, Philippsen P, Primig M: Ashbya Genome Database 3.0: a crossspecies genome and transcriptome browser for yeast biologists. *BMC Genomics* 2007, 8:9.

ADDITIONAL MATERIAL

ADDITIONAL FILE 1 - LIST OF PDRP FEATURES INCLUDED IN THIS PAPER.

Sequences of *S. cerevisiae* [12], *C. glabrata*, *K. lactis*, *D. hansenii*, *Y. lipolytica* [54] and *E. gossypii* [55] were previously published. *Z. rouxii* and *K. thermotolerans* have both been sequenced by Génoscope, and *S. kluyveri* was sequenced by the Washington University Genome Sequencing Centre. The three new genomes ZYRO, KLTH, SAKL, plus KLLA have been entirely annotated and manually curated by the Génolevures consortium [29]. ERGO annotation is by GATTIKER [56].

This file is reproduced below as an image. You may download the original PDF file at the following address:

<http://www.biomedcentral.com/content/supplementary/1471-2164-10-459-S1.PDF>

GL3 name	Abbreviated name	Subtelomeric	Phylogenetic cluster	SONS	Length	N-terminal Walker A	C-terminal ABC sig.
SACE0P04972g	PDR12		A	a	1511	VGRPGAGCSTFL	NVEQRKKLSIGV
CAGL0M07293g	CAGL M07293		A	a	1515	VGRPGAGCSTLL	NVEQRKKLSIGV
ZYRO0F08866g	ZYRO F08866		A	a	1520	VGRPGAGSSTLL	NVEQRKKLSIGV
ZYRO0F08888g	ZYRO F08888		A	a	1520	VGRPGAGCSTLL	NVEQRKKLSIGV
SAKL0C05654g	SAKL C05654		A	a	1509	VGRPGAGCSTLL	NVEQRKKLSIGV
SAKL0H10670g	SAKL H10670		A	a	1509	VGRPGAGCSTLL	NVEQRKKLSIGV
KLLA0B09702g	KLLA B09702		A		1518	VGRPGAGCSTLL	NVEQRKKLSIGV
YALI0F17996g	YALI F17996		A		1508	LGRPGSGCSTFL	NVEQRKKLSIGV
SACE0N08954g	YNR070W	St	B		1333	LGRPGAGCTSFL	NVEQRKKLSIGV
SACE0D05478g	SNQ2		B	b	1501	LGRPGAGCSSFL	NVEQRKKLSIGV
CAGL0I04862g	CAGL I04862		B	b	1507	LGRPGAGCSSML	NVEQRKKLSIGV
ZYRO0B14762g	ZYRO B14762		B		1515	LGRPGSGCSSFL	NVEQRKKLSIGV
ZYRO0A04114g	ZYRO A04114		B	b	1518	LGRPGSGCSSFL	NVEQRKKLSIGV
KLTH0A01914g	KLTH A01914		B	b	1499	LGRPGAGCSSFL	NVEQRKKVLSIGV
SAKL0H21010g	SAKL H21010		B		1447	LGRPGAGCSSFL	NVEQRKKLSIGV
SAKL0C11616g	SAKL C11616		B	b	1484	LGRPGAGCSSFL	NVEQRKKLSIGV

KLLA0D03432g	KLLA D03432	B	b	1483	LGRPGSGCSTFL	NVEQRKKLSIAV
ERGO0B08140g	ERGO B08140	B	b	1488	LGRPGAGCTSL	NVEQRKKLSIGV
DEHA2A03696g	DEHA A03696	B		1477	LGRPGAGCSSLL	NVEQKKKLSIGV
SACE0O11000g	PDR10	C		1564	LGRPGAGCTTLL	NVEQRKRLTIGV
SACE0D14146g	PDR15	C	c	1529	LGRPGSGCTTLL	NVEQRKRLTIGV
SACE0O07216g	PDR5	C	c	1511	LGRPGSGCTTLL	NVEQRKRLTIGV
CAGL0M01760g	CAGL M01760	C		1499	LGRPGSGCTTLL	NVEQRKRLTIGV
CAGL0F02717g	CAGL F02717	C	c	1542	LGRPGSGCTTLL	NVEQRKRLTIGV
ZYRO0D17710g	ZYRO D17710	St	C	1486	LGRPGSGCTTLL	NVEQRKRLTIGV
ZYRO0D11836g	ZYRO D11836	C	c	1462	LGRPGSGCTTLL	NVEQRKRLTIGV
ZYRO0D11858g	ZYRO D11858	C	c	1503	LGRPGSGCTTLL	NVEQRKRLTIGV
ZYRO0D11880g	ZYRO D11880	C	c	1498	LGRPGSGCTTLL	NVEQRKRLTIGV
KLTH0G19448g	KLTH G19448	C	g	1486	LGRPGSGCSTLL	NVEQRKRLTVGV
SAKL0C06996g	SAKL C06996	C		1545	LGRPGAGCSTLL	NVEQRKRLTIGV
SAKL0C11704g	SAKL C11704	C	b	1510	LGRPGAGCSTLL	NVEQRKRLTIGV
SAKL0G08008g	SAKL G08008	C	f	1549	LGRPGSGCSTLL	NVEQRKRLTIGV
KLLA0D03476g	KLLA D03476	C	b	1560	LGRPGAGCSTLL	NVEQRKRLTIGV
KLLA0F21692g	KLLA F21692	C	f	1525	LGRPGSGCTTLL	NVEQRKRLTIGV
ERGO0G05126g	ERGO G05126	C		1497	LGRPGAGCSTLL	NVEQRKRLTIGV
ERGO0B08162g	ERGO B08162	C	b	1512	LGRPGAGCSTLL	NVEQRKRLTIGV
DEHA2B16610g	DEHA B16610	St	C	1508	LGRPGSGCSTLL	NVEQRKRLTIGV
DEHA2F16478g	DEHA F16478	C		1500	LGRPGSGCSTLL	NVEQRKRLTIGV
DEHA2G14894g	DEHA G14894	C		1481	LGRPGAGCSTFL	NVEQRKRLTIGV
DEHA2G14916g	DEHA G14916	C		1463	LGRPGAGCSTFL	NVEQRKRLTIGV
YALI0B02544g	YALI B02544	C		1485	LGRPGSGCTTLL	NVEQRKRLTIGV
YALI0B12980g	YALI B12980	C		1472	LGRPGSGCTTFL	NVEQRKRLTIGV
YALI0C20265g	YALI C20265	C		1469	LGRPGSGCTTFL	NVEQRKRLTIGV
YALI0E14729g	YALI E14729	C		1659	LGRPGSGCTTFL	NVEQRKRLTIGV

B. Homologs of YOL075c

SACE0O02112g	YOL075C	D	d	1294	MGGSGSGKTTLL	SGGEKRRVTMGV
CAGL0I08019g	CAGL I08019	D	d	1285	MGGSGSGKTTLL	SGGEKRRVSMGI
ZYRO0B05588g	ZYRO B05588	D	d	1270	MGGSGSGKTTLL	SGGEKRRVTIGV
KLTH0E15796g	KLTH E15796	D	d	1247	VGGSGSGKTTLL	SGGEKRRVSMGT
SAKL0C08074g	SAKL C08074	D	d	1257	VGGSGSGKTTLL	SGGEKRRVSLGV
KLLA0C04477g	KLLA C04477	D	d	1253	IGGSGSGKTTML	SGGEKRRVSI
YALI0D25828g	YALI D25828	D		1328	LGGSGSGKTSLL	SGGEKRRVSICV

C. Homologs of AUS1 and Pdr11

SACE0I03608g	PDR11	E		1411	LGNP----TSAL	SPTQRKLLSIGV
SACE0O04180g	AUS1	E	e	1394	LGYP----TSTL	NPTQRKLLSIGV
CAGL0F01419g	CAGL F01419	E	e	1398	LGAP----TSGI	NPTQRKLLSIGV
SAKL0F04312g	SAKL F04312	E		1338	NYKD----SSIL	NPAEKKLLSIGV
ERGO0F13266g	ERGO F13266	E		1309	LVVG----DKSS	CLQEKRILSIGP

D. Fragments and pseudogenes

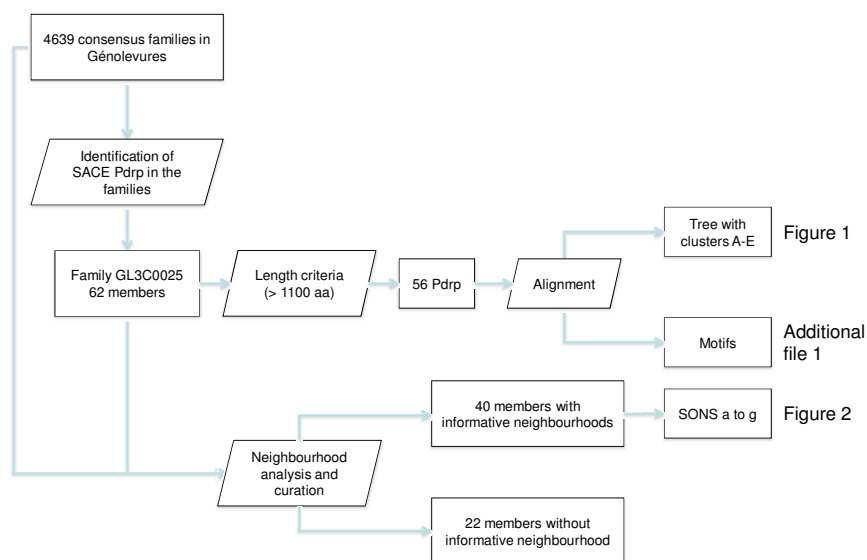
KLTH0E17138g	KLTH E17138	St	g	405	/	-
KLTH0H00110g	KLTH H00110	St	g	660	LGRPGSGCTTLL	
SAKL0D15356g	SAKL D15356	St		72		
DEHA2C03784g	DEHA C03784			626	/	-
DEHA2F00154g	DEHA F00154	St		1099	-	NVEQRKRLTIGV
DEHA2F16500g	DEHA F16500			79	-	/

ADDITIONAL FILE 2- SCHEMA OF ANALYSIS TOOLS AND DATABASES.

Rectangles represent databases and results while parallelograms represent tools used to obtain them.

This file is reproduced below as an image. You may download the original PPT file at the following address:

<http://www.biomedcentral.com/content/supplementary/1471-2164-10-459-S2.PPT>

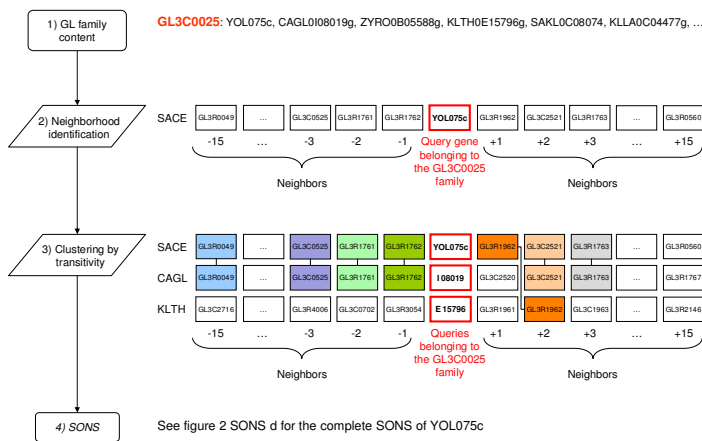


ADDITIONAL FILE 3 - STEPS OF THE IDENTIFICATION OF ORTHOLOGS BY NEIGHBORHOOD AND SIMILARITY METHOD.

(1) The building of SONS (Subset of Orthologs defined by Neighborhood and Similarity) begins by taking a family of genes whose translation products show sequence similarity for example, the family GL3C0025. (2) For each query gene of this family we identify 15 genes, which we call neighbors, on each side and especially the family to which these neighbors belong. (3) Two query genes whose translation products belong to the same family (homologs by similarity) are members of the same SONS if they share at least one pair of neighbors that are also homologous to each other by similarity. These neighbors are highlighted in the same color. In the example, YOL075c shares some neighbors with the query gene of CAGL and has one neighbor in common with the query gene of KLTH. CAGL has no gene in common with KLTH, however YOL075c must be in the same SONS as the query genes of CAGL and KLTH. In consequence, the three query genes are clustered in the same SONS. Homologs that do not share a pair of homologous neighbors are separated into two distinct SONS. (4) See Figure 2 SONS d for the complete SONS of YOL075c.

This file is reproduced below as an image. You may download the original PPT file at the following address:

<http://www.biomedcentral.com/content/supplementary/1471-2164-10-459-S3.PPT>



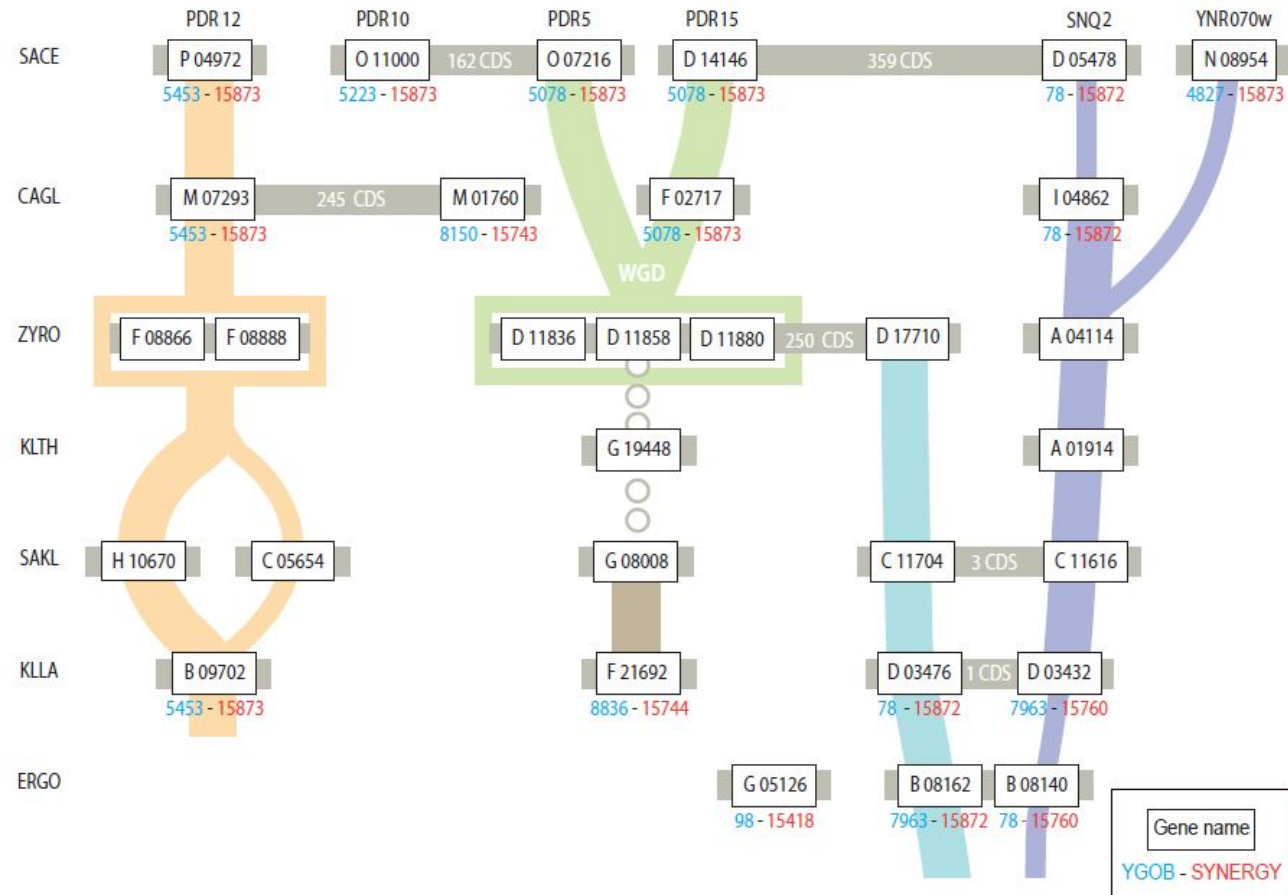
ADDITIONAL FILE 4 - RELATIONSHIPS BETWEEN THE PDRP SENSU STRICTO BASED ON THE NEIGHBOURHOOD ANALYSIS: COMPARISON WITH YGOB AND SYNERGY.

Each box represents a PDR. Grey lines join PDR genes located on the same chromosome. When non-adjacent, the number of coding sequences between two PDR genes appears in white. Shared neighborhoods are represented by colored connectors. A large connector links PDR genes sharing more than one common neighbor while a thin connector links PDR genes sharing one neighbor only. The numbers shown in blue is the number of the pillar in which the PDR gene has been classified on the Yeast Gene Order Browser [30]*. The number shown in red is the number of the orthogroup in which the PDR gene has been classified by Synergy [46].

This file is reproduced below as an image. You may download the original PDF file at the following address:

<http://www.biomedcentral.com/content/supplementary/1471-2164-10-459-S4.PDF>

* The misassignment of the gene KLLA0D03436g to the evolutive story of SNQ2 done by the YGOB has been corrected since version v3 ((Gordon et al., 2009). Nowadays the YGOB agrees with IONS to consider the gene KLLA0D03432g (rather than KLLA0D03436g) as the ortholog of SNQ2.



IX. EXTENSION DE LA CLASSIFICATION

DES TRANSPORTEURS AUX CINQ

NOUVELLES ESPÈCES SÉQUENCÉES

CLASSIFICATION DES TRANSPORTEURS

La classification des transporteurs pose deux questions essentielles : qu'est-ce qui permet de déterminer qu'une protéine est un transporteur et qu'entend-on exactement par « transporteurs » ?

1. QU'EST-CE QUI PERMET DE DÉTERMINER QU'UNE PROTÉINE EST UN TRANSPORTEUR ?

En dehors des cas d'études expérimentales en laboratoire permettant directement d'identifier un processus de transport, la prédiction du rôle de transporteur d'une protéine est basée sur une curation manuelle. Celle-ci s'appuie fréquemment sur les prédictions de motifs, en particulier l'identification des traverses transmembranaires (ou TMS), et/ou sur la similarité avec une protéine de transport précédemment caractérisée expérimentalement.

Au niveau de la prédiction des TMS, on notera cependant que, d'une part, les protéines présentant des TMS ne sont pas toutes des transporteurs : près d'un tiers des protéines membranaires présentant des TMS sont impliquées dans d'autres fonctions (Von Heijne, 2007) ; et que, d'autre part, certaines protéines appartenant à un complexe impliqué dans le transport transmembranaires ne possèdent pas de TMS. De plus, les différentes méthodes de prédictions de TMS, ne prédisent pas toujours le même nombre de TMS. La prédiction d'un grand nombre de TMS dans la structure de la protéine (≥ 6 TMS) est considérée comme un bon indicateur du rôle de transporteur d'une protéine. Par contre, une prédiction d'un faible nombre de TMS voir une absence de prédiction ne permet pas d'exclure l'implication d'une protéine dans un système de transport. C'est pourquoi ce type d'analyse nécessite une expertise manuelle ou d'être combinée avec d'autres approches, comme la similarité de séquence avec des transporteurs connus, la prédiction d'autres types de motifs caractéristiques de certaines familles de transporteurs, la prédiction de structure tridimensionnelle,...

La détermination du rôle de transporteur d'une protéine basée sur la similarité avec une protéine de transport précédemment caractérisée expérimentalement présente l'avantage d'être plus facilement automatisable. Elle présente néanmoins l'inconvénient de ne pas permettre l'identification de transporteur d'un nouveau type.

2. QU'ENTEND-ON EXACTEMENT PAR « TRANSPORTEURS » ?

Dans la littérature, lorsqu'on parle de transporteurs, certains se limitent aux transporteurs encodés dans le génome nucléaire, d'autres prennent également en compte les transporteurs mitochondriaux. Le statut des protéines qui facilitent le transport sans y participer directement est également un peu particulier. Doivent-elles être considérées comme des transporteurs ou non ? L'avantage du système de classification TC est qu'il prend en compte ces différences en attribuant des digits et permet de tenir compte des niveaux de certitudes différents. Les homologues séquencés dont la fonction ou le mécanisme de transport est inconnu et les transporteurs caractérisés fonctionnellement mais dont la séquence n'est pas disponible sont inclus temporairement dans une classe distincte, la classe 9, jusqu'à leur caractérisation complète ou leur séquençage. Une classe supplémentaire (la classe 8) est réservée aux protéines auxiliaires de transport, c'est-à-dire aux protéines qui d'une certaine façon facilitent le transport à travers une membrane sans y participer directement. Par exemple, elles peuvent avoir une fonction liée au couplage d'énergie nécessaire pour le transport, jouer un rôle dans la formation de la structure d'un complexe ou dans la régulation. Les digits des classes 1 à 5 correspondent aux transporteurs caractérisés.

EXTENSION DE LA CLASSIFICATION

La base de données exhaustive des transporteurs réalisée pour cinq espèces d'hémiascomycètes par De Hertogh *et al.* (2006) en suivant le système de classification hiérarchique TC est ici étendue à cinq espèces supplémentaires nouvellement séquencées par le Consortium Génolevures. La méthode utilisée pour l'extension de cette classification, qui s'appuie sur les familles Génolevures ou les sous-ensembles d'homologues obtenus par la méthode IONS, et les résultats obtenus sont présentés ci-dessous sous forme d'un article, soumis à *FEMS Yeast Research* le 30 mai 2012 et en révisions mineures.

Mes contributions personnelles à cet article :

- *Vérification du principe de base de la méthode développée par Inês Azevedo Isidro dans le cadre de son mémoire*
- *Corrections des scripts MySQL et application à une base de données reprenant les génomes de 10 espèces de levures*
- *Comparaison avec les résultats de De Hertogh et al. (2006)*
- *Rédaction*

METHODOLOGY FOR THE CLASSIFICATION OF TRANSPORTER PROTEINS AND ITS APPLICATION TO TEN HEMIASCOMYCETE YEASTS

Marie-Line Seret¹, Inès Isidro² and Philippe V. Baret¹

¹ Université catholique de Louvain, Earth and Life Institute (ELI), 1348 Louvain-la-Neuve, Belgium.

² Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisbon, Portugal

ABSTRACT

We built a database of transporter proteins from ten hemiascomycete species: *Saccharomyces cerevisiae*, *Candida glabrata*, *Zygosaccharomyces rouxii*, *Kluyveromyces thermotolerans*, *Saccharomyces kluyveri*, *Kluyveromyces lactis*, *Eremothecium gossypii*, *Debaryomyces hansenii*, *Arxula adeninivorans* and *Yarrowia lipolytica*. The classification is based on the hierarchical classification of the Transport Classification database (TCDB) and is performed by controlled propagation of the classification to homologous genes within a reference clustering. The database contains a total of 5,171 proteins classified in 94 transporter families. Most of them are ubiquitous (75.5%) containing members in all species. The percentages of transporter proteins are relatively similar in each genome (from 8.32 to 9.8%). The proposed method can be of used for the propagation of functional information in other genomes.

KEYWORDS

Transmembrane transporters, yeasts, hemiascomycete, hierarchical classification

INTRODUCTION

Transport proteins fulfill essential functions in living cells by allowing comings and goings of metabolites, ions and macromolecules across membranes of all organisms. They constitute more or less 10% of most yeast proteomes. Attempts to provide (partial) inventories of transporter proteins (Andre, 1995; Nelissen *et al.*, 1995) did not wait the completion of sequencing of the complete genome of *Saccharomyces cerevisiae* (Goffeau *et al.*, 1996) or just followed it (Andre, 1995; Nelissen *et al.*, 1997; Paulsen *et al.*, 1998). These first inventories, only included transporters with at least six transmembrane spans.

A first attempt to provide a complete database of transporters and other membrane proteins of *Saccharomyces cerevisiae* was made by manual curation (De Hertogh *et al.*, 2002), on the basis of functional and phylogenetic criteria, and was designed as an extension of the Transporter Classification (TC) system (Busch & Saier, 2002).

The TC system (www.tcdb.org) is a hierarchical classification with five levels that uses both functional and phylogenetic criteria. A TC identification number (TC ID) has five components: V.W.X.Y.Z. V corresponds to the transporter class (i.e., channel, carrier (porter)); W corresponds to the transporter subclass or energy source used to drive transport; X corresponds to the transporter family (or superfamily sometimes); Y corresponds to the subfamily in which a transporter is found, and Z corresponds to a specific transporter with a particular substrate or range of substrates transported (Busch & Saier, 2002). If multiple dissimilar subunits comprise a transport system, all are listed under a single 5 digit TC ID.

TCDB is a representative database of transporters, which means that all the functionally characterized transport systems are not included. For example, characterized orthologues with the same function are not included. Sequenced homologues of unknown function are not assigned a TC identification number (TC ID) neither (<http://www.tcdb.org/faq.php#tc-system>).

Genes of the TCDB as well as its hierarchical classification already served as starting point to build a comprehensive database of transporters. The system was called YETI for Yeast Transporter Information. This database was built in a semi-automated manner and contained 2,480 transporters (De Hertogh *et al.*, 2006) encoded by the genomes of five hemiascomycete species, namely *S. cerevisiae*, *C. glabrata*, *K. lactis*, *D. hansenii* and *Y. lipolytica*.

With the exponential increase in the number of genome sequencing projects, we are facing a sea of data from which biological meaningful information needs to be carefully distilled. However, it is neither feasible to proceed to a manually curated annotation nor even less to perform a functional assay for every new gene sequence.

The objective of this work is to update and expand the YETI database of yeast transporters by adding new proteins from five phylogenetically related hemiascomycetous yeasts, based on homology-based propagation of the annotation within a reference clustering. We choose to assign genes to the fourth digit of the TC classification, as the degree of similarity between homologous genes is not sufficient to ensure that the encoded proteins transport the same particular substrate or range of substrates; indeed previous studies showed that a single substitution can modify substrate specificity. For example, the modification of the Val216 of isomaltase of *S. cerevisiae* to the maltase-type amino acid (Thr) modified its substrate specificity from isomaltose to maltose (Yamamoto *et al.*, 2004).

MATERIAL AND METHOD

THE GÉNOLEVURES DATA

The Génolevures' database contains the CDS of ten completely sequenced species, for a total of 55,782 genes among which 51,298 were classed in 5,482 GL families.

The coding sequences translated into amino-acids were retrieved in FASTA format from the Génolevures public website for *Candida glabrata*, *Zygosaccharomyces rouxii*, *Kluyveromyces thermotolerans*, *Saccharomyces kluyveri*, *Kluyveromyces lactis*, *Debaryomyces hansenii* and *Yarrowia lipolytica* (<http://www.genolevures.org/download.html>, release date 2008/09/10). The amino-acids sequences of *Eremothecium gossypii*, *Arxula adeninivorans* and *Saccharomyces cerevisiae* were retrieved from the Génolevures private website as well as the classification in families (April 8th, 2010).

UPDATE OF THE GENE CONTENT OF THE YETI DATABASE

The YETI database provided in supplementary data 1 of the publication by de Hertogh *et al.* (De Hertogh *et al.*, 2006) contained 2,480 transmembrane transporters. Due to modifications in databases (see table 1), only 2,384 of them are useable to search for transporters in the other species.

Table 1. Update of the gene content of the YETI database, due to modifications in databases since 2005.

Categories of genes	Number of genes
<i>YETI database by de Hertogh et al.</i>	2,480
Identified as pseudogenes	-14
Merged with another gene or pseudogene	-2
Disappeared following the re-annotation of DEHA	-45
Not found on the Génolevures website	-27
Not classed in Génolevures families	-8
<i>YETI genes classed in GL families and used</i>	2,384*
<i>Additional genes from TCDB</i>	34
<i>Total</i>	2,418*

* The names of a part of those genes were modified due to the re-annotation of *Debaryomyces hansenii* as well as the re-annotation of chromosome E of *Kluyveromyces lactis*.

34 hemiascomycetous genes identified subsequently as transporters on the TCDB website and belonging to genomes studied by the Génolevures consortium were added to the database of transporters, raising the total number of genes potentially useable for the update to 2,418.

UPDATE OF THE TC ID OF THE YETI DATABASE

The TC ID of genes of the YETI database was updated according to the modifications made in the TC database following the procedure explained below.

The procedure to update the TC ID of genes of the YETI database (see figure 1) may be divided in three major steps corresponding to the relationship of each gene of the YETI database with genes of the TC database:

- 1) For each gene included in the TC database, if its TC ID in the YETI database is different of the one mentioned in the updated version of the TCDB database, its TC ID in the YETI database is modified accordingly.
- 2) For the genes that are not included in the TC database, we look at their Génolevures family or at the subset obtained by the IONS method (Seret & Baret, 2011) which cluster genes found in chromosomal regions of common evolutionary ancestry, based on chromosomal neighborhood and on similarity.

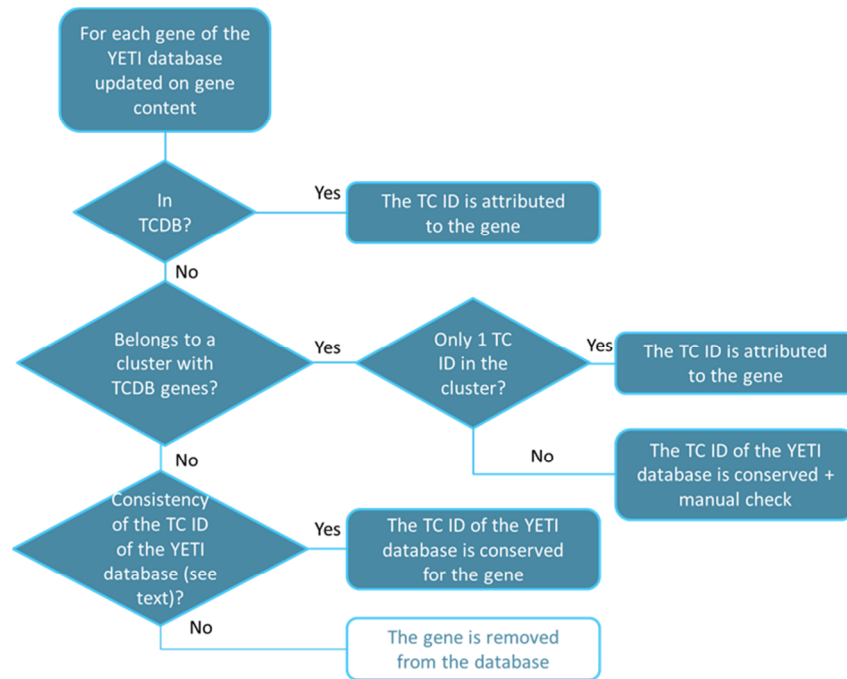


Figure 1. Flowchart of the different steps of the update of the TC ID of genes of the YETI database.

First, we insure that each cluster (Génolevures family or IONS subset if the family points to more than one TC ID) contains genes from maximum one TC ID.

If they belong to a cluster that includes gene(s) from a single TC ID, this TC ID is attributed to all YETI genes of the cluster.

If they belong to a cluster that includes genes belonging to several TC IDs, the TC ID attributed in the YETI database is conserved and a manual check is performed to ensure that there is no contradiction.

3) For the genes that belong to a cluster that does not include any gene of the TC database, a check on the TC ID attributed in the YETI database is made to ensure that no gene belonging to this TC ID in the old YETI database was attributed a new TC ID after following the rules defined in the previous steps and so that there is no ambiguities on the TC ID to assign. If the TC ID passes the check, the TC ID attributed in the YETI database is conserved for those genes. In the other case, those genes are removed from the updated YETI database.

After this update, The YETI database contains 2,096 genes belonging to 267 Génolevures families with a 3rd-digit level of classification, among which 2,044 are also classified at the 4th-digit level (see table 2).

Table 2. Number of genes treated by the 3 major steps of the procedure to update the TC ID of genes of the YETI database.

The penultimate line indicates the total number of genes useable to search for transporters in the other species, while the last line gives the number of genes removed from the database due to ambiguities on the TC ID to attribute them.

Categories of genes	Genes number at the 3 rd TC digit	Genes number at the 4 th TC digit
1. Genes included in the TC database		
- Same TC ID in TCDB and YETI	274	220
- YETI ID adapted to the TC ID	20	74
- Genes not included in YETI in 2002	34	34
2. Genes from clusters including genes from the TC database		
- YETI ID correspond to the TC ID of the genes included in the TCDB database	1,068	772
- YETI ID adapted to the TC ID of the genes included in the TCDB database	100	297
- YETI ID conserved (more than one TC ID)	0	93
3. Genes from clusters not including genes from the TC database		
YETI ID conserved (consistent)	600	554
Total in the updated YETI database	2,096	2,044
Genes removed from the YETI database (YETI ID not consistent)	322 (2,418-2,096)	374 (2,418-2,044)

AUTOMATIC ASSIGNMENT METHOD

The objective was to develop a method to allow the automatic assignment of a TC ID to genes of the five new yeast species available on the Génolevures website (Figure 2).

The method relies on the establishment of correspondence between TC ID attributed in the YETI database and Génolevures families or IONS subset, when the Génolevures family contains genes from more than one TC ID.

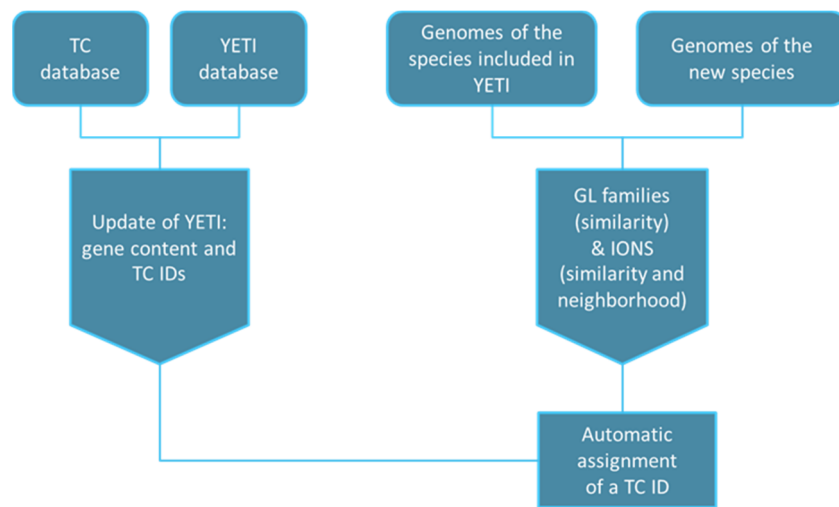


Figure 2. Steps for the automatic assignment of a TC ID.

The rationale of the method is that when all the proteins of a particular cluster of genes (Génolevures families or IONS subset) which are already included in the YETI database are classified in the same TC ID, all the proteins of this cluster of genes are assigned to this TC ID (Fig. 3 a and b). The assignment is independently established for the 3rd and 4th-digit level.

Among the 5,173 genes belonging to Génolevures families, which include at least one gene in the updated YETI database, the automatic assignment method allowed to rise the assignments at the 3rd digit to 5,171 genes (99,96%), and to 4,781 genes (92,42%) at the 4th digit (the sum of 328, 1,716 and 2,737, see table 3) .

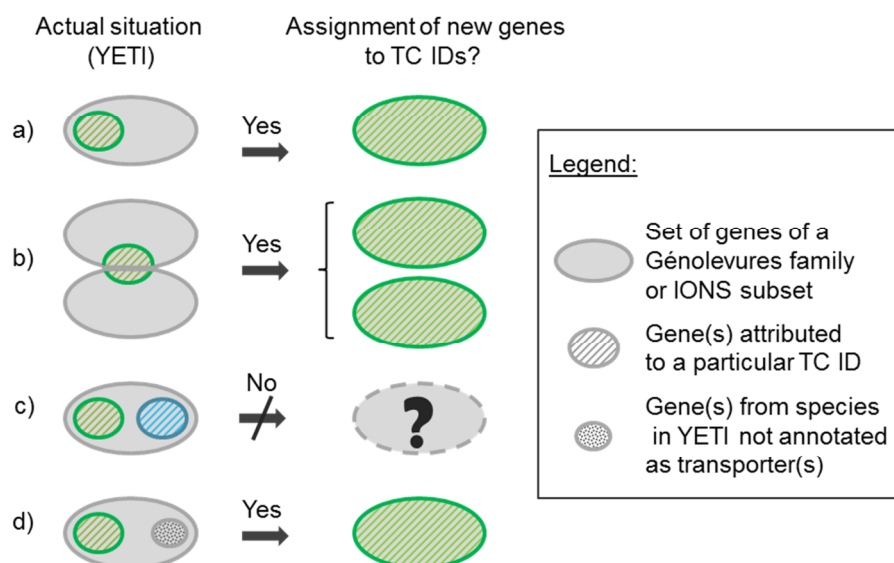


Figure 3. Working basis of the automatic assignment method.

(a and b) The Génolevures family or IONS subset is used as a reference clustering to expand the current YETI classification when all the proteins of a particular cluster of genes (Génolevures families or IONS subset) which are already included in the YETI database are classified in the same TC ID. (c) The cluster contains genes attributed to different TC IDs in the YETI database (represented by different colors). As the correspondence to a TC ID is dubious, the automatic assignment cannot be done. (d) The cluster includes genes annotated as transporters as well as gene(s) not annotated as transporter(s) belonging to species already included in the YETI database. The automatic assignment is made but the information is recorded.

In some cases (Fig. 3 c) the correspondence to a TC ID in the YETI database is dubious. Genes corresponding to these cases are considered as transporters but the automatic assignment cannot be done, or only partially. At the 3rd digit, two genes (ARAD0B20196g and YALIOF16093g) are in this situation. They belong to the same IONS subset, which does not contain any gene in the TC nor in the YETI database. However their Génolevures family contains two other IONS subsets: one containing YDR438w, classified as 2.A.7 in the TCDB, and the other one containing DEHA2B15422g, classified as 9.B.2 in YETI, so that the assignment cannot be done automatically. At the 4th digit, these cases concern 338 supplementary genes (Table 3).

In other cases (Fig. 3 d), the Génolevures family or IONS subset includes both YETI genes and non-YETI genes belonging to species already included in the YETI database. The automatic assignment is made but the information is recorded as it raises some questions since these genes may encode non-transport proteins and the method is based on the assumption that

Génolevures families including genes from the YETI database are made up of transport proteins.

RESULTS

Out of the 5,171 assigned genes, 328 were already classified in the TCDB and 1,768 in the updated YETI database at the 3rd or 4th digit (see table 3). So, our method allowed adding 3,075 genes to the classification (2,737 at the 4th digit and 338 at the 3rd digit only, Table 3). From these 3,075 genes nearly 20% belong to species already included in the YETI database, while 80% belong to the five newly added species. The number of additional transporters automatically assigned ranges from 82 (72+10) in *Saccharomyces cerevisiae*, which was already in the YETI database, to 521 in *Saccharomyces kluyveri*, a newly added species. These newly added transporters were classified into 93 distinct families (3rd digit) and 143 distinct subfamilies (4th digit).

The total number of assigned transporters (previously and with our method) ranges from 412 in *Eremothecium gossypii* to 617 in *Yarrowia lipolytica*, a difference of more than two hundred transporters. However when expressed in percentages of their genome size, assigned transporters diverge of a maximum of 1.48% from 8.32% in *S. cerevisiae* to 9.80% in *S. kluyveri* for identified transporters.

Table 3. Global statistics of transporter proteins currently in the TC and YETI databases and automatically assigned in ten Hemiascomycete species.

	SACE	CAGL	ZYRO	KLTH	SAKL	KLLA	ERGO	DEHA	ARAD	YALI	Total	
ORFs	6,660	5,201	4,991	5,092	5,318	5,076	4,715	6,270	6,012	6,447	55,782	
In the TC database (TCDB)	319	1	1	0	0	4	0	2	0	1	328	} 2,096
In the updated YETI database at the 4th digit (YETI_upd4)	144	337	0	0	0	367	0	421	0	447	1,716	
In the updated YETI database at the 3rd digit only (YETI_upd3)	9	4	0	0	0	9	0	11	0	19	52	
Additional transporters assigned at the 4th digit (Assigned_d4)	72	112	423	435	460	86	381	147	495	126	2,737	} 3,075
Additional transporters only assigned at the 3rd digit (Assigned_d3)	10	8	40	63	61	8	31	19	74	24	338	
<i>Additional transporters assigned in %</i>	1.23	2.31	9.28	9.78	9.80	1.85	8.74	2.65	9.46	2.33	5.51	
Total number of assigned transporters	554	462	464	498	521	474	412	600	569	617	5,171	
<i>Total number of assigned transporters in %</i>	8.32	8.88	9.30	9.78	9.80	9.34	8.74	9.57	9.46	9.57	9.27	
Additional unassigned transporters at the 3rd digit due to ambiguities on the TC ID (Unassigned)	0	0	0	0	0	0	0	0	1	1	2	
Total	554	462	464	498	521	474	412	600	570	618	5,173	

The abbreviation given in parenthesis refers to the one used in supplementary table H which gives the list of transporters and their classification.

Table 4. Transporter proteins in ten Hemiascomycete species per subclasses and per species.

	SACE	CAGL	ZYRO	KLTH	SAKL	KLLA	ERGO	DEHA	ARAD	YALI	Total	Mean $\pm$ SD
1.A: α -type channels	35	29	23	26	26	25+1	23	21+6	24	31+4	274	27.4 $\pm$ 4.40
1.B: β -barrel porins	7	3+4	7	7	7	2+5	7	2+5	7	2+5	70	7.0 $\pm$ 0.00
2.A: Porters (uniporters, symporters, antiporters)	212+9	151+17	191	224	235	192+16	157	248+57	284	281+32	2,306	230.6 $\pm$ 54.51
3.A: P-P-bond-hydrolysis-driven transporters	104+20	82+31	1+110	103	110	82+24	104	73+40	104	77+34	1,099	109.9 $\pm$ 6.26
3.B: Decarboxylation-driven transporters	0+2	2	1	1	1	1	1	1	2	2	14	1.4 $\pm$ 0.52
3.D: Oxidoreduction-driven transporters	21+4	15+8	23	23	23	18+4	23	26+4	26	20+7	245	24.5 $\pm$ 2.51
3.E: Light absorption-driven transporters	3	3	2	1	2	0	0	0+1	1	0	13	1.3 $\pm$ 1.16
4.C: Acyl CoA ligase-coupled transporters	1	1	1	1	1	1	1	1	2	1	11	1.1 $\pm$ 0.32
5.B: Transmembrane 1-electron transfer carriers	6+1	1	3	5	7	5	3	5+2	10	7+6	61	6.1 $\pm$ 3.54
8.A: Auxiliary transport proteins	15+9	6+6	11	13	14	7+4	7	7+5	17	6+2	129	12.9 $\pm$ 4.82
9.A: Recognized transporters of unknown biochemical mechanism	55+29	42+38	73	74	76	40+32	69	40+37	74	31+47	757	75.7 $\pm$ 4.30
9.B: Putative transport proteins	13+8	7+16	18	20	19	7+8	17	10+9	18	9+13	192	19.2 $\pm$ 2.39
Total	554	462	464	498	521	474	412	600	569	617	5,171	517.1 $\pm$ 66.71

Sums shows the number of transporters in the updated YETI database (1st term) and the automatically assigned ones (2nd term).

Based on the 2nd digit of the TCDB classification, 4,093 transporters are categorized according to their class (i.e. channel, porter) and subclass or energy source (Table 4). More than half of them (2.A) are porters (uniporters, symporters, antiporters) and a quarter (3.A) hydrolyse the diphosphate bond of a nucleoside triphosphate (ATP or another one) to drive the active uptake and/or extrusion of (a) solute(s). Besides, 129 are auxiliary transport proteins and 949 are categorized in the digit 9 of the TC system, with 757 recognized transporters of unknown biochemical mechanism (9.A) and 192 putative transport proteins. A more detailed view is given in supplementary table 1 at the third digit of the TC classification (TC ID3).

The TC ID of all transporters is given in supplemental data (supplementary table 2: Classification of transporters in ten hemiascomycetous yeasts) as well as the number of predicted transmembrane spans (TMS).

Taking interest in the transport of a specific substrate, the relationship of orthology to an experimentally annotated gene can help to find the gene that most probably transports the same substrate in another species. For this purpose, orthologs are given for the 167 genes classed in the TC database (supplementary table 3) for which it was possible to infer them by neighborhood and similarity (Seret & Baret, 2011).

Variation of the size of families (TC ID3): A chi-square test was used to test the hypothesis that the number of transporter per species in each family is equivalent in each species but proportional to its genome size. The hypothesis was accepted for 89 of the 94 TC ID3 with a p-value of 0.05 (Table 5), which means that in a large majority of cases, the observed variation of number of transporters among species is likely to have occurred by chance.

Table 5. Number of TC ID3 per type of evolution pattern and listing of the TC ID3 for which the chi-square test was significant at p=0.05.

Evolution pattern	Number of TC ID3	NRH ₀	RH ₀	TC ID3 with RH ₀
Ubiquitous	71	69	2	2.A.1, 2.A.67
No member in one species	6	5	1	8.A.5
No member in recent species (lost)	6	5	1	3.D.1
No member in ancestral species (gained)	3	3	0	
Others	8	7	1	8.A.9

The five TC ID3 showing significant variations are:

- The Major Facilitator Superfamily or MFS (2.A.1) which comprises between 44 members in *E. gossypii* and 151 members in *A. adenivorans*

- The Oligopeptide Transporter (OPT) Family (2.A.67) which comprises between 1 member in *C. glabrata* and 18 members in *Y. lipolytica*.
- The H⁺ or Na⁺-translocating NADH Dehydrogenase (NDH) Family (3.D.1) which comprises 8 members in the three most ancestral species (*Y. lipolytica*, *A. adeninivorans* and *D. hansenii*) and is completely absent in the seven more recent species
- The Voltage-gated K⁺ Channel β -subunit (Kv β) Family (8.A.5) which comprises between 1 member in *K. lactis* and 9 members in *A. adeninivorans* but is absent in *E. gossypii*
- The rBAT Transport Accessory Protein (rBAT) Family (8.A.9) which comprises from 2 members in *A. adeninivorans* and *D. hansenii* to 7 members in *S. cerevisiae* but is absent in three species (*E. gossypii*, *Z. rouxii* and *C. glabrata*).

Thirteen families only comprise the same number of transporters in each species (1.A.15, 2.A.9, 2.A.30, 2.A.54, 2.A.57, 2.A.89, 8.A.6, 8.A.13, 8.A.26, 9.A.6, 9.A.46, 9.B.1, 9.B.26). In twelve cases, there is one transporter in each species. The Cytochrome Oxidase Biogenesis (Oxa1) Family (2.A.9) counts two transporters in each species, one in each of the two subfamilies 2.A.9.1 and 2.A.9.4.

EVOLUTION PATTERNS OF FAMILIES (TC ID3)

Five different evolution patterns were distinguished (Table 5) based on the presence-absence of members in the different species.

The first one corresponds to ubiquitous families - families with members in all Hemisacomycete species- and comprises most of the families (71 families, 75.5%). It is quite expected as many transporters are essentials to maintain concentration gradients of various metabolites across cell membranes (plasma or intracellular organelles membranes). By contrast, the conservation of families of recognized transporters of unknown mechanism (9.A) or of putative transport proteins (9.B) in all species is more unexpected and suggests that some essential types of transporters are still to be characterized.

The second evolution pattern corresponds to an event of loss of all members of a transporter family in one species only. On the six cases observed (1.A.4, 2.A.16, 8.A.5, 9.A.26, 9.A.36, 9.B.24), the species that does not have any member in the family is always *Eremothecium gossypii*.

The third evolution pattern corresponds to a single event of loss of all members of a transporter family in the evolution history, resulting in an absence of member in this family in all recent species. Six cases are observed (1.A.6, 2.A.2, 2.A.40, 2.A.41, 2.A.72, 3.D.1); in half of them the loss of transporter occurred after the speciation event of *D. hansenii*. Interestingly, the H⁺ or Na⁺-translocating NADH Dehydrogenase (NDH) family (3.D.1) was highly present in the tree most ancestral species (8 members in each) and then disappeared completely.

The fourth evolution pattern seems to correspond to a new type of transporter arisen during the evolution time elapsed since their last common ancestor, as it is not present in *Y. lipolytica* or in some of the species that diverged first. Only three families are in this case (1.A.16, 2.A.44 and 9.A.9).

The last evolution pattern groups all other evolution pattern. In four cases (2.A.37, 3.E.1, 8.A.9, 9.B.2), the transporter family seems to have appeared after the speciation of *Y. lipolytica* or *A. adenivorans* but was lost once or twice. In four other cases (1.A.x1, 2.A.20, 2.A.59, 2.A.X2), the transporter family is present since the beginning of the evolution history but is lost at least two times independently.

The number of transporters per species in the 23 non-ubiquitous families is presented in supplementary table 4.

EVOLUTION PATTERNS OF SUBFAMILIES (TC ID4)

The study of the evolution pattern of subfamilies is meaningful only for the 105 subfamilies (TC ID4) belonging to the 76 families (TC ID3) in which all genes were assigned to a subfamily.

72 subfamilies are ubiquitous among which 18 subfamilies show one transporter in each species and two subfamilies (1.A.35.5, 2.A.4.4) show exactly two members in each species.

There are 8 cases of absence of members in one species only. Five cases concern *E. gossypii* (1.A.4.4, 2.A.16.3, 8.A.5.1, 9.A.26.1, 9.A.36.1) while there is one case in *C. glabrata* (2.A.67.1), *D. hansenii* (2.A.43.2) and *Y. lipolytica* (1.A.16.1). The two last cases were classed in another evolution pattern in (De Hertogh *et al.*, 2006).

Single event of loss of all members of a transporter family in the evolution history concern 6 subfamilies that were already classed as "lost" in (De Hertogh *et al.*, 2006). Subfamilies 2.A.43.3 and 2.A.72.2 are only present in *Y. lipolytica*. Subfamilies 2.A.2.6 and 2.A.7.12 also contain members in *A.*

adeninivorans and subfamilies 2.A.41.2 and 3.D.1.2 have members in *Y. lipolytica*, *A. adeninivorans* and *D. hansenii*.

The pattern of evolution of 5 subfamilies corresponds to a gain during the evolution of the hemiascomycetes phylum. Two of them (2.A.39.4 and 9.A.9.1) are gained since *K. lactis*, two others (1.A.56.2 and 2.A.44.Y1) since *D. hansenii* and one (9.B.2.1) since *S. kluyveri*.

There are 3 subfamilies of species specific* transporters, one in *S. cerevisiae* (3.A.3.3) and 2 in *D. hansenii* (2.A.43.Y1 and 9.B.2.Y1).

The remaining 11 subfamilies were lost once or twice during the evolution of the phylum. All of them except the subfamily 2.A.7.10 were classed in different evolution patterns in De Hertogh *et al.* (2006): 3 (1.A.X1.Y1, 2.A.40.5, 2.A.X2.Y1) were classed as “lost”, 4 others (1.A.1.7, 2.A.20.2, 2.A.37.4, 2.A.59.1) were considered as absent in one species only, two (2.A.6.Y1 and 2.A.37.Y1) were considered as species specific and one (3.E.1.4) was considered as “gained”.

DISCUSSION

Within the context of the rapid increase in genome sequencing projects, a method of identification and classification of the vast majority of transporter proteins is important as a starting point for diverse types of studies; the biochemical and molecular characterization of transporters (Leandro *et al.*, 2009), the mechanisms responsible for transport and regulation of a particular substrate (Arino *et al.*, 2010), the events involved in their evolution, like intra- and interspecific duplications and unusual relationships indicating specialization, specific deletions and maybe even horizontal gene transfer (Hellborg *et al.*, 2008), or for comparative analysis (Kovalchuk & Driessen, 2010).

The proposed method is based on a hierarchical classification of experimentally identified functions of transporters: the TC system, and propagates this classification to groups of orthologous or homologous genes.

Using our method, 3,077 additional genes were identified as transporters among which 99.94% could be classified: 88.95% (2,737) of them could be assigned to the 4th digit of classification of the TC system (143 different digits) and 10.99% (338) could only be assigned to the 3rd digit (18 different digits). The remaining two genes (0.06%) could not be assigned.

* The concept of species specific gene means that the gene actually has no identified ortholog in any other species. However it is expected to find orthologs in future genomes sequenced.

As foreseen by De Hertogh *et al.* (2006), the pattern of evolution of some transporter subfamilies were modified by the identification of new members, for example some subfamilies that were considered as species specific were found to be shared by other species.

The design of the method does not allow classifying transporters who do not have any homolog in the TCDB or YETI database. A rough estimate of the percentage of these transporters can be based on prediction of transmembrane spans which are characteristic of most transporters. The percentage of classified transporters with a minimum of 6 transmembrane spans predicted (73%) and the number of non-classified genes (because their homology with a gene of the TCDB was not established) containing at least 6 predicted transmembrane spans (1,326 genes) allows a rough estimate of the potentially missed transporters comprised between 20 and 30%.

If the percentage of genes involved in transport systems per species varies slightly (from 8.32% in *S. cerevisiae* to 9.80% in *S. kluyveri* for identified transporters), their numbers can be relatively different: *Y. lipolytica* possessing 1.5 times the number of transporters of the parasitic fungi *E. gossypii* (617 and 412 respectively).

It is reasonable to assume that the gain or loss of transporters generally results from physiological selection of diverging duplicated genes during the adaptation to new ecological niches (Kellis *et al.*, 2003; De Hertogh *et al.*, 2006) or to different fungal ways of life (parasitic, symbiotic or saprobic) (Lucic *et al.*, 2008). The fact that all the cases of non-presence of transporters of a particular family in only one species occurred in *E. gossypii*, the only parasitic fungi of this study, and that, more generally, *E. gossypii* also is the species where an absence of member is the more often observed (in total, 18 families do not contain any transporter in this species) corroborates this assumption.

Taking advantage of the hierarchical TC system, the manual curation of the YETI database and the Génolevures families, it was possible to rise the percentage of classified genes to 99,96% at the 3rd digit and to 92,42% at the 4th digit, for the 5,173 genes for which we had information of homology.

ACKNOWLEDGEMENTS

We would like to thank the Génolevures Consortium, coordinated by Jean-Luc Souciet, for access to the database of protein sequences from *Zygosaccharomyces rouxii*, *Kluyveromyces thermotolerans*, *Saccharomyces kluyveri*, *Eremothecium gossypii* and *Arxula adeninivorans*. We also thank André Goffeau and Maxime Dupriez for helpful discussion. With the support of the Fonds Spécial de Recherche (FSR) de l'Université de Louvain.

REFERENCES

- Andre, B. 1995. "An overview of membrane transport proteins in *Saccharomyces cerevisiae*." *Yeast*. 11:1575-1611.
- Arino, J, Ramos, J, & Sychrova, H. 2010. "Alkali metal cation transport and homeostasis in yeasts." *Microbiol.Mol.Biol Rev.* 74(1):95-120.
- Busch, W & Saier, MJr. 2002. "The transporter classification (TC) system, 2002." *Crit Rev.Biochem.Mol.Biol.* 37(5):287-337.
- De Hertogh, B, Hancy, F, Goffeau, A, & Baret, PV. 2006. "Emergence of species-specific transporters during evolution of the Hemiascomycete phylum." *Genetics*. 172:771-781.
- De Hertogh, B, Carvajal, E, Talla, E, Dujon, B, Baret, P, & Goffeau, A. 2002. "Phylogenetic classification of transporters and other membrane proteins from *Saccharomyces cerevisiae*." *Functional & Integrative Genomics*. 2:154-170.
- Goffeau, A, Barrell, BG, Bussey, H *et al.* 1996. "Life with 6000 genes." *Science*. 274:563-567.
- Hellborg, L, Woolfit, M, Arthursson-Hellborg, M, & Piskur, J. 2008. "Complex evolution of the DAL5 transporter family." *BMC.Genomics*. 9:164.
- Kellis, M, Patterson, N, Endrizzi, M, Birren, B, & Lander, ES. 2003. "Sequencing and comparison of yeast species to identify genes and regulatory elements." *Nature*. 423:241-254.
- Kovalchuk, A & Driessen, A. 2010. "Phylogenetic analysis of fungal ABC transporters." *BMC Genomics*. 11:177.

- Leandro, MJ, Fonseca, C, & Gonçalves, P. 2009. "Hexose and pentose transport in ascomycetous yeasts: an overview." *FEMS Yeast Research*. 9:511-525.
- Lucic, E, Fourrey, C, Kohler, A, Martin, F, Chalot, M, & Brun-Jacob, A. 2008. "A gene repertoire for nitrogen transporters in *Laccaria bicolor*." *New Phytologist*. 180:343-364.
- Nelissen, B, DeWachter, R, & Goffeau, A. 1997. "Classification of all putative permeases and other membrane plurispansers of the major facilitator superfamily encoded by the complete genome of *Saccharomyces cerevisiae*." *FEMS Microbiology Reviews*. 21:113-134.
- Nelissen, B, Mordant, P, Jonniaux, JL, DeWachter, R, & Goffeau, A. 1995. "Phylogenetic classification of the major superfamily of membrane transport facilitators, as deduced from yeast genome sequencing." *FEBS Letters*. 377:232-236.
- Paulsen, IT, Sliwinski, MK, Nelissen, B, Goffeau, A, & Saier, J. 1998. "Unified inventory of established and putative transporters encoded within the complete genome of *Saccharomyces cerevisiae*." *FEBS Letters*. 430:116-125.
- Seret, ML & Baret, PV. 2011. "IONS: Identification of Orthologs by Neighborhood and Similarity-an Automated Method to Identify Orthologs in Chromosomal Regions of Common Evolutionary Ancestry and its Application to Hemiascomycetous Yeasts." *Evolutionary Bioinformatics*. 7:123-133.
- Yamamoto, K, Nakayama, A, Yamamoto, Y, & Tabata, S. 2004. "Val216 decides the substrate specificity of α -glucosidase in *Saccharomyces cerevisiae*." *European Journal of Biochemistry*. 271:3414-3420.

ADDITIONAL FILES

SUPPLEMENTARY TABLE 1. NUMBERS OF TRANSPORTER PROTEINS IN TEN HEMIASCOMYCETE SPECIES PER FAMILIES (TC ID3) AND PER SPECIES.

The chi-square test was used to test the hypothesis that the number of transporter per species in each family is equivalent in each species but proportional to its genome size.

Due to its length this file was not reproduced here, it is joined on a CD-ROM.

SUPPLEMENTARY TABLE 2. LIST OF TRANSPORTERS (AUTOMATICALLY ASSIGNED AND CURRENTLY IN YETI) IN TEN HEMIASCOMYCETOUS YEASTS.

Due to its length this file was not reproduced here, it is joined on a CD-ROM.

Supplementary table 3. List of orthologs (inferred by the IONS method) to genes of the TC database.

Due to its length this file was not reproduced here, it is joined on a CD-ROM.

SUPPLEMENTARY TABLE 4. NUMBER OF TRANSPORTERS PER SPECIES IN NON-UBIQUITOUS FAMILIES (TC ID3).

* Families where the chi-square test used to test the hypothesis that the number of transporter per species in each family is equivalent in each species but proportional to its genome size is significant.

TC ID3	SACE	CAGL	ZYRO	KLTH	SAKL	KLLA	ERGO	DEHA	ARAD	YALI	Mean ± SD
1.A.4	1	1	1	1	1	1	0	2	3	3	1,4 ± 0,97
2.A.16	1	1	1	1	1	1	0	3	1	1	1,1 ± 0,74
8.A.5*	8	3	4	2	2	1	0	2	9	2	3,3 ± 2,95
9.A.26	4	2	2	3	3	2	0	6	5	2	2,9 ± 1,73
9.A.36	1	1	1	1	1	1	0	1	1	1	0,9 ± 0,32
9.B.24	1	1	1	1	1	1	0	1	1	1	0,9 ± 0,32
2.A.40	0	0	1	1	2	1	1	1	1	2	1 ± 0,67
1.A.6	0	0	0	0	0	0	0	2	1	2	0,5 ± 0,85
2.A.41	0	0	0	0	0	0	0	1	1	2	0,4 ± 0,7
3.D.1*	0	0	0	0	0	0	0	8	8	8	2,4 ± 3,86
2.A.2	0	0	0	0	0	0	0	0	1	1	0,2 ± 0,42
2.A.72	0	0	0	0	0	0	0	0	0	2	0,2 ± 0,63
1.A.16	1	1	1	1	1	1	1	1	1	0	0,9 ± 0,32
2.A.44	1	1	1	1	1	1	1	1	0	0	0,8 ± 0,42
9.A.9	1	1	1	1	2	1	0	0	0	0	0,7 ± 0,67
2.A.37	1	1	1	2	1	1	0	3	1	0	1,1 ± 0,88
3.E.1	3	3	2	1	2	0	0	1	1	0	1,3 ± 1,16

9.B.2	1	1	1	1	1	0	0	1	0	0	$0,6 \pm 0,52$
8.A.9*	7	0	0	4	5	3	0	2	2	0	$2,3 \pm 2,45$
2.A.20	1	0	0	2	1	1	0	1	1	1	$0,8 \pm 0,63$
2.A.59	1	0	1	0	1	2	0	1	1	1	$0,8 \pm 0,63$
2.A.X2	0	1	0	1	1	1	1	2	1	3	$1,1 \pm 0,88$
1.A.X1	0	0	0	1	1	1	1	1	0	1	$0,6 \pm 0,52$
Absence	8	10	9	6	5	7	18	3	5	7	

X. DISCUSSION ET PERSPECTIVES DE LA PARTIE APPLIQUÉE

Au vu de la rapidité actuelle de séquençage, la demande de méthodes de classification automatique permettant de pallier au manque de temps et de ressources pour les études expérimentales se fait sentir. La méthode de classification des transporteurs proposée dans cette thèse a permis d'atteindre, au sein des dix espèces d'hémiascomycètes étudiées, un total de 5171 transporteurs classés selon la classification hiérarchique de TCDB, soit entre 8 et 10% des gènes des génomes étudiés.

Cette discussion abordera tout d'abord l'intérêt de la classification des transporteurs, les problèmes pratiques rencontrés avec les données utilisées et se terminera par la discussion relative à la validité de l'hypothèse à partir de laquelle la méthode de classification présentée dans cette thèse a été développée.

INTÉRÊTS DE LA CLASSIFICATION DES TRANSPORTEURS

Une classification exhaustive des transporteurs est importante à différents niveaux, comme le montrent les 31 citations de l'article de (De Hertogh *et al.*, 2006) : pour leur attribuer une fonction (Thevenieau *et al.*, 2007), pour étudier la dynamique d'expansion et/ou de perte d'une fonction au cours de l'évolution (Dujon, 2006) et identifier d'éventuelles adaptations à une nouvelle niche écologique ou à un mode de vie différent, pour identifier leurs caractéristiques communes et ainsi permettre de mieux comprendre les mécanismes de transport et/ou faciliter les identifications ultérieures (Leandro *et al.*, 2009).

APPLICATIONS DANS LE DOMAINE MÉDICAL

L'identification des protéines de transport et la connaissance de leur mécanisme de fonctionnement peuvent être d'une grande utilité pour les avancées médicales dans le domaine du traitement des patients atteints par des infections dues à des pathogènes fongiques. Par exemple, les avancées rapides dans la compréhension de beaucoup des processus biologiques de base de *Candida albicans*, une levure pathogène opportuniste considérée comme une cause importante de mortalité chez les patients immunodéprimés, ont été faites suite à leur similitude avec des processus bien étudiés chez *S. cerevisiae* (Berman et Sudbery, 2002).

Bien que différents médicaments soient disponibles, beaucoup de pathogènes fongiques ont acquis des résistances multiples. La résistance à plusieurs antifongiques ayant des modes d'action différents est causée par l'expression accrue des gènes qui codent pour des protéines de transport qui agissent comme des pompes d'efflux des médicaments de manière non spécifique. Ces protéines de transport appartiennent à la famille des transporteurs ABC et généralement à la sous-famille des PDR (Pleiotropic Drug Resistance), comme Pdr5p par exemple (Goffeau, 2008).

Le séquençage du génome de *C. albicans* a permis de découvrir beaucoup de gènes possédant un homologue évident chez *S. cerevisiae* mais également beaucoup de gènes n'ayant aucun homologue chez *S. cerevisiae*. Parmi ces derniers, certains possèdent une meilleure similarité avec des gènes provenant d'autres champignons, tandis que d'autres codent un nouveau produit génique. Comme la plupart des souches de *S. cerevisiae* n'envahissent pas les tissus humains, les produits des gènes qui n'ont aucun homologue chez *S. cerevisiae* sont considérés comme de bons candidats lorsque l'on cherche à identifier les gènes importants pour les interactions avec l'hôte. Les gènes homologues qui sont communs à tous les champignons (et plus particulièrement ceux qui sont essentiels pour la croissance fongique) constituent quant à eux des cibles préférentielles pour l'élaboration d'antifongiques à large spectre s'ils ne possèdent pas d'homologues chez l'humain, ce qui limite la probabilité d'observer des effets secondaires négatifs (Berman et Sudbery, 2002).

PROBLÉMATIQUE DES MISES À JOUR DES BASES DE DONNÉES

La classification des transporteurs des cinq nouvelles espèces s'est appuyée sur le travail de classification précédemment réalisé sur cinq espèces par De Hertogh *et al.* (2006).

Deux problèmes ont été rencontrés lors de la mise en correspondance des bases de données :

1. Les bases de données utilisées pour la première classification ont subi des modifications. Suite à la ré-annotation génomique de *Debaryomyces hansenii* et du chromosome E de *Kluyveromyces lactis*, les séquences et/ou les noms de certains gènes ont été modifiés dans la base de données Génolevures, d'autres ont été identifiés comme étant des pseudogènes, ont disparus ou encore ont été fusionnés. La mise à jour de ces données est importante. En effet, il a été montré que la qualité de l'annotation des génomes affecte fortement le résultat des études génomiques ou

phylogénétiques (Milinkovitch *et al.*, 2010). On notera que l'utilisation de différentes dénominations des gènes (noms systématiques, noms standards ou numéros d'accès pour les gènes de *S. cerevisiae*) complexifie cette tâche.

2. La base TCDB utilisée comme référence pour la classification des transporteurs a été mise à jour.

Au niveau de la classification, plusieurs cas de non concordance ont pu être observés entre le digit TC attribué dans l'article de (De Hertogh *et al.*, 2006) et celui attribué actuellement dans la base de données TCDB. Dans certains cas, un gène classé avec un digit 4 temporaire dans la base de données YETI a reçu un digit 4 définitif dans TCDB suite aux nouvelles découvertes réalisées. De même, différents transporteurs putatifs (9.B) ou dont le mécanisme était jusqu'alors inconnu (9.A) ont également pu être classés suite aux nouvelles découvertes. Si ces modifications se comprennent aisément et sont tout à fait justifiées dans un processus de mise à jour, d'autres sont plus difficiles à expliquer. En effet, il existe quatre cas où le 4^{ème} digit de la classification attribué à certains gènes est modifié dans la base de données TCDB par rapport à leur classification de 2006 reprise dans YETI. Dans deux cas, le 4^{ème} digit est modifié mais les sous-familles étaient et restent non définies. Le digit « 3.A.8.2 The Inner Mitochondrial Membrane Protein Translocase (TIM) Subfamily » n'existe plus dans TCDB et les 5 gènes de *S. cerevisiae* qui y étaient classés dans YETI se retrouvent dans le digit 3.A.8.1 qui est non défini. Enfin, les deux gènes de *S. cerevisiae* appartenant à la sous-famille « 2.A.1.3 The Drug:H⁺ Antiporter-2 (14 Spanner) (DHA2) Family » dans YETI sont maintenant classés dans la sous-famille « 2.A.1.48 The Vacuolar Basic Amino Acid Transporter (V-BAAT) Family » dans la base de données TCDB.

De plus, certains gènes classés par YETI dans les digits 1.A.11 1.A.20.5, 2.A.49 et 3.A.8.1 se retrouvent actuellement dans TCDB avec un digit qui diffère dès le premier niveau, c'est-à-dire dès la classe de transporteurs. Les gènes du digit « 1.A.11 The Chloride Channel (ClC) Family » se retrouvent actuellement dans le digit « 2.A.49 The Chloride Carrier/Channel (ClC) Family », le digit a été modifié suite à l'observation que ces transporteurs pouvaient fonctionner comme des canaux ou comme des antiporteurs, mais la fonction reste identique. En ce qui concerne les gènes du digit 1.A.20.5 (The Human Phagocyte NADPH Oxidase-associated Cytochrome b558 (CytB) H⁺-channel Family, undefined subfamily), la différence peut être attribuée à la création d'un nouveau digit dans TCDB, le digit 5 (Transmembrane Electron Carriers), qui est apparu après la publication de l'article de (De Hertogh *et al.*, 2006). Les gènes de la famille « 2.A.49 The Ammonium Transporter (Amt) Family » ont été reclassés dans la famille

« 1.A.11 The Ammonia Channel Transporter (Amt) Family » dans TCDB. Et enfin, un des gènes classé dans la sous-famille « 3.A.8.1 The Outer Mitochondrial Membrane Protein Translocase (TOM) Subfamily » dans YETI est actuellement classé dans une sous-famille non définie de la famille « 1.B.33 The Outer Membrane Protein Insertion Porin (Bam Complex) (OmpIP) Family ».

En ce qui concerne ce dernier gène, l'impossibilité de télécharger les versions antérieures de la base de données TCDB et l'absence d'information relative à la source de la classification fournie dans l'article de (De Hertogh *et al.*, 2006) ne permettent pas de déterminer si ce gène était déjà classé dans TCDB à l'époque et a reçu un nouveau digit suite à de nouvelles découvertes contradictoires par rapport aux précédentes ou si le digit mentionné dans YETI pour ce gène lui a été attribué de manière erronée d'après la classification semi-automatique réalisée.

ORTHOLOGIE, HOMOLOGIE ET CONSERVATION DE LA FONCTION

Le travail de classification des transporteurs chez les levures présenté au chapitre IX est basé sur l'hypothèse d'équivalence fonctionnelle des gènes orthologues ou homologues. Le transfert de l'information fonctionnelle des gènes caractérisés expérimentalement à leurs orthologues ou homologues est un des rares moyens de caractérisation fonctionnelle détaillée pour la plupart des gènes des génomes nouvellement séquencés. La validité de cette hypothèse d'équivalence fonctionnelle des orthologues ou homologues est donc cruciale pour une annotation fonctionnelle fiable des génomes nouvellement séquencés.

Cependant, bien que les concepts d'orthologie et d'homologie soient bien représentés dans la littérature (près de 27000 occurrences dans les titres et résumés des articles sur Pubmed) et que les relations d'orthologie soient utilisées par de nombreuses méthodes pour propager l'annotation fonctionnelle d'un gène étudié expérimentalement à ces orthologues, seules deux études systématiques ont été réalisées pour tester cette hypothèse d'équivalence fonctionnelle des gènes provenant d'un ancêtre commun. L'une (citée dans (Koonin, 2005), non publiée) étudie uniquement les orthologues de type un-un, tandis que l'autre (Nehrt *et al.*, 2011) teste l'hypothèse d'équivalence fonctionnelle de gènes présentant différents types de relation d'homologie.

HYPOTHÈSE D'ÉQUIVALENCE FONCTIONNELLE DES ORTHOLOGUES

En ce qui concerne les orthologues, une estimation grossière a été obtenue par Koonin (cité dans (Koonin, 2005), non publié) en comparant les informations fonctionnelles disponibles pour 1330 groupes d'orthologues COG (Clusters of Orthologous Groups) contenant des orthologues de type un-un d'*Escherichia coli* et de *Bacillus subtilis*. Ces observations n'ont permis de trouver aucun cas de différences fonctionnelles majeures, bien que des différences subtiles, telle que la spécificité de certains enzymes ou transporteurs, soient communes. Le message général de cette brève étude de l'équivalence fonctionnelle des orthologues est que les fonctions fondamentales des orthologues peuvent se modifier mais que ces modifications sont rares, tendent à être associées à des transitions évolutives majeures et sont accompagnées par des accélérations substantielles de l'évolution (Koonin, 2005).

HYPOTHÈSE D'ÉQUIVALENCE FONCTIONNELLE DES HOMOLOGUES

Au niveau des homologues, suite au constat du manque d'étude à grande échelle sur la relation entre les différents types de relations d'homologie et la conservation de la fonction, Nehrt *et al.* se sont attelés à la tâche (Nehrt *et al.*, 2011). Bien que cet article ait été contesté à certains niveaux (cfr. p 11), on peut tout de même en retenir trois points importants :

- Premièrement, chez les orthologues de type un-un la divergence de séquence n'est pas corrélée avec la divergence de fonction. La probabilité d'équivalence de fonction pour une paire de gène orthologues de type un-un est sensiblement identique quelle que soit la divergence de séquence observée. Par contre, chez tous les types de paralogues, la probabilité d'équivalence de fonction pour une paire de gène diminue avec le pourcentage d'identité de séquence observée.
- Deuxièmement, pour une même identité de séquence entre les protéines de la paire étudiée, les paires d'homologues intra-espèces (inparalogue ou outparalogue) sont celles qui conservent le plus fréquemment une fonction équivalente. J'ajouterai à cette observation, que les inparalogues présentant généralement des pourcentages d'identité de séquences élevées, ils pourraient constituer un matériel de choix pour la propagation de l'annotation fonctionnelle, bien que le plus fréquemment de peu d'utilité du fait que l'annotation fonctionnelle est généralement propagée à partir d'une espèce modèle caractérisée à une nouvelle espèce séquencée, et non pas entre gènes appartenant à une même espèce.

- Troisièmement, pour une même identité de séquence entre les protéines de la paire étudiée, Nehrt *et al.* (2011) observent que les orthologues de type un-un et les outparalogues inter-espèces présentent approximativement la même probabilité d'équivalence de fonction, soit un peu moins de 70%.

Par rapport à l'implication pour la prédiction de fonction des protéines, Nehrt *et al.* (2011) concluent en disant qu'en l'absence de paralogues intra-espèce, leurs données indiquent que les orthologues et les outparalogues inter-espèce sont équivalents pour prédire la fonction d'une protéine mais ne peuvent cependant pas être considérés comme des relations idéales pour le transfert des termes de l'ontologie des gènes (GO terms) relatifs à la fonction, leur exactitude moyenne de prédictions étant systématiquement inférieure à 70%.

Or, dans le cadre de la propagation de l'annotation fonctionnelle de génomes nouvellement séquencés, seules des données inter-espèces sont disponibles, les relations utilisées pour la propagation ne peuvent donc être que des relations d'orthologie ou d'outparalogie.

Il est important de souligner ici que la comparaison des pouvoirs prédictifs des orthologues et des outparalogues inter-espèces de l'étude de (Nehrt *et al.*, 2011) est réalisée en comparant les paires d'orthologues présentant un certain pourcentage d'identité de séquence avec les paires d'outparalogues inter-espèces qui présentent le même pourcentage d'identité de séquence. Or, si on peut être sûr que l'ensemble des orthologues de deux espèces ont commencé à diverger au même moment, l'époque de divergence des outparalogues inter-espèces est quant à elle très variable. Dans cette étude, le pourcentage d'identité de séquences entre orthologues varie globalement entre 50 et 100%, mais une majorité d'entre eux présente un pourcentage d'identité de séquence de plus de 90%. L'étude ne donne pas la répartition des pourcentages d'identité entre paires d'outparalogues inter-espèces. Cependant d'après les seules données mentionnées dans les courbes de comparaison, qui ne sont fournies que pour des identités supérieures ou égale à 50 % (en raison d'un nombre insuffisant de paires d'orthologues de type un-un présentant des identités plus basses), on peut voir qu'une majorité d'entre eux présente un pourcentage d'identité de séquence de moins de 70%.

Dès lors, puisqu'on a vu au point un que la probabilité d'équivalence de fonction pour une paire de paralogues quel que soit son type (y compris les outparalogues inter-espèces) diminue avec le pourcentage d'identité de séquence observée, on peut considérer que dans le cadre de la propagation de l'annotation fonctionnelle entre deux espèces, si on considère l'ensemble des paires de séquences homologues inter-espèces quel que soit le pourcentage

d'identité qu'elles présentent, les orthologues de type un-un ont certainement un pouvoir prédictif de la fonction supérieur à celui des outparalogues inter-espèces.

EXPLICATION DES CHOIX RÉALISÉS POUR LA CLASSIFICATION DES TRANSPORTEURS

Au niveau de la classification des transporteurs, l'idée de départ était de propager la classification des gènes des bases de données TCDB et YETI à leurs orthologues positionnels (Ortho1 et Orthox) identifiés par la méthode IONS. Cette manière de procéder est la plus sûre car elle combine les critères de similarité et de conservation de voisinage. Elle présente cependant l'inconvénient de ne classer qu'une faible proportion des gènes. En effet, sur les 2096 gènes des bases de données TCDB et YETI seuls 587 font partie d'un groupe d'orthologues positionnels (les autres peuvent être des singletons car leur environnement chromosomique a été fortement modifié et ne permet plus de les relier à leurs homologues ou font partie de groupes de type « Undet »). Le nombre total de gènes repris dans ces groupes d'orthologues, et donc que l'on pourrait classer via cette méthode, est de 1451 gènes sur un total variant de 4345 à 5173 gènes si on utilise les relations d'homologie, selon la méthode d'inférence utilisée. Ce nombre est relativement faible et rendrait en outre l'étude des patterns évolutifs irréalisables (puisque la majorité des transporteurs ne sont pas classés).

Suite à cette constatation, deux stratégies pouvaient être envisagées :

1. Une stratégie hétérogène dans laquelle relations d'orthologie et relations d'homologie inférée par IONS étaient combinées pour la propagation de l'annotation fonctionnelle. Le nombre total de gènes repris dans les groupes d'homologues obtenus par la méthode IONS et qui contiennent au minimum un gène provenant d'une des bases de données TCDB et YETI est de 4345 gènes*. Cependant nous nous sommes aperçus que cette manière de faire risquait de biaiser l'étude des patterns évolutifs : près de la moitié des gènes (373/828) appartenant aux familles Génolevures qui contiennent au minimum un gène provenant d'une des bases de données TCDB et YETI et qui ne pourraient pas être classés suivant cette option appartiennent aux trois espèces les plus éloignées (*D. hansenii*, *A. adeninivorans* et *Y. lipolytica*). Ceci peut s'expliquer par la distance évolutive qui les sépare des autres espèces et qui a permis un plus grand nombre de

* Cette stratégie n'ayant pas été retenue, ce chiffre n'est pas repris dans l'article présenté précédemment.

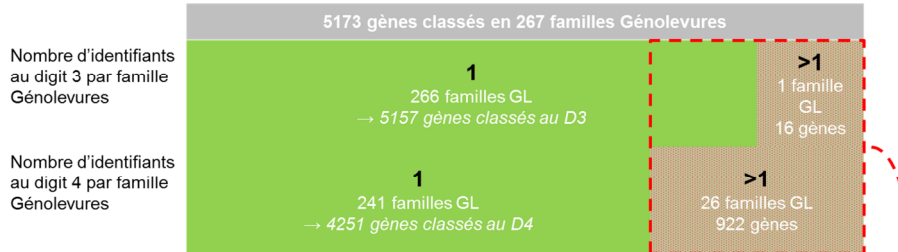
réarrangements chromosomiques effaçant ainsi la trace de leur origine commune.

2. La deuxième possibilité consiste à utiliser l'appartenance aux familles Génolevures comme inférence de l'homologie. L'avantage de cette méthode est qu'elle permet de classer le plus grand nombre de gènes (au maximum 5173) et de n'oublier aucun homologue potentiel, permettant ainsi une étude valide des patterns évolutifs des familles aux différents niveaux de la classification pour lesquels l'ensemble des gènes sont classés.

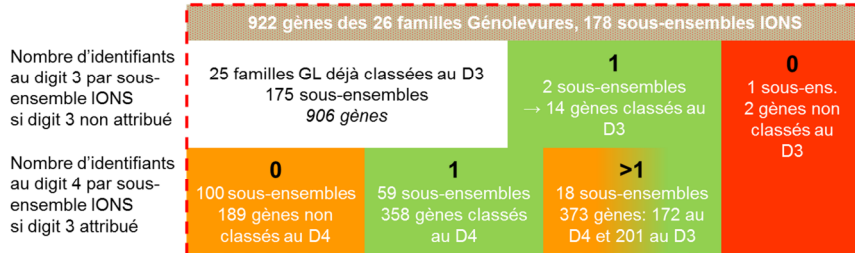
Cette dernière possibilité est celle que nous avons appliquée pour la classification des transporteurs à quelques exceptions près. Sur les 267 familles Génolevures étudiées, la classification des gènes de 241 familles, soit un total de 4251 gènes (82%), a pu être réalisée de cette manière jusqu'au digit 4 (Figure 12). Les exceptions concernent 26 familles Génolevures (soit 922 gènes), pour lesquelles la famille comptait des gènes appartenant à plusieurs identifiants TC, généralement au niveau du 4^{ème} digit de la classification. Dans ces familles hétérogènes au niveau de la fonction, nous avons ajouté la contrainte de conservation du voisinage en utilisant les sous-ensembles d'homologues créés par IONS au sein de ces familles pour propager l'annotation fonctionnelle. La méthode IONS subdivise ces 26 familles en 178 sous-ensembles d'homologues parmi lesquels il restait encore 18 sous-ensembles (soit 373 gènes) comptant des gènes appartenant à plusieurs identifiants TC au digit 4 (Figure 12). Les gènes appartenant à ces 18 sous-ensembles d'homologues créés par IONS qui n'étaient pas précédemment classés dans les bases de données YETI et TCDB ont vu leur classification limitée au digit 3.

La combinaison des familles Génolevures et des sous-ensembles d'homologues produits par la méthode IONS a permis d'atteindre un total de près de 92,42% de gènes classés jusqu'au digit 4 et de 7,54% de gènes classés jusqu'au digit 3, tandis que 0,04% (soit deux gènes) n'ont pas pu être classés.

Propagation de l'annotation fonctionnelle sur base de l'homologie inférée par les familles Génolevures



Propagation de l'annotation fonctionnelle sur base de l'homologie inférée par les sous-ensembles IONS



Totaux

4781 gènes classés au D4	390 gènes classés au D3	2 gènes non classés
--------------------------	-------------------------	---------------------

Figure 12 Utilisation des différents types d'inférence de l'homologie pour la propagation de l'annotation fonctionnelle des transporteurs.

Les gènes dont la famille comprend des gènes classés dans différents identifiants (en brun) sont ré-analysés au niveau des sous-ensembles produits par la méthode IONS. La couleur verte est attribuée aux gènes classés, l'orange aux gènes classés au digit 3 mais pas au digit 4 et le rouge aux gènes non classés par la méthode.

CONTRIBUTION DE CE TRAVAIL À L'ÉVALUATION DE L'HYPOTHÈSE D'ÉQUIVALENCE FONCTIONNELLE DES ORTHOLOGUES ET HOMOLOGUES

L'existence de ces cas d'homologues classés dans des digits TC différents pose la question de la légitimité de la propagation de l'annotation fonctionnelle aux homologues. Jusqu'à quel point les homologues conservent-ils la même fonction ? Devrions-nous arrêter la propagation de l'annotation fonctionnelle au digit 3, c'est-à-dire à la classe, la sous-classe et la source d'énergie ?

En ce qui concerne les ensembles d'orthologues définis par la méthode IONS, dans notre étude, aucun cas d'attribution de digits différents au sein d'un même ensemble n'a pu être constaté. Cette constatation est réconfortante mais n'apporte pas vraiment d'éléments permettant de vérifier l'hypothèse d'équivalence fonctionnelle des orthologues. En effet, une très

grande majorité des gènes caractérisés expérimentalement appartiennent à *S. cerevisiae* (97% des gènes de TCDB utilisés). Or les ensembles d'orthologues définis par la méthode IONS ne contiennent qu'un gène par espèce, donc un seul gène provenant de *S. cerevisiae*, il est donc assez logique de n'observer qu'un seul identifiant TC dans ces ensembles d'orthologues, soit l'identifiant attribué à l'unique gène de *S. cerevisiae*.

Par contre en ce qui concerne les homologues, 18 sous-ensembles créés par IONS contiennent des gènes présentant une classification différente au niveau du 4^{ème} digit, c'est-à-dire la sous-famille. L'avantage de l'utilisation d'une classification de type hiérarchique, est qu'elle permet de s'arrêter au niveau de classification pour lequel on a de l'information. La propagation de l'annotation fonctionnelle aux nouveaux gènes appartenant à ces sous-ensembles a donc été limitée au digit 3, la famille de transporteurs. Il est à noter que pour 10 des 18 cas observés, l'hésitation entre différents digits 4 est uniquement due à la présence d'un ou plusieurs digits 4 temporaires créés par De Hertogh *et al.* (2006) d'après l'observation d'une différence de valeur de BLAST comprise entre 1.10^{-35} and 1.10^{-65} . Ce qui laisse 8 cas où des gènes homologues ont effectivement reçu une classification différente au niveau du 4^{ème} digit sur base d'une caractérisation expérimentale. Notons également que parmi ces 8 sous-ensembles d'homologues présentant une divergence fonctionnelle entre eux, 4 font partie du 1% de sous-ensembles les plus grands et comprennent plus de 40 gènes.

Ces observations montrent que, même en utilisant une inférence de l'homologie relativement contraignante basée sur un double critère de similarité et de conservation du voisinage, il existe bel et bien des différences fonctionnelles entre homologues au niveau de la sous-famille des transporteurs (digit 4).

S'il est possible d'observer des différences fonctionnelles entre homologues dans certains sous-ensembles alors que le nombre actuel de gènes caractérisés expérimentalement est faible, cela signifie qu'il est probable qu'il en existe également dans d'autres sous-ensembles mais qu'elles ne seront découvertes qu'à la lumière de nouvelles expériences en laboratoire et/ou de nouvelles méthodes de caractérisation (étude de motifs, structure tridimensionnelles, ...). Une estimation grossière du pourcentage de divergence entre homologues au niveau de la sous-famille des transporteurs peut être obtenue en observant les gènes de la base de données TCDB : si on compare le nombre de sous-ensembles d'homologues présentant une divergence au niveau du 4^{ème} digit de la classification (soit 8) au nombre de sous-ensembles d'homologues possédant plus d'un gène de la base de données TCDB (soit 39), on peut estimer la proportion de divergence fonctionnelle entre transporteurs homologues au 4^{ème} digit de la classification

à 20%. Il est à noter que, les 8 cas de divergence étant dus à la présence de gènes de TCDB venant de *S. cerevisiae* (qui représente 97% des gènes de TCDB utilisés), cette estimation est une estimation de la divergence fonctionnelle entre transporteurs homologues intra-espèce.

PERSPECTIVES POUR L'AMÉLIORATION DE LA CLASSIFICATION

Quatre points sont à prendre en compte en ce qui concerne la confiance que l'on peut accorder à la classification des transporteurs ou à mettre en place pour limiter les erreurs :

Premièrement, il faut tenir compte du type d'inférence, de leur nombre et de la méthode par laquelle elles sont obtenues. Dans un cadre inter-espèces, la propagation de l'annotation fonctionnelle au sein d'un groupe d'orthologues de type un-un, est celle qui présente le moins de risque. D'après nos observations et assez logiquement, les groupes d'homologues de taille importante ont plus de chance de présenter des fonctions différentes, bien qu'apparentées. Enfin l'utilisation d'une inférence de l'homologie basée sur un double critère de similarité et de voisinage permet de limiter la propagation de l'annotation fonctionnelle au niveau de classification pour lequel l'information est présente.

Deuxièmement, il est important d'élargir la base de données des transporteurs caractérisés expérimentalement. Le choix des gènes à caractériser devrait idéalement être réalisé en fonction des résultats obtenus via les analyses bio-informatiques de manière à éclaircir les zones d'incertitude. On peut distinguer trois cas différents, par ordre de priorité. Le premier permettrait d'attribuer un digit 4 à un ensemble de gènes pour lesquels il existe actuellement une incertitude à ce niveau. Les deux suivants permettraient de tester la confiance à attribuer à la classification respectivement dans un cadre inter et intra-espèce.

1. Pour les 26 familles génolevures comprenant des gènes classés dans différents digits, il serait intéressant de caractériser expérimentalement la fonction d'au moins un gène appartenant à chaque grand sous-ensemble d'homologues obtenus par IONS, de manière à attribuer le bon digit à chacun de ces sous-ensembles. Ces 26 familles sont subdivisées en 178 sous-ensembles par la méthode IONS. Parmi ceux-ci, 18 comprennent plus de 7 gènes dont aucun n'est classé sur base expérimentale dans la base de données TCDB. Sur ces 18 sous-ensembles, 16 comprennent au moins un gène chez *S. cerevisiae*. La priorité numéro un serait donc de caractériser un

gène de *S. cerevisiae* appartenant à ces 16 sous-ensembles, soit 16 gènes, voire l'ensemble des gènes de *S. cerevisiae* appartenant à ces 16 sous-ensembles, soit 23 gènes. Ceci permettrait de clarifier le digit 4 à attribuer à 169 gènes des 390 gènes non classés au digit 4, soit 43%.

2. Les 73 familles comprenant exactement un gène par espèce, constituent le cadre idéal pour étudier la variabilité de fonction entre orthologues chez les transporteurs sur base du type d'information propagée au 4^{ème} digit de la classification. Pour étudier cette question tout en limitant le nombre de caractérisations à réaliser, il faudrait caractériser les gènes de *Y. lipolytica*, c'est-à-dire les gènes appartenant à l'espèce la plus éloignée d'un point de vue évolutif possible par rapport à *S. cerevisiae* (espèce dont sont issus les gènes caractérisés précédemment). En effet, si la caractérisation d'un gène appartenant à *Y. lipolytica* permet de confirmer qu'il possède la même fonction que son orthologue chez *S. cerevisiae*, cela améliore fortement la confiance en l'inférence de la fonction des gènes appartenant aux autres espèces.
3. Parmi les familles Génolevures étudiées, 41 familles possèdent un seul gène de *S. cerevisiae* caractérisé expérimentalement jusqu'à présent alors qu'elles contiennent plus d'un membre dans cette espèce. La caractérisation d'au minimum un second gène appartenant à ces familles (soit 41 gènes) voire de l'ensemble des gènes de *S. cerevisiae* non caractérisés appartenant à ces familles (soit 143 gènes) permettrait d'affiner notre estimation de la variabilité de fonction entre homologues intra-espèce.

Troisièmement, n'ayant aucune donnée relative à la probabilité de conservation de la fonction entre outparalogues inter-espèces pour lesquels le pourcentage d'identité de séquence est inférieur à 50% (non étudié dans (Nehrt *et al.*, 2011)), il serait prudent d'ajouter ce seuil dans le processus de propagation de la fonction. Les gènes qui n'atteindraient pas ce seuil ne seraient classés qu'au 3^{ème} digit.

Quatrièmement, à défaut de nouvelles caractérisations expérimentales, une curation manuelle serait particulièrement intéressante pour les 26 familles comprenant des gènes classés dans différents digits afin de vérifier que l'ensemble des gènes classés dans un même digit correspondent aux caractéristiques reconnues pour les gènes classés expérimentalement dans ce digit (gamme de nombre de TMS ou de longueur de séquence, motifs ou structure tridimensionnelle particulière).

CONCLUSION

XI. CONCLUSION ET PERSPECTIVES

Dans le cadre de cette thèse, nous avons tenté de garder un équilibre entre la participation à un projet de génomique comparative à grande échelle en collaboration avec le consortium Génolevures et des études à l'échelle moléculaire réalisées en collaboration avec différents experts de familles de transporteurs. Le travail réalisé est majoritairement du ressort de la bioinformatique. Mes contributions à la problématique de l'annotation fonctionnelle des gènes sont donc essentiellement d'ordre méthodologique via le développement d'une méthode incluant le concept de synténie pour inférer l'orthologie dans un contexte multi-espèces et d'une méthode de propagation d'une classification fonctionnelle des transporteurs à de nouvelles espèces.

LES POINTS FORTS DES MÉTHODES PROPOSÉES

Les méthodologies développées dans cette thèse ont en commun trois points forts :

- Elles sont basées sur un concept hiérarchique
- Elles sont systématiques et transparentes
- Elles permettent de prendre en compte les exceptions, la variabilité des données.

UTILISATION DE CONCEPTS HIÉRARCHIQUES

La numérotation hiérarchique des sous-ensembles d'homologues obtenus par la méthode IONS permet de retracer les différentes subdivisions subies par une famille Génolevures au cours de la procédure progressive d'étude du voisinage des gènes permettant d'identifier les orthologues positionnels. Elle permet donc d'identifier facilement quels sont les gènes homologues de ces groupes d'orthologues positionnels.

Au niveau de la classification des transporteurs, l'utilisation d'une classification de type hiérarchique permet de limiter la propagation de l'annotation fonctionnelle au niveau le plus adéquat. La similarité de séquences étant insuffisante pour permettre de déterminer le ou les substrat(s) transporté(s), la propagation de l'annotation fonctionnelle a été limitée au 4^{ème} niveau de la classification TCDB, c'est-à-dire à la classe, la sous-classe, la famille et la sous-famille.

MÉTHODES SYSTÉMATIQUES ET TRANSPARENTES

Les méthodes développées se veulent les plus systématiques possible de manière à être totalement reproductible et à ce qu'un utilisateur potentiel puisse facilement assimiler leur mode opératoire et expliquer les résultats obtenus.

En ce qui concerne la méthode IONS, cette caractéristique est due essentiellement au concept de base qui la sous-tend, à savoir la transitivité de l'homologie, et à son automatisation. Le design de la méthode permet également d'assurer la stabilité des groupes d'orthologues positionnels face à l'ajout de nouvelles espèces n'augmentant pas la distance évolutive globale considérée (c'est-à-dire d'espèces situées entre les espèces actuelles sur un arbre phylogénétique). En plus de la numérotation hiérarchique, l'algorithme produit différents fichiers permettant de retrouver le cheminement parcouru pour obtenir le résultat observé.

En ce qui concerne la classification des transporteurs, les règles de propagation de la classification sont identiques quel que soit le niveau de classification et le type d'inférence de l'homologie utilisée. Pour chaque niveau de classification, dans un premier temps la propagation de l'annotation fonctionnelle est réalisée à partir des familles Génolevures. Dans un second temps, elle est réalisée d'après les sous-ensembles d'homologues obtenus par IONS et ce, uniquement pour les gènes appartenant aux familles Génolevures pour lesquelles l'extension de la classification aux nouveaux gènes n'a pas pu être réalisée suite à la présence de gènes appartenant à des digits différents au sein d'une même famille.

« LES EXCEPTIONS FONT LA RÈGLE »

Quel que soit le domaine biologique étudié, il existe toujours une variabilité qu'il n'est pas simple de prendre en compte dans un système totalement automatisé. Cet état de fait pousse généralement les scientifiques à autoriser une paramétrisation de leurs modèles, programmes, ...

En ce qui nous concerne, les familles Génolevures étudiées présentent une variabilité au niveau de leur taille, de la représentation des espèces, de leur évolution et/ou de la fonction des gènes qui les compose.

Au niveau de la méthode IONS, il était donc illusoire de déterminer une taille de voisinage ou une contrainte synténique unique adaptée à l'étude de l'ensemble de ces familles. C'est pourquoi nous avons fait en sorte que l'utilisateur puisse définir lui-même la taille du voisinage qu'il souhaite prendre en compte au départ ainsi qu'une contrainte synténique adaptée à ses souhaits. L'utilisation d'une taille de voisinage importante augmente le

risque d'observer des liaisons entre gènes voisins dues au hasard, mais pourra éventuellement permettre d'identifier la présence de duplication segmentale. La taille du voisinage est également adaptée au sein même de l'algorithme de manière à identifier un maximum de groupes d'orthologues positionnels. Au niveau de la contrainte synténique, une faible contrainte, un voisin en commun par exemple, permettra d'identifier un maximum d'homologues notamment lorsqu'on est en présence d'espèces fort distantes d'un point de vue évolutif, tandis qu'une contrainte plus élevée améliore la confiance dans les relations d'orthologie et d'homologie identifiées.

Dans le cadre de la classification des transporteurs, l'utilisation combinée de deux types d'inférence de groupes de gènes homologues, à savoir les familles Génolevures et les sous-ensembles obtenus par la méthode IONS a permis d'optimiser le nombre de transporteurs classés tout en limitant la propagation de l'annotation fonctionnelle à un niveau de classification inférieur en cas d'incertitude.

CONTRIBUTIONS

Dans un contexte multi-espèces, la méthode IONS est la première méthode automatisée d'inférence des orthologues positionnels basée sur la similarité et le voisinage chromosomique qui ne soit pas directement dépendante d'une phylogénie des espèces.

Les orthologues positionnels présentent un intérêt majeur pour l'identification de la synténie à l'échelle des génomes et sont fréquemment utilisés comme base de données pour la réalisation d'un arbre phylogénétique robuste des espèces ainsi que pour la propagation de l'annotation fonctionnelle.

Les orthologues positionnels identifiés par la méthode IONS ont permis l'étude de la synténie entre neuf espèces d'hémiascomycètes, ce qui a notamment conduit à la constatation d'une relation non linéaire entre la divergence des séquences protéiques des orthologues et la conservation de la synténie (annexe 1, figure 5). Cette relation non-linéaire sous-tend l'hypothèse d'une diminution rapide de l'identité entre séquences orthologues précédant un réarrangement des cartes chromosomiques pour les espèces proches d'un point de vue phylogénétique. Les espèces plus distantes présentent quant à elles des réarrangements chromosomiques plus importants tandis que la divergence entre séquences protéiques tend vers un plateau ce qui peut être expliqué par des contraintes fonctionnelles et par la saturation.

Concernant les transporteurs transmembranaires des hémiascomycètes, les études de familles spécifiques de transporteurs ont permis de mettre en évidence le fait que la similarité de séquence ne permet pas directement d'identifier les orthologues positionnels.

UTILISATION DES ENSEMBLES D'ORTHOLOGUES DÉFINIS PAR LA MÉTHODE IONS POUR L'ANNOTATION FONCTIONNELLE

L'annotation fonctionnelle d'un gène est un processus difficile qui constitue le goulot d'étranglement le plus important de la génomique, la séquence d'un gène n'étant que de peu d'utilité si sa fonction n'est pas décrite correctement.

En ce qui concerne l'utilisation des méthodes d'identification des homologues ou orthologues pour l'annotation fonctionnelle, il est important de garder à l'esprit que les gènes homologues ou même orthologues ne partagent pas systématiquement la même fonction.

La probabilité que deux gènes orthologues de type un-un exercent une fonction équivalente n'est pas égale à un. La propagation de l'annotation fonctionnelle des gènes caractérisés expérimentalement au sein d'une espèce aux gènes d'une autre espèce via les relations d'orthologie amène donc des erreurs que l'utilisation d'un seuil de pourcentage d'identité minimum ne permet pas de diminuer (Nehrt *et al.*, 2011).

En effet, les mécanismes d'évolution peuvent avoir un effet sur la fonction des gènes. Une étude (Yamamoto *et al.*, 2004) montre qu'une modification de spécificité de substrat peut être liée à la substitution d'un acide aminé unique. Par exemple, chez *S. cerevisiae* une substitution de la valine située en position 216 de l'isomaltase par une thréonine (acide aminé présent chez la maltase) modifie la spécificité de l'isomaltose vers le maltose. Or une étude réalisée sur 241 groupes d'orthologues de transporteurs transmembranaire des neuf espèces de levures étudiées ici, montre que la variabilité de l'identité en acides aminés entre les paires de séquences orthologues de ces groupes est fréquemment égale ou supérieure à 30% (Desseille, 2011).

De plus, si le terme « duplication génique » semble impliquer que tous les caractéristiques pertinentes au niveau de la fonction d'un gène ancestral sont également dupliquées et que dès lors les deux copies devraient être fonctionnellement équivalentes, en réalité, il peut exister plusieurs différences entre les deux copies. La copie dérivée peut avoir une

localisation génomique différente et ne retient généralement pas le répertoire complet des éléments régulateurs du gène ancestral. Or, la régulation d'un gène peut influencer sa fonction. Ces différences suggèrent que la copie dérivée pourrait évoluer sous des contraintes différentes (due à un relâchement de la sélection naturelle ou à une sélection pour de nouveaux attributs). Une étude (Bu *et al.*, 2011) réalisée chez *S. cerevisiae* portant sur 43 paires d'ohnologues et 15 duplications à petite échelle (segmentale ou génique) montre que pour les duplications à petite échelle, dans la plupart des cas, c'est la copie dérivée qui évolue le plus vite, ce qui pourrait l'amener à évoluer vers une nouvelle fonction. Au niveau des ohnologues, l'étude n'a pas pu observer de lien entre la conservation du voisinage chromosomique et la vitesse d'évolution de la séquence ; la copie dont l'environnement chromosomique a été le moins modifié par rapport à l'environnement chromosomique ancestral, n'est pas forcément celle qui évolue le moins vite, ce qui suggère qu'elle ne conserve pas forcément la fonction ancestrale.

De plus, des cas de subfonctionnalisation (van Hoof, 2005) et de néofonctionnalisation (Byrne et Wolfe, 2007) existent chez les levures.

Par exemple, un cas de subfonctionnalisation a été décrit pour les ohnologues ORC1 et SIR3 de *S. cerevisiae*. La protéine Orc1p est essentielle pour la viabilité et fait partie du complexe de reconnaissance de l'origine intervenant lors de la réplication de l'ADN. Sir3p fait aussi partie d'un complexe de protéines qui inclut une partie des mêmes protéines que le complexe Orc1p et intervient dans l'inhibition de l'expression de certains gènes (Sir : silent information regulator). En conséquence, les mutants ne possédant pas le gène SIR3 sont vivants, mais ne peuvent pas se reproduire de manière sexuée. Des copies supplémentaires du gène SIR3 ne peuvent pas compléter un mutant ORC1 et vice versa, indiquant que les deux protéines ont des fonctions différentes. Leur orthologue chez *S. kluyveri* qui présente une identité de 48 % à Orc1p et de 24 % à Sir3p a été multiplié et inséré chez des mutants de *S. cerevisiae* dont le gène ORC1 ou SIR3 était désactivé. Le gène de *S. kluyveri* a permis de restaurer la viabilité des mutants ORC1 mais également de restaurer la capacité de reproduction des mutants SIR3, démontrant ainsi que le gène de *S. kluyveri* est capable d'effectuer les fonctions de Sir3p et de Orc1p, ce qui laisse présumer un cas de subfonctionnalisation, l'explication la plus parcimonieuse de ces données étant que les deux gènes de *S. cerevisiae* aient évolués à partir d'un gène ancestral qui exécutait deux fonctions distinctes et séparables (van Hoof, 2005).

L'existence de cas de subfonctionnalisation et de néofonctionnalisation est d'autant plus importante à garder en mémoire que les annotations

fonctionnelles réalisées expérimentalement et servant de base au transfert de l'annotation vers d'autres espèces ont été réalisées presque exclusivement chez *S. cerevisiae*, c'est-à-dire chez une espèce qui a subi une duplication génomique complète.

Cependant, par manque de données d'annotation fonctionnelle expérimentales intra-espèce, l'utilisation des relations d'orthologie pour propager l'annotation fonctionnelle est actuellement la meilleure méthode possible.

En ce qui concerne les relations d'outparalogie, leur utilisation pour la propagation de l'annotation fonctionnelle n'augmenterait pas la proportion d'erreurs pour autant que le pourcentage d'identité entre les séquences d'outparalogues soit supérieur à 50% (Nehrt *et al.*, 2011). Cet élargissement aux relations d'outparalogie pour la propagation de l'annotation fonctionnelle permet par contre d'augmenter significativement le nombre de gènes annotés.

PERSPECTIVES

Une amélioration du pourcentage d'orthologues positionnels identifiés par la méthode IONS pourrait être obtenue par un traitement adéquat des gènes dupliqués en tandem ainsi que par une évaluation de la « force » des relations reliant les environnements chromosomiques entre eux afin de pouvoir mieux discriminer le gène situé dans l'environnement chromosomique ressemblant le plus à l'environnement chromosomique de l'espèce ancestrale par rapport à une copie obtenue par une duplication segmentale ectopique. Cette évaluation de la « force » des relations reliant les environnements chromosomiques entre eux permettrait en outre de quantifier la robustesse des inférences des relations d'orthologie.

Dans le cadre de l'étude de l'histoire évolutive de familles de gène, une des perspectives pour la méthode IONS était de la développer afin d'inférer de manière automatisée les différents réarrangements chromosomiques les plus probables intervenus au cours de l'histoire évolutive séparant deux gènes orthologues ou homologues.

Une partie des perspectives établies pour la méthode IONS a reçu un écho via le développement de l'outil MultiMSOAR 2.0, une méthode d'identification des orthologues positionnels dans un contexte multi-génomes proposant la prise en compte des gènes dupliqués en tandem et réalisant une inférence des réarrangement chromosomiques (Shi *et al.*, 2011)

CONCLUSION

L'utilisation des deux critères que sont la similarité de séquence et la conservation du voisinage chromosomique a permis de développer une méthode fiable et simple d'un point de vue conceptuel pour inférer les relations d'homologie entre les gènes de plusieurs espèces. Dans le cas de notre étude portant sur les génomes de neuf levures, celle-ci a permis de confirmer la relation d'homologie de 87% des gènes analysés. Ce pourcentage peut être amélioré par l'ajout de nouvelles espèces phylogénétiquement situées entre les espèces relativement fort distantes d'un point de vue évolutif. Un essai préliminaire réalisé en ajoutant le génome d'*Arxula adenivorans* ainsi qu'une moitié du génome diploïde de *Pichia sorbitophila* permet d'atteindre un pourcentage de 95%.

L'identification des orthologues, bien que limitée aux relations de type un-un, semble présenter un taux de faux positifs très faible. Elle s'est avérée d'une grande utilité comme point d'ancrage pour une étude de la synténie à l'échelle des génomes réalisée sur cinq espèces de levures.

Les sous-ensembles, obtenus par la méthode IONS présentent un réel intérêt pour l'automatisation de l'annotation fonctionnelle des levures (ou de tout autre phylum d'espèces). Cependant, il est important de rappeler que la méthode d'annotation fonctionnelle présentée dans la seconde partie de cette thèse est basée sur des évidences informatiques et que seule l'évidence expérimentale est définitive, même si elle est en pratique impossible vu les volumes de données à annoter.

La visualisation des environnements chromosomiques d'un groupe de protéines orthologues placées dans l'ordre phylogénétique et mettant en évidence les relations d'homologie, telle que les fichiers visuels produits par la méthode IONS peut faciliter l'étude manuelle des mécanismes évolutifs à l'échelle des gènes. On notera cependant que pour cette application, de nouvelles approches, telle que celle proposée par MultiMSOAR 2.0, permettent actuellement d'aller plus loin en modélisant explicitement l'évolution du génome (par inversion, délétion, insertion, duplication, translocation, fusion et fission chromosomique).

Au niveau de la classification des transporteurs, la méthode d'extension de la classification a permis d'établir différentes priorités pour la caractérisation expérimentale de gènes et de mettre en évidence l'existence de familles contenant des gènes présentant des fonctions différentes.

L'amélioration constante des méthodes d'inférences des relations d'homologie, les efforts de curation manuelle et l'augmentation du nombre de gènes caractérisés expérimentalement de manière ciblée ne peuvent que faciliter la classification des transporteurs, même si l'annotation fine, telle que la spécificité de substrat, restera sans doute toujours l'apanage des expériences en laboratoire.

RÉFÉRENCES

- Alberts, B, A Johnson, J Lewis, M Raff, K Roberts and P Walter. 2008. *Molecular biology of the cell*. Garland. New York,
- Alexeyenko, A, Lindberg, J, Pérez-Bercoff, A and Sonnhammer, ELL. 2006. "Overview and comparison of ortholog databases." *Drug Discovery Today: Technologies*. 3:137-143.
- Altenhoff, AM, Schneider, A, Gonnet, GH and Dessimoz, C. 2011. "OMA 2011: orthology inference among 1000 complete genomes." *Nucleic Acids Res*. 39:D289-D294.
- Altenhoff, AM and Dessimoz, C. 2009. "Phylogenetic and Functional Assessment of Orthologs Inference Projects and Methods." *PLoS Comput Biol*. 5:e1000262.
- Altschul, SF, Gish, W, Miller, W, Meyers, EW and Lipman, DJ. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology*. 215:403-410.
- Ashburner, M, Ball, CA, Blake, JA, Botstein, D, Butler, H, Cherry, JM, Davis, AP, Dolinski, K, Dwight, SS, Eppig, JT, Harris, MA, Hill, DP, Issel-Tarver, L, Kasarskis, A, Lewis, S, Matese, JC, Richardson, JE, Ringwald, M, Rubin, GM and Sherlock, G. 2000. "Gene Ontology: tool for the unification of biology." *Nature Genetics*. 25:25-29.
- Barth, G and Gaillardin, C. 1997. "Physiology and genetics of the dimorphic fungus *Yarrowia lipolytica*." *Fems Microbiology Reviews*. 19:219-237.
- Berman, J and Sudbery, PE. 2002. "Candida Albicans: a molecular revolution built on lessons from budding yeast." *Nat.Rev.Genet*. 3(12):918-930.
- Boeckmann, B, Robinson-Rechavi, M, Xenarios, I and Dessimoz, C. 2011. "Conceptual framework and pilot study to benchmark phylogenomic databases based on reference gene trees." *Brief.Bioinform*. 12(5):423-435.
- Breunig, KD, Bolotin-Fukuhara, M, Bianchi, MM, Bourgarel, D, Falcone, C, Ferrero, I, Frontali, L, Goffrini, P, Krijger, JJ, Mazzoni, C, Milkowski, C, Steensma, HY, Wésolowski-Louvel, M and Zeeman, AM. 2000. "Regulation of primary carbon metabolism in *Kluyveromyces lactis*." *Enzyme and Microbial Technology*. 26:771-780.

- Bu, L, Bergthorsson, U and Katju, V. 2011. "Local synteny and codon usage contribute to asymmetric sequence divergence of *Saccharomyces cerevisiae* gene duplicates." *BMC Evol.Biol.* 11:279.
- Busch, W and Saier, MJr. 2002. "The transporter classification (TC) system, 2002." *Crit Rev.Biochem.Mol.Biol.* 37(5):287-337.
- Byrne, KP and Wolfe, KH. 2005. "The Yeast Gene Order Browser: Combining curated homology and syntenic context reveals gene fate in polyploid species." *Genome Research.* 15:1456-1461.
- Byrne, KP and Wolfe, KH. 2007. "Consistent Patterns of Rate Asymmetry and Gene Loss Indicate Widespread Neofunctionalization of Yeast Genes After Whole-Genome Duplication." *Genetics.* 175:1341-1350.
- Cabrito, TR, Teixeira, MC, Singh, A, Prasad, R and Sa-Correia, I. 2011. "The yeast ABC transporter Pdr18 (ORF YNR070w) controls plasma membrane sterol composition, playing a role in multidrug resistance." *Biochem.J.* 440(2):195-202.
- Chen, F, Mackey, AJ, Vermunt, JK and Roos, DS. 2007. "Assessing Performance of Orthology Detection Strategies Applied to Eukaryotic Genomes." *PLoS ONE.* 2:e383.
- Csank, C and Haynes, K. 2000. "Candida glabrata displays pseudohyphal growth." *FEMS Microbiol.Lett.* 189(1):115-120.
- De Hertogh, B, Hancy, F, Goffeau, A and Baret, PV. 2006. "Emergence of species-specific transporters during evolution of the Hemiascomycete phylum." *Genetics.* 172:771-781.
- Despons, L, Baret, PV, Frangeul, L, Louis, VL, Durrens, P and Souciet, JL. 2010. "Genome-wide computational prediction of tandem gene arrays: application in yeasts." *BMC Genomics.* 11:56.
- Desseille, C. 2011. "Diversité de séquences et diversité de fonctions des transporteurs chez les hémiacomycètes." Mémoire réalisé en vue de l'obtention du diplôme de Master bioingénieur : chimie et bio-industries, Université Catholique de Louvain, 98p.
- Dessimoz, C., Cannarozzi, G, Gil, M, Margadant, D, Roth, A, Schneider, A and Gonnet, G. 2005. "OMA, A Comprehensive, Automated Project for the Identification of Orthologs from Complete Genome Data: Introduction and First Achievements." In A.McLysaght and D.Huson, editors, Springer Berlin / Heidelberg, 61-72.

Dessimoz, C, Gabaldón, T, Roos, DS, Sonnhammer, E, Herrero, J and the Quest for Orthologs Consortium. 2012. "Toward Community Standards in the Quest for Orthologs." *Bioinformatics*.

Dewey, CN. 2011. "Positional orthology: Putting genomic evolutionary relationships into context." *Briefings in Bioinformatics*. 12:401-412.

Dolinski, K and Botstein, D. 2007. "Orthology and functional conservation in eukaryotes." *Annu.Rev.Genet.* 41:465-507.

Dujon, B. 2005. "Hemiascomycetous yeasts at the forefront of comparative genomics." *Current Opinion in Genetics & Development*. 15:614-620.

Dujon, B. 2006. "Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution." *Trends in Genetics*. 22:375-387.

Dujon, B. 2010. "Yeast evolutionary genomics." *Nature Reviews Genetics*. 11:512-524.

Dujon, B, Sherman, D, Fischer, G, Durrens, P, Casaregola, S, Lafontaine, I, de Montigny, J, Marck, C, Neuveglise, C, Talla, E, Goffard, N, Frangeul, L, Aigle, M, Anthouard, V, Babour, A, Barbe, V, Barnay, S, Blanchin, S, Beckerich, JM, Beyne, E, Bleykasten, C, Boisrame, A, Boyer, J, Cattolico, L, Confanioleri, F, de Daruvar, A, Despons, L, Fabre, E, Fairhead, C, Ferry-Dumazet, H, Groppi, A, Hantraye, F, Hennequin, C, Jauniaux, N, Joyet, P, Kachouri, R, Kerrest, A, Koszul, R, Lemaire, M, Lesur, I, Ma, L, Muller, H, Nicaud, JM, Nikolski, M, Oztas, S, Ozier-Kalogeropoulos, O, Pellenz, S, Potier, S, Richard, GF, Straub, ML, Suleau, A, Swennen, D, Tekaiia, F, Wesolowski-Louvel, M, Westhof, E, Wirth, B, Zeniou-Meyer, M, Zivanovic, I, Bolotin-Fukuhara, M, Thierry, A, Bouchier, C, Caudron, B, Scarpelli, C, Gaillardin, C, Weissenbach, J, Wincker, P and Souciet, JL. 2004. "Genome evolution in yeasts." *Nature*. 430:35-44.

Durrens, P and Sherman, DJ. 2005. "A systematic nomenclature of chromosomal elements for hemiascomycete yeasts." *Yeast*. 22:337-342.

Enright, AJ, Van Dongen, S and Ouzounis, CA. 2002. "An efficient algorithm for large-scale detection of protein families." *Nucleic Acids Research*. 30:1575-1584.

Fang, G, Bhardwaj, N, Robilotto, R and Gerstein, MB. 2010. "Getting started in gene orthology and functional analysis." *PLoS Comput Biol*. 6(3):e1000703.

Fischer, G, Neuveglise, C, Durrens, P, Gaillardin, C and Dujon, B. 2001. "Evolution of gene order in the genomes of two related yeast species." *Genome Research*. 11:2009-2019.

- Fischer, G, Rocha, EPC, Brunet, F, Vergassola, M and Dujon, B. 2006. "Highly variable rates of genome rearrangements between hemiascomycetous yeast lineages." *PLoS Genetics*. 2:e32.
- Fitch, WM. 1970. "Distinguishing Homologous from Analogous Proteins." *Systematic Zoology*. 19:99-113.
- Forslund, K, Pekkari, I and Sonnhammer, EL. 2011. "Domain architecture conservation in orthologs." *BMC Bioinformatics*. 12:326.
- Fu, Z, Chen, X, Vacic, V, Nan, P, Zhong, Y and Jiang, T. 2007. "MSOAR: a high-throughput ortholog assignment system based on genome rearrangement." *J.Comput Biol*. 14(9):1160-1175.
- Gabaldón, T, Dessimoz, C, Huxley-Jones, J, Vilella, AJ, Sonnhammer, EL and Lewis, S. 2009. "Joining forces in the quest for orthologs." *Genome Biol*. 10(9):403.
- Goffeau, A, Barrell, BG, Bussey, H, Davis, RW, Dujon, B, Feldmann, H, Galibert, F, Hoheisel, JD, Jacq, C, Johnston, M, Louis, EJ, Mewes, HW, Murakami, Y, Philippsen, P, Tettelin, H and Oliver, SG. 1996. "Life with 6000 genes." *Science*. 274:563-567.
- Goffeau, A. 2008. "Drug resistance: The fight against fungi." *Nature*. 452:541-542.
- Gojkovic, Z, Paracchini, S and Piskur, J. A new model organism for studying the catabolism of pyrimidines and purines. 431, 475-479. 1998.
- Gordon, JL, Byrne, KP and Wolfe, KH. 2009. "Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern *Saccharomyces cerevisiae* genome." *PLoS Genet*. 5(5):e1000485.
- Gordon, JL, Armisén, D, Proux-Wéra, E, Óhéigeartaigh, SS, Byrne, KP and Wolfe, KH. 2011. "Evolutionary erosion of yeast sex chromosomes by mating-type switching accidents." *Proceedings of the National Academy of Sciences*. 108:20024-20029.
- Hall, C, Brachat, S and Dietrich, FS. 2005. "Contribution of horizontal gene transfer to the evolution of *Saccharomyces cerevisiae*." *Eukaryot.Cell*. 4(6):1102-1115.
- Henricson, A, Forslund, K and Sonnhammer, EL. 2010. "Orthology confers intron position conservation." *BMC Genomics*. 11:412.
- Huerta-Cepas, J, Dopazo, H, Dopazo, J and Gabaldón, T. 2007. "The human phylome." *Genome Biol*. 8(6):R109.

- Huerta-Cepas, J, Dopazo, J, Huynen, MA and Gabaldón, T. 2011a. "Evidence for short-time divergence and long-time conservation of tissue-specific expression after gene duplication." *Brief.Bioinform.* 12(5):442-448.
- Huerta-Cepas, J, Bueno, A, Dopazo, Jn and Gabaldón, T. 2008. "PhylomeDB: a database for genome-wide collections of gene phylogenies." *Nucleic Acids Research.* 36:D491-D496.
- Huerta-Cepas, J, Capella-Gutierrez, S, Pryszcz, LP, Denisov, I, Kormes, D, Marcet-Houben, M and Gabaldón, T. 2011b. "PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions." *Nucleic Acids Research.* 39:D556-D560.
- Hughes, AL. 1994. "The Evolution of Functionally Novel Proteins after Gene Duplication." *Proceedings of the Royal Society of London.Series B: Biological Sciences.* 256:119-124.
- Hulsen, T, Huynen, MA, de, VJ and Groenen, PM. 2006. "Benchmarking ortholog identification methods using functional genomics data." *Genome Biol.* 7(4):R31.
- Huynen, MA and Bork, P. 1998. "Measuring genome evolution." *Proc.Natl.Acad.Sci.U.S.A.* 95(11):5849-5856.
- Inês Azevedo Isidro. 2009. "Update of the Yeast Transporter Information Database." Instituto superior técnico Universidade Técnica de Lisboa, 75p.
- Jackson, AP, Gamble, JA, Yeomans, T, Moran, GP, Saunders, D, Harris, D, Aslett, M, Barrell, JF, Butler, G, Citiulo, F, Coleman, DC, de Groot, PWJ, Goodwin, TJ, Quail, MA, McQuillan, J, Munro, CA, Pain, A, Poulter, RT, Rajandream, MA, Renauld, H, Spiering, MJ, Tivey, A, Gow, NAR, Barrell, B, Sullivan, DJ and Berriman, M. 2009. "Comparative genomics of the fungal pathogens *Candida dubliniensis* and *Candida albicans*." *Genome Research.* 19:2231-2244.
- Jansen, M, Veurink, JH, Euverink, GJ and Dijkhuizen, L. 2003. "Growth of the salt-tolerant yeast *Zygosaccharomyces rouxii* in microtiter plates: effects of NaCl, pH and temperature on growth and fusel alcohol production from branched-chain amino acids." *Fems Yeast Research.* 3:313-318.
- Katoh, K and Toh, H. 2008. "Recent developments in the MAFFT multiple sequence alignment program." *Briefings in Bioinformatics.* 9:286-298.
- Kellis, M, Birren, BW and Lander, ES. 2004. "Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*." *Nature.* 428:617-624.

- Kellis, M, Patterson, N, Endrizzi, M, Birren, B and Lander, ES. 2003. "Sequencing and comparison of yeast species to identify genes and regulatory elements." *Nature*. 423:241-254.
- Koonin, EV. 2005. "Orthologs, paralogs, and evolutionary genomics." *Annual Review of Genetics*. 39:309-338.
- Krogh, A, Brown, M, Mian, IS, Sjölander, K and Haussler, D. 1994. "Hidden Markov Models in Computational Biology: Applications to Protein Modeling." *Journal of Molecular Biology*. 235:1501-1531.
- Kurtzman, CP. 1995. "Relationships among the genera *Ashbya*, *Eremothecium*, *Holleya* and *Nematospora* determined from rDNA sequence divergence." *J.Ind.Microbiol.* 14(6):523-530.
- Kurtzman, CP. 2003. "Phylogenetic circumscription of *Saccharomyces*, *Kluyveromyces* and other members of the *Saccharomycetaceae*, and the proposal of the new genera *Lachancea*, *Nakaseomyces*, *Naumovia*, *Vanderwaltozyma* and *Zygorulasporea*." *Fems Yeast Research*. 4:233-245.
- Kurtzman, C.P., Fell, JW and Boekhout, T. 2011. "Definition, classification and nomenclature of the yeasts." *The yeasts a taxonomic study*. Elsevier. Amsterdam. 3-8.
- Kuzniar, A, van Ham, RCHJ, Pongor, S and Leunissen, JAM. 2008. "The quest for orthologs: finding the corresponding gene across genomes." *Trends in Genetics*. 24:539-551.
- Leandro, MJ, Fonseca, C and Gonçalves, P. 2009. "Hexose and pentose transport in ascomycetous yeasts: an overview." *Fems Yeast Research*. 9:511-525.
- Leh, V, Despons, L, Friedrich, A, Martin, T, Durrens, P, Casaregola, S, Neuveglise, C, Fairhead, C, Marck, C, Cruz, JA, Straub, ML, Kugler, V, Sacerdot, C, Uzunov, Z, Thierry, A, Weiss, S, Bleykasten, C, De, MJ, Jacques, N, Jung, P, Lemaire, M, Mallet, S, Morel, G, Richard, GF, Sarkar, A, Savel, G, Schacherer, J, Seret, ML, Talla, E, Samson, G, Jubin, C, Poulain, J, Vacherie, B, Barbe, V, Pelletier, E, Sherman, DJ, Westhof, E, Weissenbach, J, Baret, PV, Wincker, P, Gaillardin, C, Dujon, B and Souciet, JL. 2012. "*Pichia sorbitophila*, an Interspecies Yeast Hybrid, Reveals Early Steps of Genome Resolution After Polyploidization." *G3.(Bethesda.)*. 2(2):299-311.
- Lépingle, A, Casaregola, S, Neuveglise, C, Bon, E, Nguyen, H-V, Artiguenave, F, Wincker, P and Gaillardin, C. Genomic Exploration of the Hemiascomycetous Yeasts: 14. *Debaryomyces hansenii* var. *hansenii*. FEBS Letters 487, 82-86. 2000.

- Li, H, Coghlan, A, Ruan, J, Coin, LJ, Heriche, JK, Osmotherly, L, Li, R, Liu, T, Zhang, Z, Bolund, L, Wong, GK, Zheng, W, Dehal, P, Wang, J and Durbin, R. 2006. "TreeFam: a curated database of phylogenetic trees of animal gene families." *Nucleic Acids Res.* 34(Database issue):D572-D580.
- Li, L, Stoeckert, CJ and Roos, DS. 2003. "OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes." *Genome Research.* 13:2178-2189.
- Lodi, T, Diffels, J, Goffeau, A and Baret, PV. 2007. "Evolution of the carboxylate Jen transporters in fungi." *FEMS Yeast Res.* 7(5):646-656.
- Lynch, M and Conery, JS. 2000. "The Evolutionary Fate and Consequences of Duplicate Genes." *Science.* 290:1151-1155.
- Maresova, L and Sychrova, H. 2003. "Physiological characterization of osmotolerant yeast *Pichia sorbitophila* and comparison with a putative synonym *Pichia farinosa*." *Folia Microbiol.(Praha).* 48(2):211-217.
- Mi, H, Dong, Q, Muruganujan, A, Gaudet, P, Lewis, S and Thomas, PD. 2010. "PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium." *Nucleic Acids Research.* 38:D204-D210.
- Mi, H, Guo, N, Kejariwal, A and Thomas, PD. 2007. "PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways." *Nucleic Acids Research.* 35:D247-D252.
- Middelhoven, WJ, Hoogkamer-Te Niet, MC and Kreger-Van Rij, NJW. 1984. "Trichosporon adeninovorans sp. nov., a yeast species utilizing adenine, xanthine, uric acid, putrescine and primary n-alkylamines as the sole source of carbon, nitrogen and energy." *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology.* 50:369-378.
- Milinkovitch, MC, Helaers, R, Depiereux, E, Tzika, AC and Gabaldón, T. 2010. "2x genomes - depth does matter." *Genome Biology.* 11:R16.
- Møller, K, Sharif, MZ and Olsson, L. 2004. "Production of fungal α -amylase by *Saccharomyces kluyveri* in glucose-limited cultivations." *Journal of Biotechnology.* 111:311-318.
- Nehrt, NL, Clark, WT, Radivojac, P and Hahn, MW. 2011. "Testing the ortholog conjecture with comparative functional genomic data from mammals." *PLoS Comput Biol.* 7(6):e1002073.
- Nikolski, M and Sherman, DJ. 2007. "Family relationships: should consensus reign?--consensus clustering for protein families." *Bioinformatics.* 23:e71-e76.

- Notebaart, RA, Huynen, MA, Teusink, B, Siezen, RJ and Snel, B. 2005. "Correlation between sequence conservation and the genomic context after gene duplication." *Nucleic Acids Research*. 33:6164-6171.
- Ohno, S. 1970. *Evolution by Gene Duplication*. Springer-Verlag, New-York - Heidelberg - Berlin, 160p.
- Oliveira, RP, Lages, F and Lucas, C. 1996. "Isolation and characterisation of mutants from the halotolerant yeast *Pichia sorbitophila* defective in H⁺/glycerol symport activity." *FEMS Microbiol.Lett.* 142(2-3):147-153.
- Ostlund, G, Schmitt, T, Forslund, K, Kostler, T, Messina, DN, Roopra, S, Frings, O and Sonnhammer, EL. 2010. "InParanoid 7: new algorithms and tools for eukaryotic orthology analysis." *Nucleic Acids Res.* 38 (Database issue):D196-D203.
- Pearson, WR. 1990. "Rapid and sensitive sequence comparison with FASTP and FASTA." *Methods Enzymol.* 183:63-98.
- Peterson, ME, Chen, F, Saven, JG, Roos, DS, Babbitt, PC and Sali, A. 2009. "Evolutionary constraints on structural similarity in orthologs and paralogs." *Protein Sci.* 18(6):1306-1315.
- Piskur, J and Langkjaer, RB. 2004. "Yeast genome sequencing: the power of comparative genomics." *Molecular Microbiology*. 53:381-389.
- Powell, S, Szklarczyk, D, Trachana, K, Roth, A, Kuhn, M, Muller, J, Arnold, R, Rattei, T, Letunic, I, Doerks, T, Jensen, LJ, von, MC and Bork, P. 2012. "eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges." *Nucleic Acids Res.* 40(Database issue):D284-D289.
- Pribylová, L, Farkas, V, Slaninova, I, De Montigny, J and Sychrová, H. 2007. "Differences in osmotolerant and cell-wall properties of two *Zygosaccharomyces rouxii* strains." *Folia Microbiologica*. 52:241-245.
- Remm, M, Storm, CEV and Sonnhammer, ELL. 2001. "Automatic clustering of orthologs and in-paralogs from pairwise species comparisons." *Journal of Molecular Biology*. 314:1041-1052.
- Rockwell, N, Wolfger, H, Kuchler, K and Thorner, J. 2009. "ABC Transporter Pdr10 Regulates the Membrane Microenvironment of Pdr12 in *Saccharomyces cerevisiae*." *Journal of Membrane Biology*. 229:27-52.
- Rodrigues de Miranda, L, Appel, KR and Seyfarth, H. 1980. "*Pichia sorbitophila* sp nov." *Antonie Van Leeuwenhoek*. 46(2):157-159.
- Rolland, T and Dujon, B. 2011. "Yeasty clocks: dating genomic changes in yeasts." *C.R.Biol.* 334(8-9):620-628.

- Rolland, T, Neuveglise, C, Sacerdot, C and Dujon, B. 2009. "Insertion of Horizontally Transferred Genes within Conserved Syntenic Regions of Yeast Genomes." *PLoS ONE*. 4:e6515.
- Ruan, J, Li, H, Chen, Z, Coghlan, A, Coin, LJ, Guo, Y, Hériché, JK, Hu, Y, Kristiansen, K, Li, R, Liu, T, Moses, A, Qin, J, Vang, S, Vilella, AJ, Ureta-Vidal, A, Bolund, L, Wang, J and Durbin, R. 2008. "TreeFam: 2008 Update." *Nucleic Acids Research*. 36:D735-D740.
- Saier, MH. 2000. "A functional-phylogenetic classification system for transmembrane solute transporters." *Microbiology and Molecular Biology Reviews*. 64:354.
- Salichos, L and Rokas, A. 2011. "Evaluating Ortholog Prediction Algorithms in a Yeast Model Clade." *PLoS ONE*. 6:e18755.
- Schaffrath, R and Breunig, KD. 2000. "Genetics and molecular physiology of the yeast *Kluyveromyces lactis*." *Fungal Genet. Biol.* 30(3):173-190.
- Seiler, H and Busse, M. 1990. "The yeasts of cheese brines." *International Journal of Food Microbiology*. 11:289-303.
- Seret, ML. 25-8-2007. "Génomique comparative des levures: identification des orthologues par étude de l'environnement chromosomique." Diplôme d'Etudes Approfondies en Sciences Agronomiques et Ingénierie Biologique, Université Catholique de Louvain.
- Seret, ML, Diffels, JF, Goffeau, A and Baret, PV. 2009. "Combined phylogeny and neighborhood analysis of the evolution of the ABC transporters conferring multiple drug resistance in hemiascomycete yeasts." *BMC Genomics*. 10:459.
- Shi, G, Zhang, L and Jiang, T. 2010. "MSOAR 2.0: Incorporating tandem duplications into ortholog assignment based on genome rearrangement." *BMC Bioinformatics*. 11:10.
- Shi, G, Peng, MC and Jiang, T. 2011. "MultiMSOAR 2.0: An Accurate Tool to Identify Ortholog Groups among Multiple Genomes." *PLoS ONE*. 6:e20892.
- Smith, TF and Waterman, MS. 1981. "Identification of common molecular subsequences." *Journal of Molecular Biology*. 147:195-197.
- Sonnhammer, ELL and Koonin, EV. 2002. "Orthology, paralogy and proposed classification for paralog subtypes." *Trends in Genetics*. 18:619-620.

- Souciet, JL, Dujon, B, Gaillardin, C, Johnston, M, Baret, PV, Cliften, P, Sherman, DJ, Weissenbach, J, Westhof, E, Wincker, P, Jubin, C, Poulain, J, Barbe, V, Segurens, B, Artiguenave, F, Anthouard, V, Vacherie, B, Val, ME, Fulton, RS, Minx, P, Wilson, R, Durrens, P, Jean, G, Marck, C, Martin, T, Nikolski, M, Rolland, T, Seret, ML, Casaregola, S, Despons, L, Fairhead, C, Fischer, G, Lafontaine, I, Leh, V, Lemaire, M, De Montigny, J, Neuveglise, C, Thierry, A, Blanc-Lenfle, I, Bleykasten, C, Diffels, J, Fritsch, E, Frangeul, L, Goeffon, A, Jauniaux, N, Kachouri-Lafond, R, Payen, C, Potier, S, Pribylova, L, Ozanne, C, Richard, GF, Sacerdot, C, Straub, ML, Talla, E and Genolevures Consortium. 2009. "Comparative genomics of protoploid *Saccharomycetaceae*." *Genome Research*. 19:1696-1709.
- Souciet, JL, Aigle, M, Artiguenave, F, Blandin, G, Bolotin-Fukuhara, M, Bon, E, Brottier, P, Casaregola, S, de Montigny, J and Dujon, B. 2000. "Genomic Exploration of the Hemiascomycetous Yeasts: 1. A set of yeast species for molecular evolution studies." *FEBS Letters*. 487:3-12.
- Storm, CEV and Sonnhammer, ELL. 2002. "Automated ortholog inference from phylogenetic trees and calculation of orthology reliability." *Bioinformatics*. 18:92-99.
- Studer, RA and Robinson-Rechavi, M. 2009. "How confident can we be that orthologs are similar, but paralogs differ?" *Trends in Genetics*. 25:210-216.
- Talla, E, Anthouard, V, Bouchier, C, Frangeul, L and Dujon, B. 2005. "The complete mitochondrial genome of the yeast *Kluyveromyces thermotolerans*." *FEBS Lett*. 579(1):30-40.
- Tatusov, RL, Fedorova, ND, Jackson, JD, Jacobs, AR, Kiryutin, B, Koonin, EV, Krylov, DM, Mazumder, R, Mekhedov, SL, Nikolskaya, AN, Rao, BS, Smirnov, S, Sverdlov, AV, Vasudevan, S, Wolf, YI, Yin, JJ and Natale, DA. 2003. "The COG database: an updated version includes eukaryotes." *BMC Bioinformatics*. 4.
- Tatusov, RL, Galperin, MY, Natale, DA and Koonin, EV. 2000. "The COG database: a tool for genome-scale analysis of protein functions and evolution." *Nucleic Acids Res*. 28(1):33-36.
- Tatusov, RL, Koonin, EV and Lipman, DJ. 1997. "A genomic perspective on protein families." *Science*. 278(5338):631-637.
- Thevenieau, F, Le Dall, MT, Nthangeni, B, Mauersberger, S, Marchal, R and Nicaud, JM. 2007. "Characterization of *Yarrowia lipolytica* mutants affected in hydrophobic substrate utilization." *Fungal Genetics and Biology*. 44:531-542.

- Thomas, PD. 2010. "GIGA: A simple, efficient algorithm for gene tree inference in the genomic age." *BMC Bioinformatics*. 11.
- Thomas, PD, Wood, V, Mungall, CJ, Lewis, SE and Blake, JA. 2012. "On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report." *PLoS Comput Biol*. 8(2):e1002386.
- Thomas, PD, Campbell, MJ, Kejariwal, A, Mi, H, Karlak, B, Daverman, R, Diemer, K, Muruganujan, A and Narechania, A. 2003. "PANTHER: A Library of Protein Families and Subfamilies Indexed by Function." *Genome Research*. 13:2129-2141.
- Trachana, K, Larsson, TA, Powell, S, Chen, WH, Doerks, T, Muller, J and Bork, P. 2011. "Orthology prediction methods: a quality assessment using curated protein families." *Bioessays*. 33(10):769-780.
- van der Heijden, RT, Snel, B, van, N, V and Huynen, MA. 2007. "Orthology prediction at scalable resolution by phylogenetic tree analysis." *BMC Bioinformatics*. 8:83.
- van Hoof, A. 2005. "Conserved Functions of Yeast Genes Support the Duplication, Degeneration and Complementation Model for Gene Duplication." *Genetics*. 171:1455-1461.
- Vandepoele, K, Saeys, Y, Simillion, C, Raes, J and Van de Peer, Y. 2002. "The Automatic Detection of Homologous Regions (ADHoRe) and Its Application to Microcolinearity Between Arabidopsis and Rice." *Genome Research*. 12:1792-1801.
- Von Heijne, G. 2007. "The membrane protein universe: What's out there and why bother?" *Journal of Internal Medicine*. 261:543-557.
- Wall, DP, Fraser, HB and Hirsh, AE. 2003. "Detecting putative orthologs." *Bioinformatics*. 19:1710-1711.
- Wapinski, I, Pfeffer, A, Friedman, N and Regev, A. 2007. "Automatic genome-wide reconstruction of phylogenetic gene trees." *Bioinformatics*. 23:I549-I558.
- Wartmann, T and Kunze, G. 2000. "Genetic transformation and biotechnological application of the yeast <i>Arxula adenivorans</i>." *Applied Microbiology and Biotechnology*. 54:619-624.
- Wendland, J and Walther, A. 2005. "Ashbya gossypii: A model for fungal developmental biology." *Nature Reviews Microbiology*. 3:421-429.

Wolfe, KH. 2006. "Comparative genomics and genome evolution in yeasts." *Philosophical Transactions of the Royal Society B-Biological Sciences*. 361:403-412.

Wolfe, KH and Shields, DC. 1997. "Molecular evidence for an ancient duplication of the entire yeast genome." *Nature*. 387:708-713.

Wood, V, Gwilliam, R, Rajandream, MA, Lyne, M, Lyne, R, Stewart, A, Sgouros, J, Peat, N, Hayles, J, Baker, S, Basham, D, Bowman, S, Brooks, K, Brown, D, Brown, S, Chillingworth, T, Churcher, C, Collins, M, Connor, R, Cronin, A, Davis, P, Feltwell, T, Fraser, A, Gentles, S, Goble, A, Hamlin, N, Harris, D, Hidalgo, J, Hodgson, G, Holroyd, S, Hornsby, T, Howarth, S, Huckle, EJ, Hunt, S, Jagels, K, James, K, Jones, L, Jones, M, Leather, S, McDonald, S, McLean, J, Mooney, P, Moule, S, Mungall, K, Murphy, L, Niblett, D, Odell, C, Oliver, K, O'Neil, S, Pearson, D, Quail, MA, Rabinowitsch, E, Rutherford, K, Rutter, S, Saunders, D, Seeger, K, Sharp, S, Skelton, J, Simmonds, M, Squares, R, Squares, S, Stevens, K, Taylor, K, Taylor, RG, Tivey, A, Walsh, S, Warren, T, Whitehead, S, Woodward, J, Volckaert, G, Aert, R, Robben, J, Grymonprez, B, Weltjens, I, Vanstreels, E, Rieger, M, Schafer, M, Muller-Auer, S, Gabel, C, Fuchs, M, Fritz, C, Holzer, E, Moestl, D, Hilbert, H, Borzym, K, Langer, I, Beck, A, Lehrach, H, Reinhardt, R, Pohl, TM, Eger, P, Zimmermann, W, Wedler, H, Wambutt, R, Purnelle, B, Goffeau, A, Cadieu, E, Dreano, S, Gloux, S, Lelaure, V, Mottier, S, Galibert, F, Aves, SJ, Xiang, Z, Hunt, C, Moore, K, Hurst, SM, Lucas, M, Rochet, M, Gaillardin, C, Tallada, VA, Garzon, A, Thode, G, Daga, RR, Cruzado, L, Jimenez, J, Sanchez, M, del Rey, F, Benito, J, Dominguez, A, Revuelta, JL, Moreno, S, Armstrong, J, Forsburg, SL, Cerrutti, L, Lowe, T, McCombie, WR, Paulsen, I, Potashkin, J, Shpakovski, GV, Ussery, D, Barrell, BG and Nurse, P. 2002. "The genome sequence of *Schizosaccharomyces pombe*." *Nature*. 415:871-880.

Yamamoto, K, Nakayama, A, Yamamoto, Y and Tabata, S. 2004. "Val216 decides the substrate specificity of α -glucosidase in *Saccharomyces cerevisiae*." *European Journal of Biochemistry*. 271:3414-3420.

Zhang, J. 2003. "Evolution by gene duplication: an update." *Trends in Ecology & Evolution*. 18:292-298.

Zmasek, CM and Eddy, SR. 2002. "RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs." *BMC Bioinformatics*. 3:14.

ANNEXES

**ANNEXE 1 : COMPARATIVE GENOMICS OF PROTOPLOID
SACCHAROMYCETACEAE**

Comparative genomics of protoploid *Saccharomycetaceae*

The Génolevures Consortium¹

Our knowledge of yeast genomes remains largely dominated by the extensive studies on *Saccharomyces cerevisiae* and the consequences of its ancestral duplication, leaving the evolution of the entire class of hemiascomycetes only partly explored. We concentrate here on five species of *Saccharomycetaceae*, a large subdivision of hemiascomycetes, that we call "protoploid" because they diverged from the *S. cerevisiae* lineage prior to its genome duplication. We determined the complete genome sequences of three of these species: *Kluyveromyces (Lachancea) thermotolerans* and *Saccharomyces (Lachancea) kluyveri* (two members of the newly described *Lachancea* clade), and *Zygosaccharomyces rouxii*. We included in our comparisons the previously available sequences of *Kluyveromyces lactis* and *Ashbya (Eremothecium) gossypii*. Despite their broad evolutionary range and significant individual variations in each lineage, the five protoploid *Saccharomycetaceae* share a core repertoire of approximately 3300 protein families and a high degree of conserved synteny. Synteny blocks were used to define gene orthology and to infer ancestors. Far from representing minimal genomes without redundancy, the five protoploid yeasts contain numerous copies of paralogous genes, either dispersed or in tandem arrays, that, altogether, constitute a third of each genome. Ancient, conserved paralogs as well as novel, lineage-specific paralogs were identified.

[Supplemental material is available online at <http://www.genome.org> and at <http://www.genolevures.org/>. The sequence data for *Zygosaccharomyces rouxii* and *Kluyveromyces thermotolerans* have been submitted to EMBL-Bank (<http://www.ebi.ac.uk/emb/>) under accession nos. CU928173–CU928176, CU928178, CU928179, CU928181 and CU928165–CU928171, and CU928180, respectively. *Saccharomyces kluyveri* sequences were deposited to GenBank under accession no. AAC03000000.]

Yeasts have played a critical role in our understanding of molecular function and evolution in eukaryotes. Their small, compact genomes, their importance in a variety of fermentation processes, and the facility of manipulating them in the laboratory have led to the determination and analysis of the genome sequences of several yeast species (for review, see Dujon 2005, 2006; Kurtzman and Piskur 2006; Scannell et al. 2007a). Molecular phylogenies suggest that unicellular yeasts arose from ancestral fungal lineages several times independently, from *Basidiomycota* leading, for example, to the *Cryptococcus* and *Malassezia* species (Loftus et al. 2005; Xu et al. 2007), and from *Ascomycota* leading to the well-studied *Schizosaccharomyces* species, on one hand (Wood et al. 2002; Aslett and Wood 2006), and to the large, homogeneous class known as hemiascomycetes, or budding yeasts, on the other hand. The hemiascomycetes have enjoyed most of the attention of genomic studies, owing in part to their large number of species (more than a thousand are described [Kurtzman and Fell 2006] and many more likely exist [Boekhout 2005]), but mostly because they include *Saccharomyces cerevisiae*, the first eukaryote sequenced (Goffeau et al. 1996), and a favored experimental model for functional genomics (Hofmann et al. 2003; Ooi et al. 2006).

An exploration of 13 hemiascomycete genomes revealed their broad evolutionary range (Souciet et al. 2000), and several internal subdivisions exist, three of which have been characterized by complete sequencing of the genomes of a number of species (Dietrich et al. 2004; Dujon et al. 2004; Jones et al. 2004; Kellis et al. 2004). *Yarrowia lipolytica*, with a GC-rich genome roughly twice as large as that of other hemiascomycetous yeasts, is the representative

species of the first subdivision. It contains more protein-coding genes and more introns than other yeasts and has a larger variety of transposable elements. It also has several peculiarities, such as multiple subtelomeric rDNA loci and dispersed 5S RNA coding genes, half of which are transcriptionally fused to tRNA genes (Acker et al. 2008). A second subdivision of hemiascomycetes consists of species that translate CTG codons as serine rather than leucine, a reassignment believed to have occurred more than 170 million years ago (Miranda et al. 2006). This subdivision has been intensely studied because it contains the pathogenic yeast *Candida albicans* (Jones et al. 2004; Noble and Johnson 2007) and related *Candida* species (Magee et al. 2008). It also contains *Debaryomyces hansenii*, whose genome shows numerous tandem gene arrays (TGAs) that contribute to gene family expansion (Dujon et al. 2004). A third subdivision consists of yeasts of the "Saccharomyces complex" or *Saccharomycetaceae*, the subdivision of hemiascomycetes with the most genome sequences available. This subdivision contains a large variety of species that, although sharing a number of common physiological and genomic properties, represent a broad phylogenetic range. They were recently classified into 14 distinct clades (Kurtzman 2003; Kurtzman and Robnett 2003) whose genomes remain unequally explored.

The importance of *S. cerevisiae* as a model system (for review, see Barnett 2007; Replansky et al. 2008) has resulted in most genomic studies focusing on closely related species leaving other clades of *Saccharomycetaceae* relatively unexplored, with only few exceptions (Dietrich et al. 2004; Dujon et al. 2004; Kellis et al. 2004). Today, genome sequences are available (at various coverages) for *Saccharomyces paradoxus*, *Saccharomyces mikatae*, *Saccharomyces kudriavzevii*, *Saccharomyces bayanus*, *Saccharomyces exiguus*, *Saccharomyces servazii*, and *Saccharomyces castellii* (Bon et al. 2000a,b; Casaregola et al. 2000; Cliften et al. 2003, 2006; Kellis et al. 2003), and several strains of *S. cerevisiae* have been entirely sequenced (Wei et al. 2007; Borneman et al. 2008; http://www.broad.mit.edu/annotation/genome/saccharomyces_cerevisiae/).

¹A complete list of authors and affiliations appears at the end of the paper, before the Acknowledgments section.

^{1a}Corresponding author.

E-mail jean-luc.souciet@gem.u-strasbg.fr; fax 33-3-90-24-20-28.

Article published online before print. Article and publication date are at <http://www.genome.org/cgi/doi/10.1101/gr.091546.109>. Freely available online through the Genome Research Open Access option.

This offers a unique view on the population structure and reproductive preferences of a group of yeasts playing an important role in fermentation processes, and a powerful tool for quantitative trait analyses (Demogines et al. 2008). Recently, wild populations of *S. cerevisiae* and *S. paradoxus* have also been characterized by hybridization to genome tiling oligonucleotide arrays and by partial genome sequencing (Liti et al. 2009; Schacherer and Kruglyak 2009).

The recognition of a whole-genome duplication in the ancestry of *S. cerevisiae* (Wolfe and Shields 1997) prompted many studies on the evolutionary and functional consequences of such events in yeasts (Wong et al. 2002; Byrnes et al. 2006) and in other eukaryotes such as plants (*Arabidopsis* Genome Initiative 2000; Yu et al. 2005; Jaillon et al. 2007), fishes (Jaillon et al. 2004), or ciliates (Aury et al. 2006). The subsequent loss of redundant gene copies, which may lead to speciation bursts (Scannell et al. 2006), destroys original gene neighborhood relationships and reshapes the genetic maps of post-duplication species, leaving typical dual synteny patterns (Dietrich et al. 2004; Jaillon et al. 2004; Kellis et al. 2004). Inference of the original gene order has been attempted by analyzing duplicated genomes (Byrne and Wolfe 2005; Byrnes et al. 2006; Scannell et al. 2007b; Conant and Wolfe 2008), but an authentic image of the original genome organization of this important group of yeasts is desirable.

We report the complete sequence and manual annotation of the genomes of three novel yeast species, *Zygosaccharomyces rouxii*, a member of the *Zygosaccharomyces* clade, *Saccharomyces (Lachancea) kluyveri*, and *Kluyveromyces (Lachancea) thermotolerans*, two members of the *Lachancea* clade. We have compared these genomes with those of *Kluyveromyces lactis* and *Ashbya (Eremothecium) gossypii* (Dietrich et al. 2004; Dujon et al. 2004), providing a multispecies comparison among nonduplicated *Saccharomycetaceae*. We discovered that, far from being minimal genomes, these species contain many paralogous genes, which are often highly diverged in sequence and represent approximately a third of their total genes. We also found that, despite sharing significant conservation of synteny and a common protein repertoire of approximately 3300 families, these species span a large evolutionary range as evidenced from their important sequence divergence.

Results

General sequencing strategy and brief description of the genome sequences

Rationale of species selection

Our goal was to characterize and compare gene content and genome organization of the *Saccharomycetaceae* (as defined by Kurtzman and Robnett 2003) that did not experience the ancestral whole-genome duplication that sculpted the genomes of *S. cerevisiae* and its closest relatives. To this end, we determined the complete DNA sequence of three novel yeast genomes, *Z. rouxii*, *K. thermotolerans*, and *S. kluyveri* (see Table 1), and included in our analyses the previously published sequences of *K. lactis* (Dujon et al. 2004) and *A. gossypii* (Dietrich et al. 2004).

These five yeast species represent four distinct clades of *Saccharomycetaceae*, respectively, designated as *Zygosaccharomyces*, *Lachancea*, *Kluyveromyces*, and *Eremothecium* (Fig. 1A). Given their complex phylogenetic relationship within the *Saccharomycetaceae*, we designate them collectively as “protoploid” solely to distinguish them from the “duplicated” yeasts (clades 1–6) (Kurtzman 2003). Other protoploid genomes, such as *Kluyveromyces waltii* (Kellis et al. 2004) and *Kluyveromyces marxianus* (Llorente et al. 2000a), were not considered since our comparisons required complete genome sequences with a single contig per chromosome.

The five yeast species studied here show diverse biological and metabolic properties. *Z. rouxii* is an osmo- and halotolerant species used in some fermentation processes (Solieri and Giudici 2007). It is able to grow on high concentrations of salt and/or sugar (Jansen et al. 2003) and is often considered as a food-spoiling agent. The type strain (CBS732) is haploid, but some wild isolates of *Z. rouxii* are diploid (Solieri et al. 2008). *K. thermotolerans* is usually associated with fruits or with insects feeding on plants, but one isolate was recently obtained from a wine re-fermentation process (Vilela-Moura et al. 2008). *S. kluyveri* is often used in protein production because it wastes less glucose by aerobic fermentation than does *S. cerevisiae*, and hence is better at biomass production (Møller et al. 2002, 2004). It has been developed as a model organism to study a variety of biological processes such as pyrimidine degradation (Beck et al. 2008), metabolic flux (Blank et al. 2005), or fatty acid desaturases (Oura and Kajiwaru 2008). *S. kluyveri* appears widespread in the environment, with strains isolated from insect guts, soil, or trees in North America, Europe, and India. *K. lactis* and *A. gossypii*, are respectively, a haploid, lactose-utilizing yeast (Dujon et al. 2004), and a haploid phytopathogen able to form filaments (Dietrich et al. 2004).

Overall genome anatomy, composition, and duplications

We annotated the three novel yeast genomes sequenced here and the previously sequenced genome of *K. lactis* (see Methods and <http://www.genolevures.org/>). Annotations of *A. gossypii* were taken from AGD (<http://agd.vital-it.ch/index.html>). Table 2 lists major features and properties of these five genomes. For comparison, the genomes of *S. cerevisiae* (Goffeau et al. 1996) and *Candida glabrata*

Table 1. Sequencing and assembly of new yeast genomes

Species	Strain	No. of chromosomes	Genome ploidy	Shotgun coverage	No. of final contigs	Final assembly size (nuc.)
<i>Z. rouxii</i>	CBS 732 ^a	7	n	11.1×	7	9,764,635
<i>K. thermotolerans</i>	CBS 6340 ^a	8	2n	12.3×	8	10,392,862
<i>S. kluyveri</i>	CBS 3082	8	2n		8	11,345,820

All chromosomes were assembled as unique contigs from one end to the other, except for chromosome E of *Z. rouxii* and H of *K. thermotolerans* and *S. kluyveri*, each assembled as two contigs separated by rDNA repeats. Two tandem copies of rDNA repeat units, flanked by a series of Ns to constitute a total of 17 kb, were manually inserted into the final assembly of these chromosomes. Assembled contigs correspond to entire chromosome sequences with the exception of the telomeric repeats. Sequences of *Z. rouxii* and *K. thermotolerans* were deposited to EMBL (accession nos. CU928173, CU928174, CU928175, CU928176, CU928178, CU928179, CU928181, and CU928165, CU928166, CU928167, CU928168, CU928169, CU928170, CU928171, and CU9 28180, respectively). Sequences of *S. kluyveri* were deposited to GenBank under accession no. AAC03000000. Assemblies from diploid strains did not reveal significant heterozygosity. The assembled sequences of *S. kluyveri* and *K. thermotolerans* contain the *MATa* sequence on chromosomes C and F, respectively. A *MATalpha* sequence has been recovered from the short, unassembled contigs of *S. kluyveri* (Payen et al. 2009). Other short, unassembled contigs of *Z. rouxii* correspond to mtDNA (data not shown) and to the pSR1 plasmid (Araki et al. 1985). MtDNA of *K. thermotolerans* was previously published (Talla et al. 2005).

^aGenome surveys of the same strains were previously published for *Z. rouxii* (de Montigny et al. 2000) and *K. thermotolerans* (Malpertuy et al. 2000a).

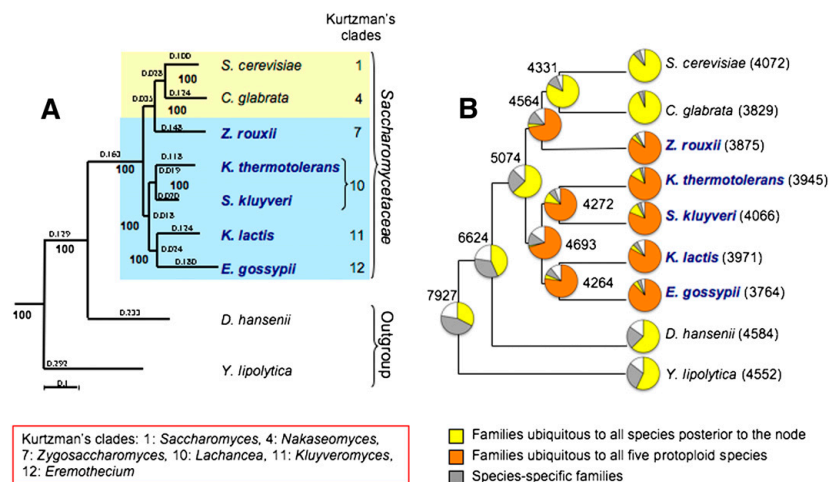


Figure 1. Phylogeny and protein-coding repertoire of *Saccharomycetaceae* and outgroups. (A) The phylogenetic tree results from the alignment of 180 proteins (66,709 amino acids), selected from all universal one-member families having a homolog in *S. pombe*, using the MAFFT algorithm (Kato et al. 2005), cleaned with Gblocks (Castresana 2000) before concatenation. Only families for which the ratio between the cleaned blocks and the initial alignment was higher than 75% were considered. The tree was constructed by maximum likelihood using PhyML (Guindon and Gascuel 2003) with a JTT substitution model corrected for heterogeneity among sites by a gamma-law distribution using four different categories of evolution rates. The proportion of invariable sites and the alpha-parameter of the gamma-law distribution were optimized according to the data. Branch length is indicated *above* or *next* to each branch, and bootstrap values (in bold) *next* to each node. *S. pombe* was used as an anchoring outgroup. Clades number and designation for *Saccharomycetaceae* are according to Kurtzman (2003) and Kurtzman and Robnett (2003). Protoploid species are highlighted (bold blue names). (B) Figures represent the total number of protein families in each species or node (pan-proteome, defined as the sum of all families present in all species posterior to the node). Pie charts indicate the proportion of families classified as “ubiquitous” (core-proteome, common to all species posterior to the node), “species-specific” (present in only one of the species posterior to the node), or other combinations. In orange, the proportion of families shared by all five protoploid species. Note the similar pie charts for the *Saccharomycetaceae* species, compared to the different pie charts of the outgroup.

(Dujon et al. 2004) were included as representative of post-genome duplication species of *Saccharomycetaceae*, while the genomes of *Debaryomyces hansenii* and *Yarrowia lipolytica* (Dujon et al. 2004) served as outgroups. The five protoploid genomes vary in size (from 8.7 Mb to 11.3 Mb) and are smaller than the two post-duplication genomes and the two outgroup genomes. Genome size reduction in *Z. rouxii* and *A. gossypii* is accompanied by an increased gene density (76.1% and 79.6% compared to 69.2% to 72.3% for the three other protoploid yeasts).

Two species, *K. thermotolerans* and *A. gossypii*, show a markedly higher GC content than other yeasts, *Y. lipolytica* excepted (Table 2). In every species, the nucleotide composition is generally uniform across the entire genome, despite local fluctuations in some cases (Supplemental Fig. 1). *S. kluyveri* represents an astonishing exception to this rule. An ~1-Mb-long region of chromosome C displays a GC content 12% higher than the rest of the genome (25% higher for the third codon position of CDS). The remaining 250-kb portion of this chromosome is identical in nucleotide composition to the rest of the genome. This heterogeneity is studied in detail in a separate work (Payen et al. 2009). In brief, it is found that the GC-rich chromosomal segment shows conserved synteny with *K. thermotolerans*, is devoid of transposable elements, and replicates later than other chromosomes during the S phase.

The protoploid yeast species have six, seven, or eight chromosomes. A single centromere composed of three short elements (CDEI, CDEII, and CDEIII) was identified for each chromosome of *Z. rouxii*, *K. thermotolerans*, and *S. kluyveri* (Supplemental Table 1; Supplemental Fig. 2B). Centromeres of *K. lactis* and *A. gossypii* were

previously described (Dietrich et al. 2004; Dujon et al. 2004). Interestingly, the AT-rich CDE II spacers vary in length (Supplemental Fig. 2A). The correspondence between centromeres of the five protoploid *Saccharomycetaceae* genomes could be deduced from synteny conservation of flanking genes (Supplemental Table 2). In all likelihood, the protoploid ancestor was a species with eight chromosomes. The eight centromeres of *K. thermotolerans* and *S. kluyveri* show a simple one-to-one congruence. Non-ambiguous correspondence with the former two species is also found for the seven centromeres of *Z. rouxii* and *A. gossypii*. The case of *K. lactis* is more complex with only six chromosomes and multiple rearrangements around several centromeres.

As anticipated from their phylogenetic position among the *Saccharomycetaceae*, the five yeast species studied here do not show traces of the whole-genome duplication that occurred in the ancestry of clades 1–6 (Kurtzman and Robnett 2003). Similarly, we can rule out the possible occurrence of another genome duplication in either of the four clades studied here since we could not find evidence of dual synteny in all pairwise map comparisons between our five species. Taking advantage of this fact, we have examined the possible presence of segmental duplications in these genomes (Supplemental Table 3). If one ignores known transposable elements and subtelomeric regions, only a few cases of segmental duplication are found in the genomes of *K. thermotolerans*, *S. kluyveri*, and *K. lactis*, and none in *Z. rouxii* or *A. gossypii*. Such a paucity of segmental duplications is surprising in view of the high frequency (10^{-7} per mitosis) of their spontaneous formation in *S. cerevisiae* (Payen et al. 2008), and contrasts with their abundance

Table 2. Summary of annotated features in yeast genomes of interest

Species	Strain	No. of chromosomes	Genome size (Mb)	Average G+C content (%)	Total no. of CDS	Genome coding coverage (%)	Gene density (no. of CDS per 10 kb)	Average G+C in CDS (%)	Average CDS size (codons)	Total tRNA genes ^a	Total snRNA genes ^b	Total snoRNA genes ^b	No. of rDNA clusters ^c
<i>S. cerevisiae</i>	S288c	16	12.1	38.3	5769	70.0	4.8	40.3	488	274	6	77	1 (I)
<i>C. glabrata</i>	CBS138	13	12.3	38.8	5204	64.2	4.2	41.7	507	207	6	ND	2 (S)
<i>Z. rouxii</i>	CBS732	7	9.8	39.1	4992	76.1	5.1	40.2	497	272	5	44	1 (I)
<i>K. thermotolerans</i>	CBS6340	8	10.4	47.3	5094	72.3	4.9	49.2	492	229	5	43	1 (I)
<i>S. kluyveri</i>	CBS3082	8	11.3	41.5	5320	69.6	4.7	43.1	497	257	5	43	1 (I)
<i>K. lactis</i>	CBS2359	6	10.7	38.8	5076	69.2	4.8	40.5	485	163	5	43	1 (I)
<i>A. gossypii</i>	ATCC10895	7	8.7	52.0	4715	79.6	5.4	52.0	491	190	5	79	1 (I)
<i>D. hansenii</i>	CBS767	7	12.2	36.3	6395	74.2	5.2	38.0	479	200	5	ND	3 (I)
<i>Y. lipolytica</i>	CBS7504	6	20.5	49.0	6580	46.0	3.1	53.8	489	510	6	ND	6 (S)

Data from *S. cerevisiae* were taken from SGD (<http://www.yeastgenome.org/>); *C. glabrata*, *D. hansenii*, and *Y. lipolytica* from Génolevures (<http://www.genolevures.org/>); and *A. gossypii* from AGD (<http://agd.vital-it.ch/index.html>). Annotations for *Z. rouxii*, *K. thermotolerans*, *S. kluyveri*, and *K. lactis* are part of this work and are available from the Génolevures online database (<http://www.genolevures.org/>). ND, Not determined.

^aSee Supplemental Tables 4 and 5 for details.

^bSee Supplemental Table 6 for details.

^crDNA clusters may be internal to chromosome arms (I) or in subtelomeric (S) locations.

in mammalian genomes (for review, see Samonte and Eichler 2002) (see Discussion).

Genome content: Protein-coding genes and spliceosomal introns

Protein-coding genes were identified as described in Methods. The protoploid genomes contain from 4715 (*A. gossypii*) to 5320 (*S. kluyveri*) protein-coding genes (Table 2). This is 8%–18% less than *S. cerevisiae*, which, itself, contains ~10% less genes than the two species used here as outgroups, *D. hansenii* and *Y. lipolytica*. With approximately 5000 protein-coding genes on average, protoploid yeasts have the smallest known gene set among hemiascomycetes. Protein size distributions and global amino acid compositions are similar for the nine yeasts considered here (mean: 492 ± 14 amino acids; median: 410 ± 11 amino acids). Spliceosomal introns are rare (3%–6% of protein-coding genes are interrupted by introns) in these five yeasts, and very few genes contain two introns (four in *Z. rouxii*, 11 in *S. kluyveri*). Splice sites and branch-point sequence motifs are similar to those of *S. cerevisiae*, and intron lengths are shorter, except for *K. lactis*. Additional information on yeast spliceosomal introns can be found at <http://genome.jouy.inra.fr/genosplicing/>.

tRNA-coding genes

The variety of tRNA species (identified by their anticodon), number of genes, and corresponding codon usage are given in Supplemental Table 4. The five protoploid yeasts follow the tRNA splicing rules previously described for other hemiascomycetes, enabling them to interpret the entire genetic code with only 43 or 44 tRNA species instead of the complete eukaryotic set of 46 (Marck et al. 2006). As expected for yeasts, a significant proportion (25%–35%) of tDNAs harbor an intron, all of which are short (except for the tDNA-Leu [CAG], where introns are 288, 150, and 134 nucleotides [nt] long for *K. lactis*, *A. gossypii*, and *S. kluyveri*, respectively).

In each yeast, most species of tRNA molecules are encoded by more than one gene (up to 17 copies), and the different paralogous gene copies are identical, or almost identical, in sequence, that is, there exists only one tRNA species for each anticodon. Two exceptions are notable here: (1) in *Z. rouxii*, the two copies of

tDNA-Val (TAC) differ from each other at 26 positions; (2) in *S. kluyveri*, one of the three copies of tDNA-Glu (CTC) differs by 17 nt changes from the two other copies of tDNA-Glu (CTC). Interestingly, this tDNA, which is located on the GC-rich left arm of chromosome C, results from a silent T-to-C mutation in the anticodon but is otherwise identical in sequence to the 13 tDNA-Glus (TTC).

As in other yeasts, tDNAs are dispersed throughout the genomes. Distances between two successive tDNAs along chromosomes vary from a few hundred base pairs to >300 kb (median values 20–30 kb). Clusters of tDNAs are rare in yeasts and tandem arrays of identical tDNAs even rarer. Thus, it is worth mentioning the array of five tandemly repeated tDNA-Glus (CTC), each separated by ~1 kb, that lies on chromosome B of *K. thermotolerans*. As previously described for other hemiascomycetes (Marck et al. 2006; Acker et al. 2008), probable di-cistronic tDNA pairs are also found in the yeast genomes we analyzed (Supplemental Table 5). In all cases, the intervals between the two successive tDNAs, that are always co-oriented, are very short and lack Pol III terminators. Some pairs, such as Arg (TCT)–Asp (GTC) or His (GTG)–Val (AAC), are common to several yeast species and probably result from ancestral fusions. Others appear specific to individual lineages.

Genes for other noncoding RNA molecules

A single rDNA locus is found in each protoploid species (Table 2). It is located within a chromosome arm, as in *S. cerevisiae*, not in a subtelomeric position as in some other yeasts. In each case, a gene encoding the 5S RNA molecule is part of the rDNA repeat, in opposite orientation of the 35S transcript (precursor of 18S, 5.8S, and 26S RNAs).

Genes for other major noncoding RNA molecules were annotated as described in Methods (Supplemental Table 6). The five spliceosomal RNAs (U1, U2, U4, U5, U6 snRNAs) are conserved in structure and size, and each is encoded by a single gene in the five yeast species. The same holds true for the U3 snoRNA, except for an interchromosomal duplication that created two identical genes in *Z. rouxii*. Similarly, the RNA moieties of the RNase P and SRP complexes are each encoded by a single gene in all yeast species, but their sequences and sizes differ owing to insertions of variable lengths around the conserved structural core. C/D snoRNA genes are highly

conserved. (H/ACA snoRNAs were not systematically investigated because of their variable size and poor sequence conservation.) The five polycistronic clusters, which encode 17 snoRNA molecules, are conserved. Also conserved are the five snoRNA genes embedded in introns of protein-coding genes. Finally, nearly all of the 21 monocistronic C/D snoRNA genes were unambiguously identified in each yeast, including snR52 transcribed by RNA polymerase III. The three absent cases (snR4 in *K. lactis* and *S. kluyveri*, snR50 in *K. thermotolerans*, and snR62 in *Z. rouxii*) were missed, probably owing to extensive sequence divergence and lack of synteny. We did not precisely define telomerase RNAs (beside the two previously annotated ones) (McEachern and Blackburn 1995; Chappell and Lundblad 2004), owing to their extensive sequence variation.

Transposable elements

Very few transposable elements are present in the five yeast genomes we analyzed (Supplemental Table 7). *Z. rouxii* and *A. gossypii* have a few degenerate Ty3-like elements and no other class I elements (retro-elements). *K. thermotolerans*, *S. kluyveri*, and *K. lactis* have several solo-LTRs of Ty1-like elements, but only *S. kluyveri* has intact copies of that retro-element. In *K. thermotolerans*, there are only two degenerated Ty1 copies, and both are located in subtelomeric regions, as previously observed for *K. lactis* (Fairhead and Dujon 2006).

Our greatest surprise came from the presence of a novel class II element in the protoploid yeast genomes. Such elements have only been found so far in *C. albicans* (Goodwin et al. 2001, 2007) and *Y. lipolytica* (Neuvéglise et al. 2005). Now, we found sequences similar to the hAT DNA transposon of plants and fungi (Rubin et al. 2001) in *K. thermotolerans*, *S. kluyveri*, *K. lactis*, and *A. gossypii* (Supplemental Table 7). Four full-length elements of this family (called *Rover*) were recognized in total. All carry a single CDS (between 631 and 867 codons), possess Terminal Inverted Repeats (TIR), and create 8-bp target site duplications. The presence of this element in four yeasts only, suggests an invasive transfer (possibly in the common ancestor of clades 10, 11, and 12 of *Saccharomycetaceae*, as the element is not found in *Z. rouxii*). Judging from the limited number of *Rover* sequences or their remnants, this element had a limited evolutionary success.

Protein families and functional repertoire

Overall description of protein families and core protein repertoire

The predicted proteomes of the five protoploid yeast species (see Methods) were classified into families based on all-to-all sequence comparisons and consensus clustering, as previously described (Nikolski and Sherman 2007). Proteomes of *S. cerevisiae*, *C. glabrata*, *D. hansenii*, and *Y. lipolytica* were included in the comparisons, bringing the total to 48,889 yeast proteins. Among these, 98% could be partitioned into 7927 protein families (Table 3). The remaining 1015 proteins (2%) could not be unambiguously classified using our clustering parameters. The list of protein families and their content can be found at <http://www.genolevures.org/>. A third of the protein families are

common to all nine yeasts. In each species, they are represented either by a single gene (1689 families) or by several paralogous genes (902 families). Another 45% of protein families are species-specific, and the remaining 22% are represented in various subsets of species. We examined the distribution of the three classes of families in the different yeast species and at each evolutionary node according to the phylogenetic tree (Fig. 1B). As expected, most species-specific families originate from the two outgroup species. Yet about a quarter of the 5074 families found in all *Saccharomycetaceae* is made up of species-specific proteins. The five protoploid proteomes are quite homogeneous, with only small proportions of species-specific families. They share additional protein families in addition to the ones common to all nine species, bringing the total of their common families to 3295. This “core protein repertoire” represents 81%–88% of the protein families present in each of the five protoploid species. As expected, most families of this core repertoire are represented in *S. cerevisiae* and *C. glabrata*, making it a characteristic feature of *Saccharomycetaceae* whose functions are worth elucidating.

Functional categorization of the core protein repertoire

Because few functional data are available for yeasts other than *S. cerevisiae*, we restricted our analysis of the core repertoire to families having, among them, at least one *S. cerevisiae* representative with a functional annotation. A total of 4097 distinct *S. cerevisiae* proteins with an associated GO-term for biological processes were retrieved from GO resources at <http://www.yeastgenome.org/>. For each of the 32 informative GO-terms recovered, the proportion of families belonging to the core repertoire was computed (Supplemental Fig. 3). Genes involved in ribosome biogenesis, protein translation or modification, transport, RNA metabolism, or cellular respiration are highly represented among the core repertoire families. In contrast, processes such as sporulation, meiosis, or conjugation are more frequently based on species- or lineage-specific genes.

Paralogous genes, genome redundancy, and tandem arrays

Paralogous genes

As noted above, a large number of protein families (1479) are represented by more than one protein in a given yeast species, revealing paralogs derived from ancestral duplications. Figure 2A

Table 3. Numerical distribution of protein families and corresponding numbers of protein-coding genes

	Families present in all nine yeast species	Families present in a subset of species (2–8)	Families present in only one species	Total
Families with ≤1 gene	1689	1416^a	3343	6448
per species	15,201	7640	3343	26,184
Families with >1 gene	902	362^b	215	1479
per species	17,518	3378	794	21,960
Total	2591	1778	3558	7927
	32,719	11,018	4137	47,874

Protein families were computed from the nine yeast species as explained in Methods, and were classified according to their presence in all species (column 1), a subset of species (column 2), or only one species (column 3). In each case, a family may be represented by one (line 1) or by several genes (line 2). For each category, the table indicates the total number of distinct protein families (bold) and the corresponding number of protein-coding genes. The complete list of families and corresponding genes can be found at <http://www.genolevures.org/>.

^aIncluding 498 families (1562 genes) absent from *S. cerevisiae*.

^bIncluding 99 families (607 genes) absent from *S. cerevisiae*.

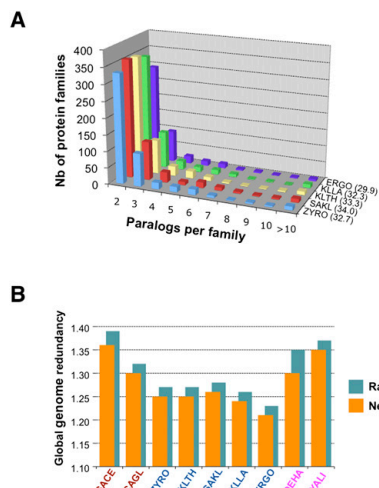


Figure 2. Genome redundancy. (A) Shown for each yeast species is the total number of protein families distributed according to their size (nb of paralogs per family; families of one member are not shown). The proportion (in percent) of the number of CDS belonging to multigene families to the total number of CDS is indicated in brackets next to the species abbreviation. (B) Compared genome redundancies for all nine yeast genomes. (SACE) *S. cerevisiae*; (CAGL) *C. glabrata*; (ZYRO) *Z. rouxii*; (KLTH) *K. thermotolerans*; (SAKL) *S. kluyveri*; (KLLA) *K. lactis*; (ERGO) *A. gossypii*; (DEHA) *D. hansenii*; (YALI) *Y. lipolytica*. Global genome redundancy is calculated as the ratio of total number of protein-coding genes in a genome vs. the total number of protein families in the same species. Raw redundancy counts all gene copies within tandem gene arrays (TGA); net redundancy considers only one gene-equivalent per TGA.

shows that the distributions of protein family sizes for the five protoploid species are globally similar. With 300–350 families of two paralogous genes per species, 95–115 families of three paralogs, and a few larger families, a total of 30%–34% of protein-coding genes are members of multigene families in the protoploid genomes (compared to 44% for *S. cerevisiae*). Thus, ancestral duplications of various sorts (see below) contribute to a third of the genomes of *Saccharomycetaceae*, while the whole-genome duplication in the ancestry of *S. cerevisiae* contributes to only an additional 10% of paralogs. We also compared the global genome redundancy of all nine yeast species (Fig. 2B) (note that redundancy only refers here to sequence similarity, not to function). The five protoploid yeasts are similar in this regard (slightly lower for *A. gossypii*) and different from other yeasts. The higher genome redundancy of *S. cerevisiae* and *C. glabrata* reflects their ancestral whole-genome duplication. In contrast, the higher genome redundancy of *D. hansenii* and *Y. lipolytica* results from the expansion of specific gene families (if the 104 multigene families specific to *Y. lipolytica* were ignored, then its net redundancy would drop to

the same figure as *D. hansenii*). In both cases, the total gene number increases without a corresponding increase in the protein repertoire.

Ignoring TGAs (see below), paralogs appear randomly dispersed in the genomes of protoploid yeasts (Supplemental Table 8). This dispersion is compatible with interchromosomal as well as intrachromosomal duplication events that reshuffled the genomes over a long time. Global distributions of sequence identities between paralogous proteins are remarkably similar for the species studied (Fig. 3A). A major mode is observed at ~27% amino acid identity, which corresponds to highly diverged proteins (probably very ancient gene duplications). Higher identity (~35%–75%), representing more recent duplications or stronger functional constraints, is observed with slowly decreasing frequency. Note the relative excess of highly similar paralogs (>90% amino acid identity), representing even more recent gene duplications or stringent functional constraints (slightly more important in *K. lactis* and *K. thermotolerans* than for the three other species).

Tandem gene arrays (TGAs)

The total number of TGAs, with either no (most cases) or a few intervening genes between the successive paralogous copies, ranges from 31 for *A. gossypii* to 51 for *S. kluyveri*, most of which consist of only two or three gene copies (Supplemental Table 9). The few TGAs of four and five genes may correspond to functional adaptations. Some TGAs consist of a gene and a pseudogene, indicating functional inactivation of a gene copy after the duplication. Interestingly, paralogs in TGAs are generally less diverged in sequence than dispersed paralogs (Fig. 3A), in agreement with the idea that they represent more recent gene duplication events, possibly of adaptive type. Accordingly, several TGAs appear as species-specific, as if the tandem duplications occurred independently in the various phylogenetic branches (Supplemental Table 10). They encompass a large variety of functions and, for the most part, protein families of limited sizes (one to three members per species). Other TGAs are conserved within the protoploid yeasts. Many consist of large multigene families (with more than 10 members per species). Using synteny conservation as an indication of common ancestry, a minimum of 18 TGAs were formed prior to species divergence and conserved (despite copy number variation; Supplemental Table 10). For example, genes encoding

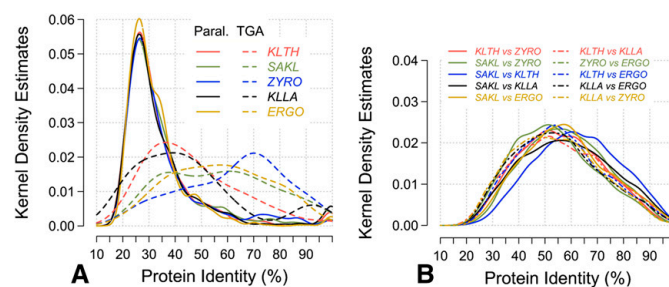


Figure 3. Distribution of amino acid sequence identities between pairs of homologous proteins. Pairs of orthologous (as defined from SONS) or paralogous proteins (defined from protein families) were used to compute the distributions. (A) Paralogs; (B) orthologs. Amino acid identities were calculated from BLAST alignments with low complexity filter. Each distribution was computed from all pairwise alignments between two species for orthologs, and from pairwise alignments within families of two and three members for paralogs. (Solid lines) Dispersed paralogs; (dashed lines) TGAs; species abbreviations as in Figure 2.

the TAF14 subunit of the polymerase II transcriptional machinery, protein kinases involved in mitotic exit network, and B-type cyclins comprise two-copy TGAs conserved in all five protoploid yeasts (the last one is duplicated in *S. cerevisiae* and *C. glabrata*). TGA expansion and conservation among all hemiascomycetous yeasts will be published elsewhere (L. Despons, P. Baret, L. Frangeul, V. Leh-Louis, P. Durrens, and J.L. Souciet, in prep.).

Conservation of dispersed paralogous gene copies

We also examined the conservation of dispersed paralogs to pinpoint events of gene duplication or loss during the evolution of the five protoploid yeasts and to try to evaluate their functional consequences. Many families of two members (204) or three members (52) are conserved among the five species studied. Considering, as a conservative underestimate, only pairs of paralogs for which orthology of both members is clearly demonstrated for all five species by synteny conservation, a total of 114 cases could be retrieved (Supplemental Table 11). They are involved in a variety of functions, as deduced from their *S. cerevisiae* homologs, and have diverse degrees of sequence conservation. In addition to the genes encoding histones H4 and H2B, and mitochondrial, cytoplasmic, or peroxysomal protein isoforms that were expected, we found genes encoding a variety of protein kinases, translational elongation factors, thioredoxins, a nuclear pore component, an ATPase of the AAA family, the 1,3-beta-D-glucan synthase, factors involved in fatty acid metabolism, and specific components of the proteasome and ubiquitination pathway. Some conserved duplicated genes are of unknown function. Among the conserved paralogous proteins having undergone the greatest sequence variation (~20% of amino acid identity) are a component of the SSU processosome containing the U3 snoRNA, a ubiquitin-specific protease, and a protein involved in structural maintenance of chromosomes.

Variation of dispersed paralogous gene copies

At the other extreme, events of gene duplication or loss that occurred during the evolution of protoploid yeasts are illustrated by several species-specific families of paralogs. Given the diversity of the resulting situations, we have examined three simple cases. If one considers (Supplemental Table 12) genes present in one copy in four species but in families of two or more paralogs in the fifth, the largest amplification is represented by five paralogs (including one tandem) in *S. kluyveri* that encode proteins with weak similarity to the Amn1 proteins of *S. cerevisiae* involved in daughter cell separation and chromosome stability. Other series of two or three paralogs are involved in a variety of functions. If one considers (Supplemental Table 13) genes present in families of two or more paralogs in one species but absent from all others, the most spectacular cases are two families of, respectively, 20 and 10 members observed in *Z. rouxii*, the first one related to the COS/DUP family of mostly subtelomeric genes in *S. cerevisiae*. Similarly, there are two families of seven members in *K. lactis*, two families of five members, and one family of four members in *K. thermotolerans*. Among this species-specific category, families of two or three paralogs are also found in all species, but, unfortunately, almost no functional inference is possible owing to lack of homologs. Finally, there are a variety of gene family expansions in some species that are probably related to functional adaptation (Supplemental Table 14). For example, *K. lactis* has 12 copies (compared to zero to four for other species) of a gene similar to the Mch2 protein of *S. cerevisiae* that may be involved in importing monocarboxylic acids. Similarly, *S. kluyveri* has 12 copies (compared to three to six for other species) of a gene encoding proteins weakly similar to cell surface

proteins, and *Z. rouxii* has 10 copies (compared to zero to four for other species) of genes potentially encoding oxydoreductases involved in the formation of chiral alcohols.

Orthologous genes and synteny conservation

Definition of orthologs

The proper identification of orthologous genes is a major challenge (Kuzniar et al. 2008) because accumulation of gene-loss and duplication events tends to blur the recognition of true orthologs among the set of remaining homologs. Reciprocal best hits are often used to assign orthologs among a set of homologs (Rivera et al. 1998), but true orthologs are not necessarily those with the most similar sequences among all homologs (Lynch and Conery 2000). The present set of yeast species offered us a chance to define orthologous genes based on conservation of their genetic location. We started from a series of homologous genes identified from the corresponding protein families (see above), and examined their flanking chromosomal regions for the presence of other homologs (see Methods). In this way, we defined subsets of orthologous genes from each inspected family of homologs. The procedure was reiterated exhaustively until all protein-coding genes of all yeast species had been examined, defining SONS (Subset of Orthologs by Neighborhood and Similarity). In total, we identified 3896 SONS out of 3493 families (80% of families present in more than one species). A total of 27,926 genes (64% of the genes) are assigned to SONS, that is, belong to groups of orthologs confirmed by a shared neighborhood. Most of the remaining genes belong to *Y. lipolytica* and *D. hansenii* as expected from the very low conservation of synteny between those species and other yeasts. The complete list of genes in SONS can be found at <http://www.genolevures.org/>.

Conservation of orthologs

A total of 1609 SONS (41%) are common to all yeasts (i.e., they comprise a series of 1:1:1:1 orthologs confirmed by synteny). For the rest, some genes are missing in some species owing to gene loss during evolution. Using all alignments of orthologous proteins (as defined from cognate SONS), we computed the distributions of sequence identities for all pairwise comparisons of our five yeasts (Fig. 3B). Strikingly, their mean or median values (all distributions are monomodal, as expected for orthologs) underline the large evolutionary distances between protoploid species of *Saccharomycetaceae*. Even species belonging to the same clade, such as *S. kluyveri* and *K. thermotolerans*, show an average amino acid identity as low as 58.2%. Other pairwise comparisons show even lower similarity (48%–53%), consistent with their proposed phylogeny (see Fig. 1). Thus, even a homogeneous phylogenetic group such as the *Saccharomycetaceae* spans a very broad evolutionary range.

Synteny conservation

Using gene orthologs (as defined from SONS), we computed the maximal synteny conservation between the five protoploid yeast genomes as described in Methods. Pairwise comparisons between the five species (Supplemental Fig. 4) show the significant amount of map reshuffling that occurred during the evolution of this group of species, leaving numerous and relatively short synteny blocks scattered throughout these genomes. However, recognizable synteny blocks cover significant proportions of each genome (~80%–90% of all protein-coding genes). The distributions of block size (14–26 genes) and number (180–291) vary only slightly between pairs of species, except for the *S. kluyveri* vs. *K. thermotolerans*

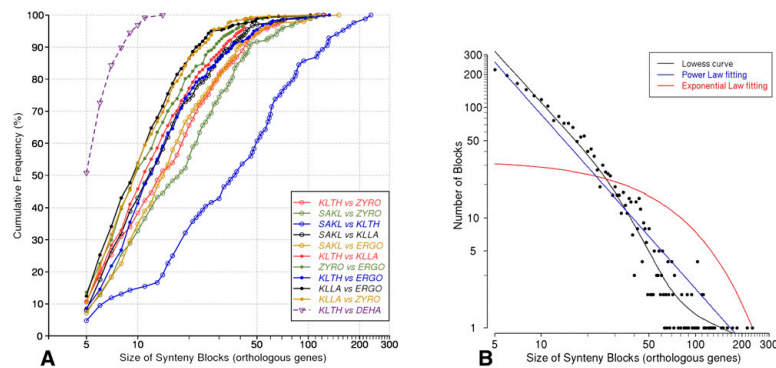


Figure 4. Size of synteny blocks. (A) Distribution of sizes of synteny blocks (log scale abscissae) between all pairs of protoploid *Saccharomycetaceae*. The same distribution for synteny blocks between *K. thermotolerans* and an outgroup species (*D. hansenii*) is shown for comparison. Species abbreviations as in Figure 2. (B) Frequency distribution of synteny blocks (log scale ordinate) between all pairs of protoploid *Saccharomycetaceae* according to size (log scale abscissa).

comparison that shows larger (mean size 59 genes) and fewer (84) synteny blocks. An example of synteny block conservation is illustrated by Supplemental Figure 5. Note that, despite extensive map reshuffling, the five protoploid species clearly form a homogeneous group: When a more distantly related yeast such as *D. hansenii* is compared to any of them, the size of synteny blocks and their coverage of the genome diminish considerably (Supplemental Fig. 6). Distributions of size of synteny blocks in all pairwise species comparisons (Fig. 4) are consistent with the phylogenetic relationships between species. Interestingly, the frequency distribution of synteny blocks according to their size fits a power law rather than the expected exponential law implied by the random breakage model of Nadeau and Taylor (1984).

Our definition of synteny blocks allows for the presence of “intervening” genes in them (see Methods). The number of such genes is roughly proportional to block sizes (Supplemental Fig. 6). Figures are similar for *Z. rouxii*, *K. lactis*, and *A. gossypii* (approximately seven intervening genes for 100 orthologs in syntenic blocks), but are significantly higher for *S. kluyveri* (approximately 12). An intermediate figure is observed for *K. thermotolerans* (approximately nine).

Relationship between sequence divergence and synteny conservation

We examined the relationship between the rates of sequence divergence and chromosomal rearrangements during the evolution of hemiascomycetes (Fig. 5). There is a nonlinear relationship between average amino acid identity of orthologous proteins and the proportion of genes remaining in synteny. The five protoploid species are grouped in a sector of the curve in which sequence identity decreases rapidly while synteny conservation decreases slowly. Pairwise comparisons of the same species with those of the outgroup (*D. hansenii* or *Y. lipolytica*) extrapolate well with results of the most distant species among protoploids (*Z. rouxii* vs. *K. lactis* or *A. gossypii*). The nonlinear form of this relationship indicates that significant sequence divergence occurs before chromosomal maps become extensively rearranged. At larger evolutionary distances, chromosome reshuffling catches up to protein-sequence divergence, which becomes limited by saturation and functional constraints (see Discussion).

Genome rearrangements from inferred ancestors

The slow pace of genome rearrangements enabled us to infer ancestral events from contemporary genomes based on conservation of synteny blocks. Such an inference is possible because, in protoploid yeasts, the probability of independent formation of such blocks in different lineages is likely to remain negligible compared to their inheritance from a common ancestor. Each genome was factored into a sequence of ordered synteny blocks common to all genomes, and super-blocks were computed using the method developed by G Jean, DJ Sherman, and M Nikolski (in prep.). From this, we computed median genomes for the five protoploid genomes, as well as rearrangement trees containing intermediate

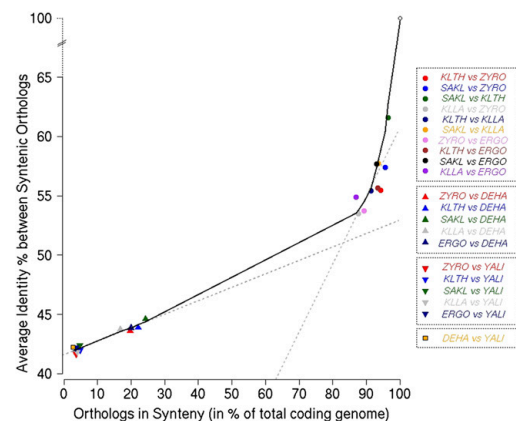


Figure 5. Relationship between protein sequence divergence and synteny conservation. Average sequence identities between orthologous proteins (ordinate) in all pairwise species comparisons were calculated from Figure 3. The percent of orthologs remaining in synteny (abscissa) for the same pairs of yeast species is the ratio of the total number of orthologs within syntenic blocks over the total number of orthologs between the two species considered.

ancestral candidates for different branches (Goëffon et al. 2008). Figure 6 shows pairwise rearrangement distances and a rearrangement tree. The rearrangement tree separates *Z. rouxii* from the other four species, in agreement with the phylogenetic tree based on sequence alignments (see Fig. 1). Again, in agreement with phylogeny, *S. kluyveri* and *K. thermotolerans* are the closest in terms of rearrangements. However, in the rearrangement tree, *A. gossypii* and *K. lactis* are as diverged from each other as *Z. rouxii* is from the ancestor of all five species. This underlines the highly dynamic nature of yeast genomes, even within a related group like the *Saccharomycetaceae*. (The difference in rearrangement distance between the binary rearrangement tree and a simple star topology rooted at the median is only 10 events [3.5%].)

Discussion

With the complete genome sequencing and analysis of three novel yeast species, *Z. rouxii*, *K. thermotolerans*, and *S. kluyveri*, we have been able to perform the first comparative genomic study involving four distinct clades (and five species) of *Saccharomycetaceae*, that all separated from *S. cerevisiae* before its ancestral genome duplication. In two previous studies, yeasts of these clades, namely, *A. gossypii* (Dietrich et al. 2004) and *K. waltii* (Kellis et al. 2004), were individually compared to *S. cerevisiae* but not between themselves. In another study (Dujon et al. 2004), *K. lactis* was involved in multispecies comparative genomics with four other hemiascomycetes, but none was a protoploid *Saccharomycetaceae*. The present study was facilitated by the development of new methods that may be of general interest for other comparative genomic studies between species having sufficient conservation of

synteny (see Methods). Among them is the method we used to define gene orthology (ML Seret and PV Baret, in prep.). Compared to YGOB, which implies manual curation of data (Byrne and Wolfe 2005), and to SYNERGY, which requires a priori definition of parameters (Wapinski et al. 2007), our method (IONS) simply subdivides precomputed sets of homologs (based on sequence similarity) using gene neighborhood evidence in a reiterative process with gradually decreasing neighborhood sizes until series of homologs with only one gene per species are obtained.

Genomes of the five species we examined reveal common signatures that confirm the monophyletic origin of *Saccharomycetaceae* and distinguish them from representatives of other subdivisions of hemiascomycetes, used here as outgroups for comparison. Among those signatures are short centromeres (Meraldi et al. 2006), triplication of mating cassettes (except for *S. kluyveri*) (Butler et al. 2004), a very limited number of spliceosomal introns (Bon et al. 2003), usage of the universal genetic code (Miranda et al. 2006), and a single rDNA locus containing the genes for the 5S RNA molecule (Acker et al. 2008). Additional discriminatory signatures are also found in their mitochondrial genomes: The genes encoding subunits of complex I of the respiratory chain are missing, while a gene for a variable ribosomal subunit and another one for an abnormal tRNA-Thr (UAG) altering the mitochondrial genetic code are present (<http://www.ncbi.nlm.nih.gov/genomes/OrganelleResource.cgi?opt=organelle&taxid=4751>).

The five protoploid genomes show only limited variation in the total number of protein-coding genes (<7% relative to the mean) and almost no variation for some noncoding RNA genes. But their chromosome number varies from six to eight, the number of tRNA genes varies from 163 to 272, and the average genome composition from ~39% to 52% GC, indicating important lineage-specific evolution that may be correlated with the different lifestyles and properties of these five species. Such diversity is not surprising given the large evolutionary distances separating the five species, as revealed by the significant sequence divergence among orthologous proteins. This was not known before. Even the two species classified within the same *Lachancea* clade, *S. kluyveri* and *K. thermotolerans*, show average amino acid identity of only ~58%. This is much less than between the two most distant species of the *Saccharomyces sensu stricto* clade (Bon et al. 2000a; Cliften et al. 2003; Kellis et al. 2003), which has nevertheless been more extensively studied, illustrating again previous emphasis on yeasts involved in biotechnological fermentation.

Evolutionary distances among protoploid yeasts are such that our sequence-based phylogeny (Fig. 1), which is in agreement with that of Fitzpatrick et al. (2006), is not entirely congruent with the phylogeny of *Saccharomycetaceae* published by Kurtzman (2003), where clade 12 (*Eremothecium*) is separated from clade 11 (*Kluyveromyces*). In both cases, however, *Z. rouxii* branches separately from the four other yeasts and shares a common origin with the duplicated yeasts of the *Saccharomyces* and *Nakaseomyces* clades. Results presented here definitely demonstrate that *Z. rouxii* separated from them before the whole-genome duplication. Given the tree topology of Kurtzman (2003) and the fact that the *Vanderwaltozyma* clade was recently demonstrated to have emerged after the duplication (Scannell et al. 2007b), *Zygosaccharomyces* represents the last clade having diverged from its relatives prior to the whole-genome duplication. Of all protoploid yeasts studied here, *Z. rouxii* is, therefore, the most closely related to the putative ancestral genome of *S. cerevisiae*.

Despite their long evolutionary branches, the five protoploid species we analyzed show a high degree of synteny conservation.

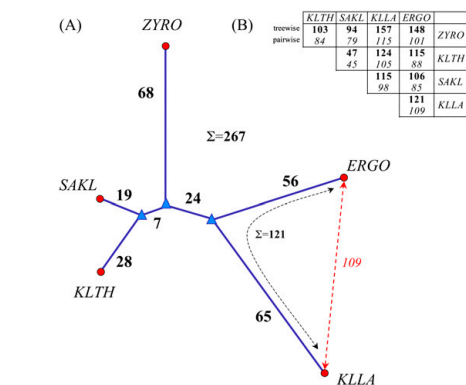


Figure 6. Genome rearrangements. (A) Minimum rearrangement tree computed using the FAUCILS stochastic local search method (Goëffon et al. 2008); (●) contemporary genomes; (▲) inferred ancestral genomes; bold figures show the number of rearrangements in this minimum tree. (B) Contrast between rearrangement distances summing the branches in the tree (in bold, upper row), and pairwise distances (in italics, lower row) computed using the GRIMM method (Tesler 2002). Note that the requirement of going through a common ancestor usually increases tree-wise distances compared to pairwise distances owing to the triangle inequality, but decreases the sum of distances for the entire tree when all genomes are taken into account. This is illustrated in A for the pair ERGO and KLLA, $\Sigma = 121 > 109$, while the overall sum for the tree is 267. The rearrangement tree independently corroborates the phylogenetic tree calculated using sequence similarity (see Discussion).

Even if chromosome rearrangement breakages are relatively numerous (a few hundred, except between *S. kluyveri* and *K. thermotolerans*), they leave relatively long synteny blocks of one or two dozen genes on average, and large proportions of genomes (>80%) are present in still recognizable synteny blocks in all pairwise comparisons. Such a range of synteny conservation was not available for yeasts before. The *Saccharomyces sensu stricto* yeasts share almost completely conserved synteny with few breakpoints (Fischer et al. 2000, 2001), and other yeasts are too distantly separated to retain much synteny (Fischer et al. 2006). The nonlinear relationship between sequence divergence and synteny conservation is compatible with the idea that genetic maps are more robust than DNA sequences over evolutionary periods corresponding to entire families or even orders of the Linnean taxonomical hierarchy. The effects of genome rearrangements become significant over longer evolutionary times when sequence changes saturate. Similar ideas based on much more limited sets of data were already discussed (Langkjaer et al. 2000; Llorente et al. 2000b; Malpertuy et al. 2000b). The frequency distribution of synteny blocks according to their size is similar to results obtained from the 12 *Drosophila* genomes and other insects (Zdobnov and Bork 2007), suggesting that chromosome rearrangements are not random in either of these two groups of eukaryotes despite very distinct genome organizations. Accordingly, genome rearrangement trees built from inferred ancestors are not entirely coincident with phylogenetic trees built on sequence conservation of gene products. Differences in the rate of genome rearrangements between yeast lineages have been reported (Fischer et al. 2006). Note that conserved synteny blocks often contain short internal gene deletions and/or single gene insertions ("intervening" genes). Many "intervening" genes have homologs at ectopic locations in either the same or other genomes and may have arisen by single gene transfer (possibly by retroposition) or by dispersed short segmental duplications similar to the few that were recognized. A few "intervening" genes correspond to horizontally transferred genes, as previously illustrated by the presence of a bacterial transposase-type of sequence in *A. gossypii* (Hall et al. 2005). Gene transfer from bacteria to yeasts is also documented in *S. kluyveri* by the presence of six copies of a bacterial IS element (family IS607) on chromosomes B, G, and H (Rolland et al. 2009).

The genomes of protoploid yeasts contain remarkably few segmental duplications (except in subtelomeric locations), in contrast to genomes of higher eukaryotes, where they are a major source of genome rearrangements (Johnson et al. 2001; Samonte and Eichler 2002), and to the high frequency of their spontaneous formation in *S. cerevisiae* (Payen et al. 2008). The limited role of segmental duplications in the evolution of yeast genomes may be related to the important instability of duplicated copies of identical sequences. Compared to their paucity in recent duplications of nearly identical sequences, genomes of protoploid yeasts contain a surprisingly high number of paralogous genes: about a third of their gene content. Thus, we can say that there is no such thing as "nonduplicated" yeasts. Families of paralogs issue from ancient duplication events that leave diverged gene copies at dispersed genomic locations and from TGAs, in which gene copies are globally less diverged. Although some of the dispersed duplications appear ancestral and conserved in all protoploid species (some are also conserved in outgroup yeasts, suggesting they may be ancestral to all hemiascomycetes), we recognized a significant number of lineage-specific events of gene duplication and loss, some of which resulted in significant gene family expansions that may be correlated with specific biological properties of each yeast. But the

remarkable point is that, despite such dynamics, the diverse protoploid species have similar levels of global genome redundancy (the figure for *A. gossypii* is slightly smaller, in agreement with a reductive and adaptive evolution [Dietrich et al. 2004]). The net redundancy, ~1.25 on average, reflects an equilibrium between gene duplication and loss over long evolutionary periods, when this equilibrium is not affected by accidental events such as whole-genome duplication or adaptive events such as gene family expansions.

The genome sequences of the *Saccharomycetaceae* we described, which offer a more balanced representation of the different clades, provide the framework to understand the origin of the common properties and individual variations of these yeasts. Their core genetic repertoire (core proteome) consists of approximately 3300 protein families, within a pan-proteome of approximately 5000 families for all *Saccharomycetaceae*. The fact that the pan repertoire dramatically increases when outgroup species are considered underlines the major evolutionary gaps between the different subdivisions of hemiascomycetes and the need for additional genomic studies of this interesting fungal group.

Methods

Genome sequencing

Complete genome sequences of *K. thermotolerans* (CBS6340) and *Z. rouxii* (CBS732; determined at Genoscope) and of *S. kluyveri* (CBS3082; determined at the Washington University Genome Sequencing Center) were automatically assembled from shotgun Sanger sequencing reads using ARACHNE (Jaffe et al. 2003) and finished by dedicated sequencing for gaps and low-quality regions.

Genome annotation

CDS

Gene models for protein-coding genes were constructed using GeneMark (Borodovsky and McIninch 1993) trained with a conservative set of coding and noncoding sequences for each of the four genomes: *Z. rouxii*, *K. thermotolerans*, *S. kluyveri*, and *K. lactis*. All open reading frames (ORFs) greater than 80 codons were examined. Overlapping conflicts were solved by best GeneMark prediction and/or existence of BLASTP alignments with the UniProtKB database. Gene models were manually examined and simultaneously annotated for the four species after automated identification of homologs (DJ Sherman, T Martin, and P Durrens, in prep.).

Spliceosomal introns

Consensus splice sites and branch point sequences were defined for each yeast species from introns of ribosomal protein-coding genes (Bon et al. 2003). All combinations of the three motifs separated by appropriate distances (also defined for each yeast species) were examined for the possible formation of CDS greater than 80 codons after splicing. Note that introns in 5'-UTRs were, therefore, not systematically predicted. Overlapping intron predictions were manually curated.

Genes for noncoding RNAs

Genes for tRNAs were identified as described in Marck et al. (2006). Ribosomal DNA repeats were identified by comparison to

other yeast genomes (Chindamporn et al. 1993; Rustchenko and Sherman 1994) and mapped to chromosomes using previous karyotype data (Sor and Fukuhara 1989; Neuvéglise et al. 2000; Sychrova et al. 2000). Genes for C/D snoRNAs and snRNAs were annotated by a combination of sequence similarity search with *S. cerevisiae*, secondary structure, and synteny conservation. H/ACA snoRNAs were not systematically identified.

Mobile genetic elements

Transposable elements were detected by TBLASTN on chromosome sequences using known fungal transposable elements as queries, and manually curated. Identified elements were then used to detect partial or degenerate elements and solo LTRs.

Protein families

Four complementary distance matrices were computed between the predicted translation products for the 47,874 protein-coding genes of the seven yeasts annotated by Génolevures (*C. glabrata*, *Z. rouxii*, *K. thermotolerans*, *S. kluyveri*, *K. lactis*, *D. hansenii*, and *Y. lipolytica*) (<http://www.genolevures.org/>), plus *S. cerevisiae* (<http://www.yeastgenome.org/>) and *A. gossypii* (<http://agd.vital-it.ch/>), combining BLAST and Smith-Waterman alignments with and without filtering for homeomorphy (Wu et al. 2004). Symmetric matrices derived from amino acid identities were constructed and submitted to MCL clustering (Enright et al. 2002) with a range of inflation parameters. These competing partitions were reconciled using the consensus method of Nikolski and Sherman (2007) and manually curated using literature search and systematic comparisons with COG (Tatusov et al. 2003) and PIRSF (Wu et al. 2004) classifications.

Orthologous genes

SONS were defined by the combination of two dimensions: sequence similarity of gene products (defined from families; see above), and conservation of gene neighborhood along chromosomes (Supplemental Fig. 7). Two genes of different yeast species whose translation products belong to the same family (homologs by similarity) will be members of the same SONS if they share at least one pair of neighbors that are also homologous to each other by similarity. The process is reiterated for all possible hetero-specific pairwise comparisons of homologs. At an initial step, neighboring is examined for 15 genes on each side of the two homologs considered, giving rise to $(2 \times 14)^2/2$ pairwise comparisons. Homologs that do not share a pair of homologous neighbors are separated in two distinct SONS. The process is reiterated (reducing the neighborhood size in order to decrease the probability of spurious connections) until each SONS contains ≤ 1 gene for each yeast species. Homologs in such a case are considered as orthologs confirmed by synteny. Details of the method and comparisons with other methods to infer orthology will be published separately (ML Seret and PV Baret, in prep.).

Syntenic block construction

Syntenic blocks between two yeast genomes were constructed from the physical adjacency of orthologous genes (defined by SONS, above) along chromosomes (deduced from sequence-based map), using two parameters: *A*, the minimum number of ortholog pairs forming anchor points; and *I*, the maximum number of non-

orthologous genes ("intervening") between two successive anchor points. (*A* and *I* were set, respectively, to 5 and 10, following Fischer et al. [2006].) Note that TGAs (see text) were considered as equivalent to a single gene.

List of participants and affiliations

Overall coordination

Jean-Luc Souciet,^{2,18} Bernard Dujon,³ Claude Gaillardin,⁴ Mark Johnston,^{5,17} Philippe V. Baret,⁶ Paul Cliften,⁷ David J. Sherman,⁸ Jean Weissenbach,⁹ and Eric Westhof¹⁰

Genome sequence and assembly

Patrick Wincker,⁹ Claire Jubin,⁹ Julie Poulain,⁹ Valérie Barbe,⁹ Béatrice Ségurens,⁹ François Artiguenave,⁹ Véronique Anthouard,⁹ Benoît Vacherie,⁹ Marie-Eve Val,⁹ Robert S. Fulton,¹¹ Patrick Minx,¹¹ and Richard Wilson¹¹

Analysis and annotation

Pascal Durrens,⁸ Géraldine Jean,⁸ Christian Marck,¹² Tiphaine Martin,⁸ Macha Nikolski,⁸ Thomas Rolland,³ Marie-Line Seret,⁶ Serge Casarégola,⁴ Laurence Despons,² Cécile Fairhead,³ Gilles Fischer,³ Ingrid Lafontaine,³ Véronique Leh,² Marc Lemaire,¹³ Jacky de Montigny,² Cécile Neuvéglise,⁴ Agnès Thierry,³ Isabelle Blanc-Lenle,⁴ Claudine Bleykasten,² Julie Diffels,⁶ Emilie Fritsch,² Lionel Frangeul,¹⁴ Adrien Goëffon,⁸ Nicolas Jauniaux,² Rym Kachouri-Lafond,¹⁰ Célia Payen,³ Serge Potier,² Lenka Pribylova,^{2,15} Christophe Ozanne,⁴ Guy-Franck Richard,³ Christine Sacerdot,³ Marie-Laure Straub,² and Emmanuel Talla¹⁶

Acknowledgments

We thank E. Beyne and H. Soueidan for help in programming and data management, Y. Boulard for help in tRNA gene search, M. Santos for sharing unpublished results on tRNA genes, A.N.

²Université de Strasbourg, CNRS UMR7156, F-67000 Strasbourg, France.

³Institut Pasteur, CNRS URA2171, University Pierre et Marie Curie, Paris 6 UFR927, F-75724, Paris-CEDEX15, France.

⁴AgroParisTech, CNRS UMR2585, INRA UMR1238, Microbiologie et Génétique Moléculaire, F-78850 Thiverval-Grignon, France.

⁵Washington University School of Medicine, Department of Genetics, St. Louis, Missouri 63110, USA.

⁶Université Catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium.

⁷Department of Biology, Utah State University, Logan, Utah 84322, USA.

⁸University of Bordeaux 1, CNRS UMR5800, LaBRI INRIA Bordeaux Sud-Ouest (MAGNOME) F-33405 Talence, France.

⁹CEA, DSV, IG, Génomique; CNRS UMR 8030; Université d'Evry Val d' Essonne, F-91057 Evry, France.

¹⁰Université Louis Pasteur, Architecture et Réactivité de l'ARN, Institut de Biologie moléculaire et cellulaire du CNRS, F-67084 Strasbourg, France.

¹¹Washington University School of Medicine, Department of Genetics and Genome Sequencing Center, St. Louis, Missouri 63108, USA.

¹²Institut de Biologie et de Technologies de Saclay (IBiTec-S), CEA, F-91191 Gif-sur-Yvette CEDEX, France.

¹³Université de Lyon 1, CNRS, UMR5240 Microbiologie, Adaptation et Pathogénie, INSA de Lyon, Villeurbanne, F-69621 Villeurbanne, France.

¹⁴Institut Pasteur, Platform Intégration et Analyse génomique, F-75015, Paris, France.

¹⁵Institut of Physiology AS CR, Department of Membrane Transport, Videnska 1083, 14220 Prague 4, Czech Republic.

¹⁶Université de la Méditerranée, Laboratoire de Chimie Bactérienne, CNRS-UPR9043, F-13402 Marseille CEDEX 20, France.

¹⁷Present address: Department of Biochemistry and Molecular Genetics, University of Colorado—Denver, Anschutz Medical Campus, Mail Stop 8101, P.O. Box 6511, Aurora, Colorado 80045, USA.

Nikolskaya for expertise in curating protein families, and M. Aigle and F. Tekaia for useful discussions. This work was supported in part by funding from the Consortium National de Recherche en Génomique (CNRG) to Génoscope, from CNRS (GDR 2354; Génolevures), ANR (ANR-05-BLAN-0331; GENARISE), ARC (grant 3738), Région Aquitaine (Pôle Recherche en Informatique; 20051306001AB), and Bureau des ressources génétiques (grant 347; Diversité fongique), and in part by grants awarded to R.W. by the National Human Genome Research Institute of the U.S. National Institutes of Health, and by funds provided to M.J. by the James S. McDonnell Foundation. B.D. is a member of Institut Universitaire de France.

References

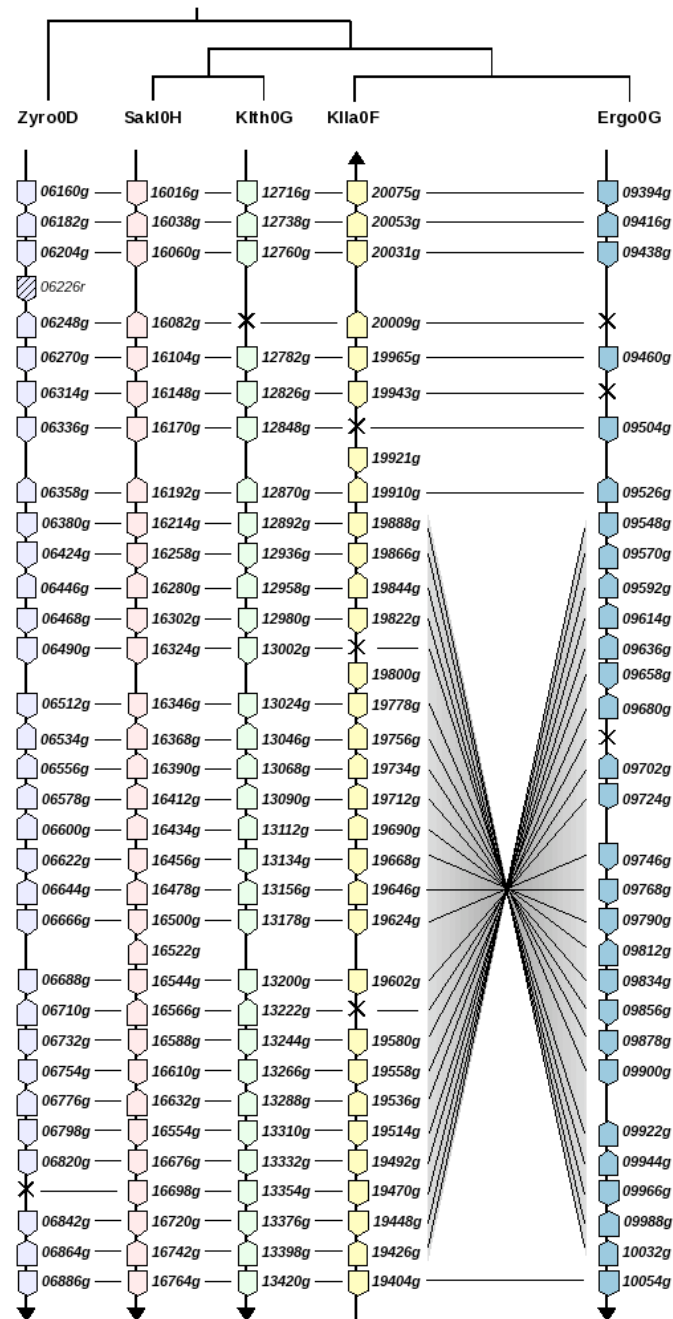
- Acker J, Ozanne C, Kachouri-Lafond R, Gaillardin C, Neuvéglise C, Marck C. 2008. Diclitrionic tRNA-5S genes in *Yarrowia lipolytica*: An alternative TFIIA-independent way for expression of 5S rRNA genes. *Nucleic Acids Res* **18**: 5832–5844.
- Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796–815.
- Araki H, Jeampipatkul A, Tatsumi H, Sakurai T, Ushio K, Muta T, Oshima Y. 1985. Molecular and functional organization of yeast plasmid pSR1. *J Mol Biol* **182**: 191–203.
- Aslett M, Wood V. 2006. Gene ontology annotation status of the fission yeast genome: Preliminary coverage approaches 100%. *Yeast* **23**: 913–919.
- Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Séguens B, Daubin V, Anthouard V, Aïach N, et al. 2006. Global trends of whole-genome duplication revealed by the ciliate *Paramecium tetraurelia*. *Nature* **444**: 171–178.
- Barnett JA. 2007. A history of research on yeasts 10: Foundations of yeast genetics. *Yeast* **24**: 799–845.
- Beck H, Dobritzsch D, Piskur J. 2008. *Saccharomyces kluyveri* as a model organism to study pyrimidine degradation. *FEMS Yeast Res* **8**: 1209–1213.
- Blank LM, Lehmbeck F, Sauer U. 2005. Metabolic flux and network analysis in fourteen hemiascomycetous yeasts. *FEMS Yeast Res* **5**: 545–558.
- Boekhout T. 2005. Biodiversity: Gut feeling for yeasts. *Nature* **434**: 449–451.
- Bon E, Neuvéglise C, Casarégola S, Artiguenave F, Wincker P, Aigle M, Durrens P. 2000a. Genomic exploration of the hemiascomycetous yeasts: 5. *Saccharomyces bayanus* var. *uvarum*. *FEBS Lett* **487**: 37–41.
- Bon E, Neuvéglise C, Lépingle A, Wincker P, Artiguenave F, Gaillardin C, Casarégola S. 2000b. Genomic exploration of the hemiascomycetous yeasts: 6. *Saccharomyces exiguus*. *FEBS Lett* **487**: 42–46.
- Bon E, Casarégola S, Blandin G, Llorente B, Neuvéglise C, Munsterkotter M, Guldener U, Mewes HW, Van Helden J, Dujon B, et al. 2003. Molecular evolution of eukaryotic genomes: Hemiascomycetous yeast spliceosomal introns. *Nucleic Acids Res* **31**: 1121–1135.
- Bormeman AR, Forgan AH, Pretorius IS, Chambers PJ. 2008. Comparative genome analysis of a *Saccharomyces cerevisiae* wine strain. *FEMS Yeast Res* **8**: 1185–1195.
- Borodovsky M, McIninch J. 1993. Recognition of genes in DNA sequence with ambiguities. *Biosystems* **30**: 161–171.
- Butler G, Kenny C, Fagan A, Kurischko C, Gaillardin C, Wolfe KH. 2004. Evolution of the MAT locus and its HO endonuclease in yeast species. *Proc Natl Acad Sci* **101**: 1632–1637.
- Byrne KP, Wolfe KH. 2005. The Yeast Gene Order Browser: Combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res* **15**: 1456–1461.
- Byrnes JK, Morris GP, Li WH. 2006. Reorganization of adjacent gene relationships in yeast genomes by whole-genome duplication and gene deletion. *Mol Biol Evol* **23**: 1136–1143.
- Casarégola S, Lépingle A, Bon E, Neuvéglise C, Huu-Vang N, Artiguenave F, Wincker P, Gaillardin C. 2000. Genomic exploration of the Hemiascomycetous Yeasts: 7. *Saccharomyces servazzii*. *FEBS Lett* **487**: 47–51.
- Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**: 540–552.
- Chappell AS, Lundblad V. 2004. Structural elements required for association of the *Saccharomyces cerevisiae* telomerase RNA with the Est2 reverse transcriptase. *Mol Cell Biol* **24**: 7720–7736.
- Chindamporn A, Iwaguchi S, Nakagawa Y, Homma M, Tanaka K. 1993. Clonal size-variation of rDNA cluster region on chromosome XII of *Saccharomyces cerevisiae*. *J Gen Microbiol* **139**: 1409–1413.
- Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterson R, Cohen BA, Johnston M. 2003. Finding functional features in *Saccharomyces cerevisiae* by phylogenetic footprinting. *Science* **301**: 71–76.
- Cliften PF, Fulton RS, Wilson RK, Johnston M. 2006. After the duplication: Gene loss and adaptation in *Saccharomyces* genomes. *Genetics* **172**: 863–872.
- Conant GC, Wolfe KH. 2008. Probabilistic cross-species inference of orthologous genomic regions created by whole-genome duplication in yeast. *Genetics* **179**: 1681–1692.
- Demogines A, Smith E, Kruglyak L, Alani E. 2008. Identification and dissection of a complex DNA repair sensitivity phenotype in Baker's yeast. *PLoS Genet* **4**: e1000123. doi: 10.1371/journal.pgen.1000123.
- de Montigny J, Straub ML, Potier S, Tekaia F, Dujon B, Wincker P, Artiguenave F, Souciet JL. 2000. Genomic exploration of the hemiascomycetous yeasts: 8. *Zygosaccharomyces rouxii*. *FEBS Lett* **487**: 52–55.
- Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, Steiner S, Mohr C, Puhlmann R, Luedi P, Choi S, et al. 2004. The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science* **304**: 304–307.
- Dujon B. 2005. Hemiascomycetous yeasts at the forefront of comparative genomics. *Curr Opin Genet Dev* **6**: 614–620.
- Dujon B. 2006. Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. *Trends Genet* **22**: 375–387.
- Dujon B, Sherman D, Fischer G, Durrens P, Casarégola S, Lafontaine I, De Montigny J, Marck C, Neuvéglise C, Talla E, et al. 2004. Genome evolution in yeasts. *Nature* **430**: 35–44.
- Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**: 1575–1584.
- Fairhead C, Dujon B. 2006. Structure of *Kluyveromyces lactis* subtelomeres: Duplications and gene content. *FEMS Yeast Res* **6**: 428–441.
- Fischer G, James SA, Roberts IN, Oliver SG, Louis EJ. 2000. Chromosomal evolution in *Saccharomyces*. *Nature* **405**: 415–454.
- Fischer G, Neuvéglise C, Durrens P, Gaillardin C, Dujon B. 2001. Evolution of gene order in the genomes of two related yeasts species. *Genome Res* **11**: 2009–2019.
- Fischer G, Rocha EP, Brunet F, Vergassola M, Dujon B. 2006. Highly variable rates of genome rearrangements between hemiascomycetous yeast lineages. *PLoS Genet* **2**: e32. doi: 10.1371/journal.pgen.0020032.
- Fitzpatrick DA, Logue ME, Stajich JE, Butler G. 2006. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evol Biol* **6**: 99. doi: 10.1186/1471-2148-6-99.
- Goëffon A, Nikolski M, Sherman DJ. 2008. An efficient probabilistic population-based descent for the median genome problem. In *Proceedings of the 2008 GECCO Conference Companion on Genetic and Evolutionary Computation*, pp. 315–322. ACM, New York.
- Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, et al. 1996. Life with 6000 genes. *Science* **274**: 546–567.
- Goodwin TJ, Ormandy JE, Poulter RT. 2001. L1-like non-LTR retrotransposons in the yeast *Candida albicans*. *Curr Genet* **39**: 83–91.
- Goodwin TJ, Busby JN, Poulter RT. 2007. A yeast model for target-primed (non-LTR) retrotransposition. *BMC Genomics* **8**: 263. doi: 10.1186/1471-2164-8-263.
- Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**: 696–704.
- Hall C, Brachat S, Dietrich FS. 2005. Contribution of horizontal transfer to the evolution of *Saccharomyces cerevisiae*. *Eukaryot Cell* **4**: 1102–1115.
- Hofmann G, McIntyre M, Nielsen J. 2003. Fungal genomics beyond *Saccharomyces cerevisiae*. *Curr Opin Biotechnol* **14**: 226–231.
- Jaffe DB, Butler J, Gnerre S, Mauceli E, Lindblad-Toh K, Mesirov JP, Zody MC, Lander ES. 2003. Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res* **13**: 91–96.
- Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, et al. 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* **431**: 946–957.
- Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**: 463–476.
- Jansen M, Veurink JH, Euverink GJ, Dijkhuizen L. 2003. Growth of the salt-tolerant yeast *Zygosaccharomyces rouxii* in microtiter plates: Effects

- of NaCl, pH and temperature on growth and fusel alcohol production from branched-chain amino acids. *FEMS Yeast Res* **3**: 313–318.
- Johnson ME, Viggiano L, Bailey JA, Abdul-Rauf M, Goodwin G, Rocchi M, Eichler EE. 2001. Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413**: 514–519.
- Jones T, Federspiel NA, Chibana H, Dungan J, Kalman S, Magee BB, Newport G, Thorstenson YR, Agabian N, Magee PT, et al. 2004. The diploid genome sequence of *Candida albicans*. *Proc Natl Acad Sci* **101**: 7329–7334.
- Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* **33**: 511–518.
- Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**: 241–254.
- Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428**: 617–624.
- Kurtzman CP. 2003. Phylogenetic circumscription of *Saccharomyces*, *Kluyveromyces* and other members of the *Saccharomycetaceae*, and the proposal of the new genera *Lachancea*, *Nakaseomyces*, *Naumovia*, *Vanderwaltozyma* and *Zygorotulasporea*. *FEMS Yeast Res* **4**: 233–245.
- Kurtzman CP, Fell JW. 2006. Yeast systematics and phylogeny—implications of molecular identification methods for studies in ecology. In *The yeast handbook*, pp. 11–30. Springer-Verlag, New York.
- Kurtzman C, Piskur J. 2006. Taxonomy and phylogenetic diversity among the yeasts. In *The yeast handbook*, pp. 29–46. Springer-Verlag, New York.
- Kurtzman CP, Robnett CJ. 2003. Phylogenetic relationships among yeasts of the “*Saccharomyces* complex” determined from multigene sequence analyses. *FEMS Yeast Res* **3**: 417–432.
- Kuzniar A, van Ham RC, Pongor S, Leunissen JA. 2008. The quest for orthologs: Finding the corresponding gene across genomes. *Trends Genet* **24**: 539–551.
- Langkjaer RB, Nielsen ML, Dagaard PR, Li W, Piskur J. 2000. Yeast chromosomes have been significantly reshaped during their evolutionary history. *J Mol Biol* **304**: 271–288.
- Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts I, Burt A, Koufopanou V, et al. 2009. Population genomics of domestic and wild yeasts. *Nature* **458**: 337–341.
- Llorente B, Malpertuy A, Blandin G, Artiguenave F, Wincker P, Dujon B. 2000a. Genomic exploration of the hemiascomycetous yeasts: 12. *Kluyveromyces marxianus* var. *marxianus*. *FEBS Lett* **487**: 71–75.
- Llorente B, Malpertuy A, Neuveglise C, de Montigny J, Aigle M, Artiguenave F, Blandin G, Bolotin-Fukuhara M, Bon E, Brottier P, et al. 2000b. Genomic exploration of the hemiascomycetous yeasts: 18. Comparative analysis of chromosome maps and synteny with *Saccharomyces cerevisiae*. *FEBS Lett* **487**: 101–112.
- Loftus BJ, Fung E, Roncaglia P, Rowley D, Amedeo P, Bruno D, Vamathevan J, Miranda M, Anderson IJ, Fraser JA, et al. 2005. The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans*. *Science* **307**: 1321–1324.
- Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151–1155.
- Magee BB, Sanchez MD, Saunders D, Harris D, Berriman M, Magee PP. 2008. Extensive chromosome rearrangements distinguish the karyotype of the hypovirulent species *Candida dubliniensis* from the virulent *Candida albicans*. *Fungal Genet Biol* **45**: 338–350.
- Malpertuy A, Llorente B, Blandin G, Artiguenave F, Wincker P, Dujon B. 2000a. Genomic exploration of the hemiascomycetous yeasts: 10. *Kluyveromyces thermotolerans*. *FEBS Lett* **487**: 61–65.
- Malpertuy A, Tekai F, Casarégola S, Aigle M, Artiguenave F, Blandin G, Bolotin-Fukuhara M, Bon E, Brottier P, de Montigny J, et al. 2000b. Genomic exploration of the hemiascomycetous yeasts: 19. Ascomycete-specific genes. *FEBS Lett* **487**: 113–121.
- Marck C, Kachouri-Lafond R, Lafontaine I, Westhof E, Dujon B, Grosjean H. 2006. The RNA polymerase III-dependent family of genes in hemiascomycetes: Comparative RNomics, decoding strategies, transcription and evolutionary implications. *Nucleic Acids Res* **34**: 1816–1835.
- McEachern MJ, Blackburn EH. 1995. Runaway telomere elongation caused by telomerase RNA gene mutations. *Nature* **376**: 403–409.
- Meraldi P, McAlain AD, Rheinbay E, Sorger PK. 2006. Phylogenetic and structural analysis of centromeric DNA and kinetochore proteins. *Genome Biol* **7**: R23. doi: 10.1186/gb-2006-7-3-r23.
- Miranda I, Silva R, Santos MA. 2006. Evolution of the genetic code in yeasts. *Yeast* **23**: 203–213.
- Møller K, Bro C, Piskur J, Nielsen J, Olsson L. 2002. Steady-state and transient-state analyses of aerobic fermentation in *Saccharomyces kluyveri*. *FEMS Yeast Res* **2**: 233–244.
- Møller K, Sharif MZ, Olsson L. 2004. Production of fungal alpha-amylase by *Saccharomyces kluyveri* in glucose-limited cultivations. *J Biotechnol* **111**: 311–318.
- Nadeau JH, Taylor BA. 1984. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc Natl Acad Sci* **81**: 814–818.
- Neuveglise C, Bon E, Lépingle A, Wincker P, Artiguenave F, Gaillardin C, Casarégola S. 2000. Genomic exploration of the hemiascomycetous yeasts: 9. *Saccharomyces kluyveri*. *FEBS Lett* **487**: 56–60.
- Neuveglise C, Chalvet F, Wincker P, Gaillardin C, Casarégola S. 2005. Mutator-like element in the yeast *Yarrowia lipolytica* displays multiple alternative splicings. *Eukaryot Cell* **4**: 615–624.
- Nikolski M, Sherman DJ. 2007. Family relationships: Should consensus reign? Consensus clustering for protein families. *Bioinformatics* **23**: e71–e76.
- Noble SM, Johnson AD. 2007. Genetics of *Candida albicans*, a diploid human fungal pathogen. *Annu Rev Genet* **41**: 193–211.
- Ooi SL, Pan X, Peyser BD, Ye P, Meluh PB, Yuan DS, Irizarry RA, Bader JS, Spencer FA, Boeke JD. 2006. Global synthetic-lethality analysis and yeast functional profiling. *Trends Genet* **22**: 56–63.
- Oura T, Kajiwara S. 2008. Substrate specificity and regioselectivity of delta12 and omega3 fatty acid desaturases from *Saccharomyces kluyveri*. *Biosci Biotechnol Biochem* **72**: 3174–3179.
- Payen C, Koszul R, Dujon B, Fischer G. 2008. Segmental duplications arise from pol32-dependent repair of broken forks through two alternative replication-based mechanisms. *PLoS Genet* **4**: e1000175. doi: 10.1371/journal.pgen.1000175.
- Payen C, Fischer G, Marck C, Proux C, Sherman DJ, Coppée J-Y, Johnston M, Dujon B, Neuveglise C. 2009. Unusual composition of a yeast chromosome arm is associated with its delayed replication. *Genome Res* (in press). doi: 10.1101/gr.090605.108.
- Replansky T, Koufopanou V, Greig D, Bell G. 2008. *Saccharomyces sensu stricto* as a model system for evolution and ecology. *Trends Ecol Evol* **23**: 494–501.
- Rivera MC, Jain R, Moore JE, Lake JA. 1998. Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci* **95**: 6239–6244.
- Rolland T, Neuveglise C, Sacerdot C, Dujon B. 2009. Insertion of horizontally transferred genes within conserved syntenic regions of yeast genomes. *PLoS ONE* (in press).
- Rubin E, Lithwick G, Levy AA. 2001. Structure and evolution of the hAT transposon superfamily. *Genetics* **158**: 949–957.
- Rustchenko EP, Sherman F. 1994. Physical constitution of ribosomal genes in common strains of *Saccharomyces cerevisiae*. *Yeast* **10**: 1157–1171.
- Samonte RV, Eichler EE. 2002. Segmental duplications and the evolution of the primate genome. *Nat Rev Genet* **3**: 65–72.
- Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* **440**: 341–345.
- Scannell DR, Butler G, Wolfe KH. 2007a. Yeast genome evolution—the origin of the species. *Yeast* **24**: 929–942.
- Scannell DR, Frank AC, Conant GC, Byrne KP, Woolfit M, Wolfe KH. 2007b. Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. *Proc Natl Acad Sci* **104**: 8397–8402.
- Schacherer J, Kruglyak L. 2009. Comprehensive polymorphism survey elucidates population structure of *S. cerevisiae*. *Nature* **458**: 342–346.
- Solieri L, Giudici P. 2007. Yeasts associated to traditional balsamic vinegar: Ecological and technological features. *Int J Food Microbiol* **125**: 36–45.
- Solieri L, Cassanelli S, Croce MA, Giudici P. 2008. Genome size and ploidy level: New insights for elucidating relationships in *Zygosaccharomyces* species. *Fungal Genet Biol* **45**: 1582–1590.
- Sor F, Fukuhara H. 1989. Analysis of chromosomal DNA patterns of the genus *Kluyveromyces*. *Yeast* **5**: 1–10.
- Souciet JL, Aigle M, Artiguenave F, Blandin G, Bolotin-Fukuhara M, Bon E, Brottier P, Casarégola S, de Montigny J, Dujon B, et al. 2000. Genomic exploration of the hemiascomycetous yeasts: 1. A set of yeast species for molecular evolution studies. *FEBS Lett* **487**: 3–12.
- Sychrova H, Braun V, Potier S, Souciet JL. 2000. Organization of specific genomic regions of *Zygosaccharomyces rouxii* and *Pichia sorbitophila*: Comparison with *Saccharomyces cerevisiae*. *Yeast* **16**: 1377–1385.
- Talla E, Anthouard V, Bouchier C, Frangeul L, Dujon B. 2005. The complete mitochondrial genome of the yeast *Kluyveromyces thermotolerans*. *FEBS Lett* **579**: 30–40.
- Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, et al. 2003. The COG database: An updated version includes eukaryotes. *BMC Bioinformatics* **4**: 41. doi: 10.1186/1471-2105-4-41.
- Tesler G. 2002. GRIMM: Genome rearrangements web server. *Bioinformatics* **18**: 492–493.
- Vilela-Moura A, Schuller D, Mendes-Faia M, Côte-Real M. 2008. Reduction of volatile acidity of wines by selected yeast strains. *Appl Microbiol Biotechnol* **80**: 881–890.

- Wapinski I, Pfeffer A, Friedman N, Regev A. 2007. Automatic genome-wide reconstruction of phylogenetic gene tree. *Bioinformatics* **23**: 549–558.
- Wei W, McCusker JH, Hyman RW, Jones T, Ning Y, Cao Z, Gu Z, Bruno D, Miranda M, Nguyen M, et al. 2007. Genome sequencing and comparative analysis of *Saccharomyces cerevisiae* strain YJM789. *Proc Natl Acad Sci* **104**: 12825–12830.
- Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**: 708–713.
- Wong S, Butler G, Wolfe KH. 2002. Gene order evolution and paleopolyploidy in hemiascomycete yeasts. *Proc Natl Acad Sci* **99**: 9272–9277.
- Wood V, Gwilliam R, Rajandream MA, Lyne R, Stewart A, Sgouros J, Peat N, Hayles J, Baker S, Basham D, et al. 2002. The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415**: 871–880.
- Wu CH, Nikolskaya A, Huang H, Yeh LS, Natale DA, Vinayaka CR, Hu ZZ, Mazumder R, Kumar S, Kourtesis P, et al. 2004. PIRSF: Family classification system at the Protein Information Resource. *Nucleic Acids Res* **32**: D112–D114.
- Xu J, Saunders CW, Hu P, Grant RA, Boekhout T, Kuramae EE, Kronstad JW, DeAngelis YM, Reeder NL, Johnstone KR, et al. 2007. Dandruff-associated *Malassezia* genomes reveal convergent and divergent virulence traits shared with plant and human fungal pathogens. *Proc Natl Acad Sci* **104**: 18730–18735.
- Yu J, Wang J, Lin W, Li S, Li H, Zhou J, Ni P, Dong W, Hu S, Zeng C, et al. 2005. The genomes of *Oryza sativa*: A history of duplications. *PLoS Biol* **3**: e38. doi: 10.1371/journal.pbio.0030038.
- Zdobnov EM, Bork P. 2007. Quantification of insect genome divergence. *Trends Genet* **23**: 16–20.

Received January 23, 2009; accepted in revised form April 20, 2009.

Supplementary figure 5



**ANNEXE 2 : COMBINED PHYLOGENETIC AND
NEIGHBOURHOOD ANALYSIS OF THE HEXOSE
TRANSPORTERS AND GLUCOSE SENSORS IN YEASTS**

RESEARCH ARTICLE

Combined phylogenetic and neighbourhood analysis of the hexose transporters and glucose sensors in yeasts

Margarida Palma¹, Marie-Line Seret² & Philippe V. Baret²

¹Life Sciences Department, Centro de Recursos Microbiológicos, New University of Lisbon, Caparica, Portugal; and ²Unité de génétique (GENA-POPS), Faculté d'ingénierie biologique, agronomique et environnementale, Université catholique de Louvain, Louvain-la-Neuve, Belgium

Correspondence: Margarida Palma, Life Sciences Department, Centro de Recursos Microbiológicos, New University of Lisbon, 2829-516 Caparica, Portugal. Tel./fax: +35 12 1294 8530; e-mail: mip@fct.unl.pt

Received 22 January 2009; revised 21 February 2009; accepted 24 February 2009.
First published online May 2009.

DOI:10.1111/j.1567-1364.2009.00511.x

Editor: André Goffeau

Keywords

sugar transporters; glucose sensors; neighbourhood analysis; phylogeny; HXT; evolutionary lineages.

Abstract

The sugar porter family in yeasts encompasses a wide variety of transporters including the hexose transporters and glucose sensors. We analysed a total of 75 members from both groups in nine hemiascomycetous species, with complete and well-annotated genomes: *Saccharomyces cerevisiae*, *Candida glabrata*, *Zygosaccharomyces rouxii*, *Kluyveromyces thermotolerans*, *Saccharomyces kluyverii*, *Kluyveromyces lactis*, *Eremothecium gossypii*, *Debaryomyces hansenii* and *Yarrowia lipolytica*. We present a model for the evolution of the hexose transporters and glucose sensors, supported by two types of complementary evidences: phylogeny and neighbourhood analysis. Five lineages of evolution were identified and discussed according to different mechanisms of gene evolution: lineage A for *HXT1*, *HXT3*, *HXT4*, *HXT5*, *HXT6* and *HXT7*; lineage B for *HXT2* and *HXT10*; lineage C for *HXT8*; lineage D for *HXT14*; and lineage E for *SNF3* and *RGT2*.

Introduction

Saccharomyces cerevisiae natural habitat is sugar-rich fruit juices. The uptake of sugars across the plasma membrane is a decisive step in sugar metabolism and a rate-limiting step of glycolysis (Becker & Betz, 1972; Gancedo & Serrano, 1989). Since their discovery (Lewis & Bisson, 1991; Ko *et al.*, 1993; Wendell & Bisson, 1993; Theodoris *et al.*, 1994; Reifemberger *et al.*, 1995), the phylogeny of the sugar porter (SP) family of *S. cerevisiae* has been analysed by different authors (Kruckeberg, 1996; Nelissen *et al.*, 1997; Wiczorke *et al.*, 1999). More recently, 214 putative sugar porters were identified from eight hemiascomycetous yeasts *Candida glabrata*, *Kluyveromyces lactis*, *Eremothecium gossypii*, *Debaryomyces hansenii*, *Yarrowia lipolytica*, *Candida albicans*, *Pichia stipitis* and *S. cerevisiae* (Palma *et al.*, 2007). The transported substrates were classified according to the transport classification system (Saier *et al.*, 2006) in nine clusters: hexose transporters, glucose sensors, general α -glucoside: H⁺ symporter, fructose: H⁺ symporter, lactose: H⁺ symporter, high-affinity glucose transporters, myo-inositol: H⁺ symporter, glycerol: H⁺ symporter and quinate: H⁺ symporter. Approximately one-third of the 214 putative sugar porters still has unknown function.

In *S. cerevisiae*, the members of the SP family span a size range of 486–884 amino acids. The largest proteins are the glucose sensors, which have a long C-terminal tail (Özcan *et al.*, 1996). The 12 putative transmembrane domains of the hexose transporters are highly conserved in contrast to the variable cytosolic amino- and carboxyl-termini (Kruckeberg, 1996).

In *S. cerevisiae*, the hexose transporters are encoded by the *HXT1* to *HXT17* and *GAL2* genes (for reviews, see Lagunas, 1993; Bisson *et al.*, 1993; Boles & Hollenberg, 1997; Leandro *et al.*, 2009) and operate through a facilitated diffusion mechanism (Jansen *et al.*, 2002; Maier *et al.*, 2002). The most important hexose transporters are Hxt1p, Hxt2p, Hxt3p, Hxt4p, Hxt6p and Hxt7p; each of these six transporters is able to support growth on glucose (Reifemberger *et al.*, 1995). They also transport fructose and mannose (Wiczorke *et al.*, 1999). Two other genes in SP family, *SNF3* and *RGT2*, act as glucose sensors (Özcan *et al.*, 1996), which do not transport glucose, but are responsible for sensing and signalling the availability of this substrate (Kruckeberg *et al.*, 1998; Özcan *et al.*, 1998). The *HXT12* is a pseudo-gene that will not be mentioned in this study (Wiczorke *et al.*, 1999).

The increasing number of yeast genome sequences provides a wealth of new information on speciation and

genome evolution. The widely used phylogenetic analysis does not *per se* show the whole history of a specific group of genes. Other evidences such as synteny (the conserved presence of neighbouring genes along a given chromosome or chromosomal segment) and conservation of functional motifs or pathways (Dujon, 2005) will provide an improved understanding of gene lineages.

The aim of this work is to outline the evolution of hexose transporters and glucose sensors considering the availability of nine complete and well-annotated genome sequences, including *S. cerevisiae*, *C. glabrata*, *K. lactis*, *E. gossypii*, *D. hansenii* and *Y. lipolytica*, and three novel yeast genomes (J.-L. Souciet *et al.*, unpublished data) provided by the Génolevures project: *Saccharomyces kluyverii*, *Zygosaccharomyces rouxii* and *Kluyveromyces thermotolerans*.

The evolution of hexose transporters and glucose sensors was inferred taking into consideration the complementary phylogeny and gene neighbourhood analyses.

Materials and methods

Genomes

The protein sequences of *S. cerevisiae* (SACE), *C. glabrata* (CAGL), *Z. rouxii* (ZYRO), *K. thermotolerans* (KLTH), *S. kluyverii* (SAKL), *K. lactis* (KLLA), *E. gossypii* (ERGO), *D. hansenii* (DEHA) and *Y. lipolytica* (YALI) were retrieved from the Génolevures database. The *Z. rouxii* and *K. thermotolerans* genomes were recently sequenced by the Génoscope, and the *S. kluyverii* genome was sequenced by the Washington University Genome Sequencing Centre. We limited our analysis to fully annotated genomes. The Génolevures consortium annotated the new genomes of ZYRO, KLTH and SAKL. The annotation of KLLA was updated by Génolevures (J.-L. Souciet *et al.*, unpublished data). The annotation of ERGO was recently updated by Gattiker *et al.* (2007).

Subset of orthologues defined by neighbourhood and similarity (SONS)

The Génolevures family GL3C002 (Nikolski & Sherman, 2007) comprised 218 members of sugar porters (Supporting Information, Table S1), within the species studied. These protein sequences were identified by identification of orthologues by neighbourhood and similarity (IONS) (M.-L. Seret *et al.*, unpublished data). The rationale of the method is that two genes of different yeast species whose translation products belong to the same family (homologues by similarity) will be members of the same SONS if they share at least one pair of neighbours that are also homologous to each other by similarity. The process is reiterated for all possible heterospecific pairwise comparisons of homologues deduced from the protein families. Homologues that do not share a pair of homologous neighbours are divided into two distinct SONS.

Within the 218 members of sugar porters considered in the present study, 54 were selected as 'hexose transporters' and 21 as 'glucose sensors' according to phylogeny and similarities to *S. cerevisiae* members.

Phylogeny

Multiple alignment of 75 selected hexose transporters and glucose sensors sequences was calculated by MUSCLE (Edgar, 2004) and processed using the PHYLIP package (Felsenstein, 2005). In a first step, SEQBOOT, a bootstrapping tool, allowed to generate 100 data sets that were analysed by PROML. In a second step, the CONSENSE program generates a consensus phylogenetic tree. We chose a 50% threshold for the consensus tree, meaning that all the generated branches have a frequency of occurrence of > 50% in the 100 generated trees. The DENDROSCOPE application (Huson *et al.*, 2007) was used to visualize the phylogenetic trees.

Results

Figure 1 shows a phylogenetic tree of the selected 75 hexose transporters or glucose sensors of the GL3C002 family. The phylogenetic clusters are defined based on the tree nodes and the bootstrap scores and labelled a to j.

By combining the phylogenetic information contained in Fig. 1 and that given by neighbourhood analysis, a model for the differentiation of hexose transporters and glucose sensors can be proposed. Five different lineages are illustrated in Fig. 2: A gave rise to the SACE transporters Hxt1p, Hxt3p, Hxt4p, Hxt5p, Hxt6p and Hxt7p; B gave rise to Hxt2p and Hxt10p; C gave rise to Hxt8p; D is the lineage of Hxt14p; and E gave rise to Snf3p and Rgt2p. Table 1 illustrates the distribution of genes in each lineage. The column labelled 'Others' lists genes that were not included in the lineages presented because there was no evidence of a relationship either by neighbourhood or phylogeny.

Lineage A: the relationship of the Hxt1p, Hxt3p, Hxt4p, Hxt5p, Hxt6p and Hxt7p transporters (cluster a, c, d and e)

The lineage A comprises 26 members and is highly represented in SACE (six members) and CAGL (seven members). Figure 3 illustrates the relationship among lineage A transporters. Neighbour genes belonging to the same family are considered homologues by similarity. The SACE *HXT3*, *HXT6* and *HXT7* genes are on chromosome D. Hxt7p is very similar to Hxt6p, as their protein sequences differ only by one amino acid out of 570. On SACE chromosome H, three adjacent copies of hexose transporters (*HXT5*, *HXT1* and *HXT4*) share identity with the triplet *HXT3*, *HXT6* and *HXT7* located on chromosome D.

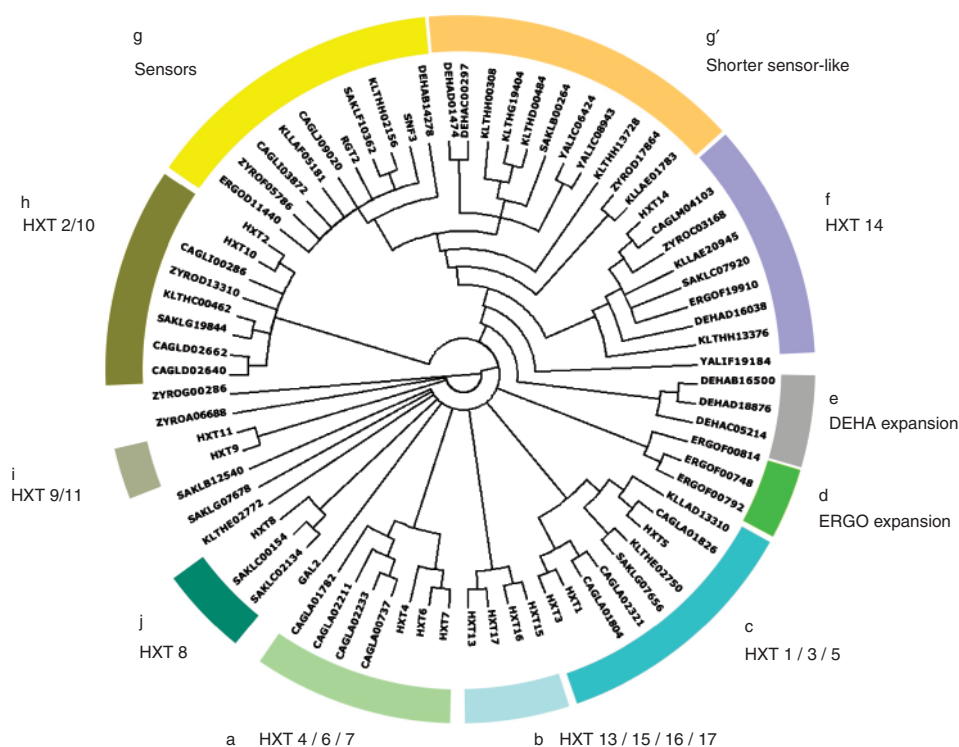


Fig. 1. A phylogenetic tree of the 75 genes based on a maximum likelihood model applied to 100 bootstrap data sets. Only nodes with > 50% support are represented.

Chromosome A from CAGL bears three members (A02321, A02233 and A02211) that share high homology with the *HXT3*, *HXT6* and *HXT7* genes, respectively. The transporters A02233 and A02211 are more different from each other than Hxt6p and Hxt7p (96% identity vs. 99% identity, respectively). Still, in CAGL chromosome A, but located 14 coding sequences apart from the A02321, A02233 and A02211 triplet, there is another set of three similar genes (A01826, A01804 and A01782). At the other end of chromosome A, the CAGL transporter A00737 is phylogenetically related to the six other CAGL members of the lineage A, sharing 79% amino acid identity with CAGL A02233. According to the SONS analysis, it is a singleton, sharing no neighbours with other members of the lineage.

In ERGO, SAKL, KLTH and ZYRO, a very significant neighbourhood (five to seven shared neighbours) connect all members of lineage A. Despite being classified in the species-specific phylogenetic cluster d, the three ERGO members share neighbours with the remaining genes in this lineage. All the members shown in Fig. 3a share at least two

common families in their neighbourhood. A tandem repeat is present in SAKL and KLTH. YALI and DEHA comprise members of this lineage (amino acid identity with SACE Hxt transporters > 50%), but they are not in the same phylogenetic cluster. Figure 3b demonstrates the neighbour relationships between ZYRO, a pre-whole genome duplication (WGD) species and the post-WGD species SACE and CAGL.

Lineage B: the relationship of the Hxt2p and Hxt10p transporters (cluster h)

Lineage B consists of eight members. The SAKL subtelomeric member G19844 shares four neighbours with KLTH C00462. Other members are phylogenetically related, but cannot be confirmed by neighbourhood analysis. In CAGL and SACE, we observe once more an increase in the number of glucose transporters per species: three for CAGL and two for SACE. The ZYRO member and the CAGL I00286 share a

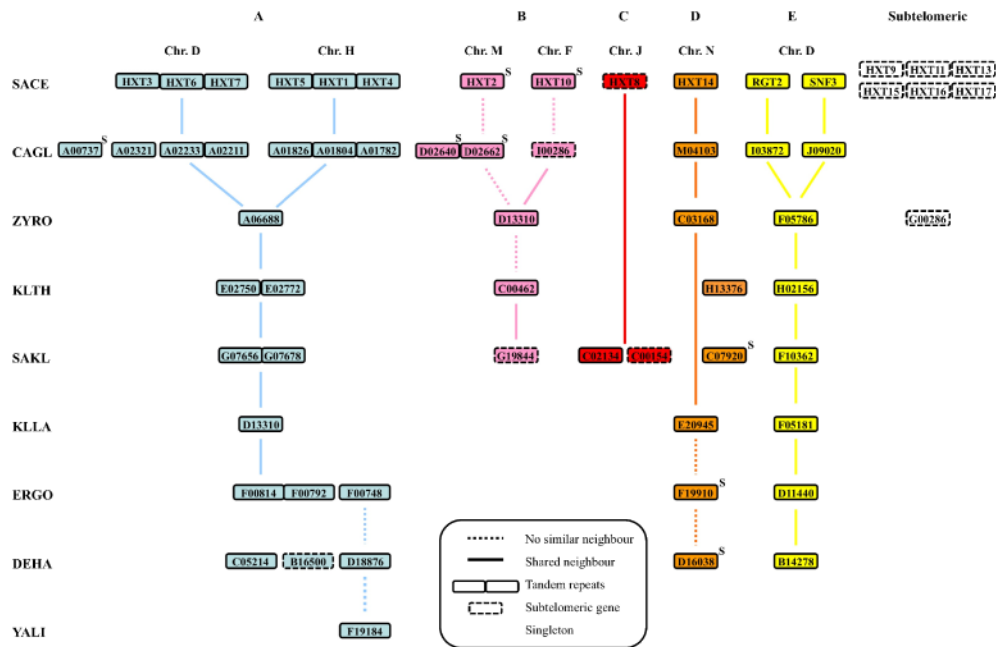


Fig. 2. The five different lineages (A to E) of hexose transporters and glucose sensors. The members shown in dashed boxes were encoded by genes located in subtelomeric regions of the chromosome. Boxes with a superscript 'S' represent genes considered as singletons by neighbourhood analysis. Full lines represent the genes that share a neighbourhood, and dashed lines illustrate relationships based on phylogeny or identity only.

Table 1. Number of genes in each lineage

Species	Lineage						Total
	A	B	C	D	E	Others*	
SACE	6	2	1	1	2	7	19
CAGL	7	3		1	2		13
ZYRO	1	1		1	1	2	6
KLTH	2	1		1	1	4	9
SAKL	2	1	2	1	1	2	9
KLLA	1			1	1	1	4
ERGO	3			1	1		5
DEHA	3			1	1	2	7
YALI	1					2	3
Total	26	8	3	8	10	20	75

*The genes in this column were not included in the lineages described in Fig. 2.

common neighbour and the latter is located in the subtelomeric region of chromosome I.

Lineage C: the relationship of the Hxt8p transporter (cluster j)

The lineage C comprises three members: SACE Hxt8p, SAKL C00154 and SAKL C02134. Both SAKL transporters share 78% and 74% amino acid identity with Hxt8p,

respectively. They share both the same phylogenetic cluster and one neighbour. HXT8 and SAKL C00154 are located in subtelomeric regions.

Lineage D: the relationship of the Hxt14p transporter (cluster f)

Lineage D consists of eight members and was established by a combination of neighbourhood and phylogeny. The

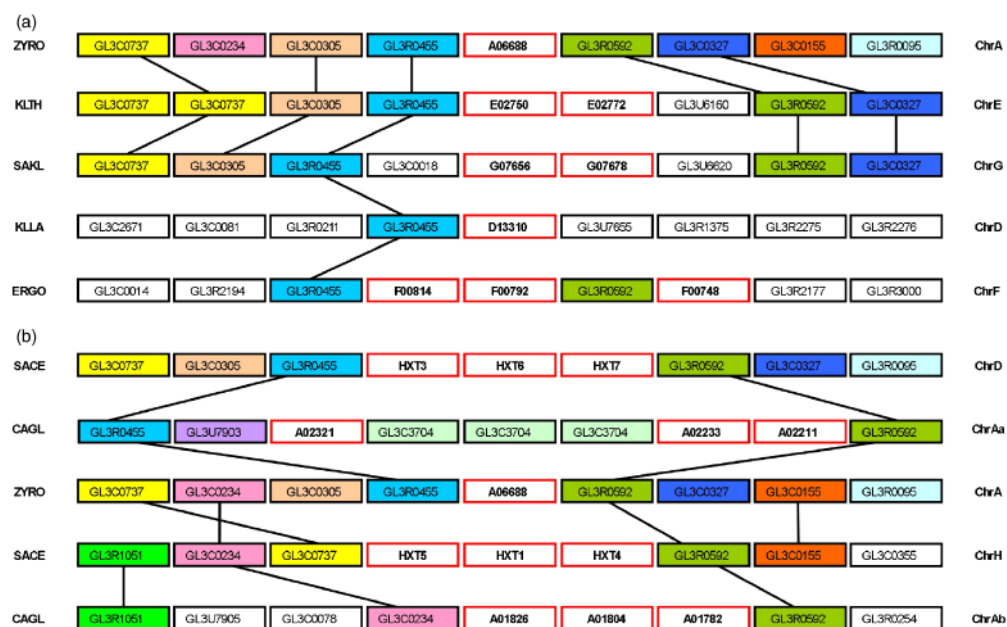


Fig. 3. Neighbourhood analysis of lineage A in (a) ERGO, KLLA, SAKL, KLTH and ZYRO and (b) ZYRO, CAGL and SACE. Each line represents a distinct chromosomal region around the hexose transporter (in red). The ZYRO chromosomal segment is repeated on both parts of the figure and connects the pre-WGD species (a) and the post-WGD species (b). Neighbours with equally coloured boxes represent members of the same family.

hexose transporters from KLLA, ZYRO, CAGL and SACE share three or more neighbours, which make this a clear lineage. The DEHA D16038, ERGO G19910 and SAKL C07920 are singletons in neighbourhood analysis but they are part of phylogenetic cluster f. They share 39%, 41% and 48% amino acid identity with SACE Hxt14p, respectively. KLTH H13376, which shares 36% amino acid identity with Hxt14p, was added to this lineage based on phylogeny only.

Lineage E: the relationship of the glucose sensors Snf3p and Rgt2p (cluster g)

Lineage E includes cluster g members that share sequence similarities with Snf3p and Rgt2p. Figures 1 and 4 show that cluster g comprises 10 members, one per species in DEHA to ZYRO, and two copies in CAGL and SACE that clearly result from WGD. No YALI member is detected in this cluster. The genes in lineage E share a consistent neighbourhood. The encoded proteins comprise from 664 to 884 amino acid residues (compared with 533–592 residues for the HXT sequences) with a 100–300 amino acid cytoplasmic C-terminus including the 25-amino acid sequence defined as the Özcan motif (Özcan *et al.*, 1998).

Although cluster g' shares sequence similarities with this lineage, its 11 members are shorter (457–578-amino acid residues) than the glucose sensors of cluster g. Members from the g' cluster are identified in YALI, DEHA, KLLA, SAKL, KLTH and ZYRO, but not in ERGO, CAGL and SACE. A transition between clusters g and g' is observed in DEHA. Despite sharing neighbours with the other members of cluster g and being characterized by its long sequence of 731 amino acids, DEHA B14278 lacks the Özcan sensor signature. On the other hand, DEHA C00297 and DEHA D01474, comprising 457 and 527 amino acid residues, respectively, clearly belong to cluster g'.

Other genes

A total of 20 transporters were not included in any lineage because there were no reliable neighbourhood and phylogenetic data. The location of some genes in subtelomeres does not allow neighbourhood analysis because these are dynamic regions.

An important number (21%) of the sugar porters analysed is located in subtelomeric regions of chromosomes. As proposed by Fairhead & Dujon (2006), we considered

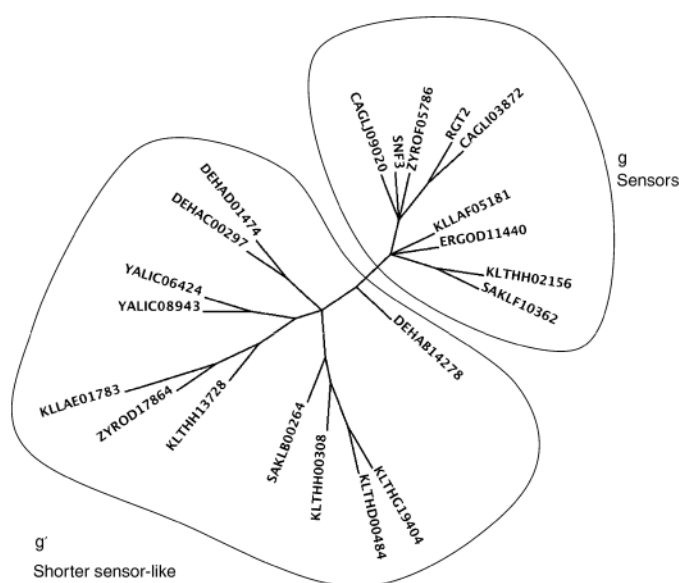


Fig. 4. Phylogeny of sensors and sensor-like proteins based on a maximum likelihood model applied to 100 bootstrap data sets. Only nodes with > 50% support are represented.

subtelomeres being the last 30 kb of each chromosome. SACE possesses seven genes in subtelomeric regions: *HXT8*, *HXT9*, *HXT11*, *HXT13*, *HXT15*, *HXT16* and *HXT17* (clusters b, j and i). The *HXT15* and *HXT16* genes are located in the subtelomeric regions of chromosomes D and J, respectively. They share 99% amino acid identity and a considerable number of neighbours of the same family. The same occurs with *HXT13* and *HXT17*, which share 98% amino acid identity and are located in chromosome E and N arms, respectively. *HXT9* and *HXT11* are in chromosomes J and O, respectively, and they share 97% amino acid identity. They are also allocated in the same phylogenetic cluster and share a neighbour of the same family.

Despite sharing 76% amino acid identity, SACE Gal2p and SAKL B12540 were not considered as a lineage because there was no neighbourhood or phylogenetic evidence to justify it.

Discussion

It is thought that yeasts have separated from other organisms about one billion years ago and differentiated into *Euascomycetes*, *Hemiascomycetes* and *Archiascomycetes*. The present work considers only the 'Saccharomycetes' complex from the *Hemiascomycetes* phylum (Kurtzman & Robnett, 2003) represented by seven species whose full genome sequences offer a unique opportunity to explore eukaryotic evolution mechanisms. We present a model for the evolution of two clusters from the large SP family, the hexose

transporters and glucose sensors, supported by two types of evidences: phylogeny and neighbourhood analysis.

Five lineages of evolution were identified from YALI to SACE. Each lineage can be described according to three different mechanisms: (1) conservation, gain or loss of genes during differentiation of each species, (2) WGD and (3) single-gene or segmental duplications within species.

The lineage of the Hxt1p, Hxt3p, Hxt4p, Hxt5p, Hxt6p and Hxt7p transporters involves the occurrence of the three mechanisms. We assume that the YALI putative hexose transporter F19184 represents the ancestor of the whole lineage, despite the fact that it shares no neighbours with the other studied species. Regarding the other members of the lineage, we identified a gene triplication in ERGO. Two of these three genes were lost in KLLA. This could be related to the low concentrations of hexoses in the natural lactose-rich habitat of this species (Naumova *et al.*, 2004). In SAKL and KLTH, a duplication event occurred, which is consistent with the presence of hexoses in their natural habitat (Kurtzman & Fell, 1998). The ZYRO species has lost one member. In post-WGD species, seven and six hexose transporters of the same lineage were observed in CAGL and SACE, respectively. The quasi identity of Hxt6p and Hxt7p (one amino acid residue different) may indicate that they are recently duplicated genes. In fact, Brown *et al.* (1998) reported the occurrence of multiple duplications of the high-affinity glucose transporters HXT6 and HXT7 when a population of yeast cells underwent 450 generations of glucose-limited growth. This emphasizes the potential role

of environmental pressure on the number of copies of hexose transporters. The dynamics that led to this large number of hexose transporters cannot be detailed as multiple rearrangements masks the real sequence of evolutionary events. In SACE, lineage A comprises several of the most efficient hexose transporters. As reported by Reifengerger *et al.* (1995), each of the Hxt1p, Hxt2p, Hxt3p, Hxt4p, Hxt6p and Hxt7p is able to support glucose growth on its own. Hxt2p was considered part of another lineage, together with Hxt10p. Both were considered singletons by neighbourhood analysis probably because they were translocated to another neighbourhood.

Gene loss occurred in the Hxt8p lineage as SAKL and SACE are the only species that have members of this lineage.

The 10 members of the phylogenetic cluster g (the glucose sensors) as well as the 11 members of cluster g' (the shorter sensor-like proteins) were identified by similarity to the SACE glucose sensors Snf3p and Rgt2p. Their distribution in the studied hemiascomycetous species suggests the following scenario. The YALI members of cluster g' comprising from 457- to 578-amino acid residues may belong to a subfamily of putative hexose transporters of undetermined specificity. In DEHA, a copy of a short g' gene fused to a gene fragment encoding a cytoplasmic 'signal sequence' of 100–300-amino acid residues that progressively acquired the Özcan motif. The fused gene produced the 'glucose sensors'

of cluster g that was conserved from DEHA to SACE. In contrast, the original 'short g' sensor' genes present in YALI, DEHA, KLLA, KLTH, SAKL and ZYRO were lost in species such as ERGO, CAGL and SACE.

As presented in Fig. 2 by dashed-line boxes, 13 hexose transporters are located in subtelomeric regions. Seven of them are from *S. cerevisiae*. Moreover, the huge number of subtelomeric hexose transporter genes is a characteristic of *S. cerevisiae* strains, and are absent in older species such as *Saccharomyces castellii*. It was reported by Kellis *et al.* (2003) that 69% of species-specific genes in *S. cerevisiae* are located in subtelomeres. This was confirmed in a broader analysis involving other hemiascomycetous yeasts (Fabre *et al.*, 2004). Subtelomeric gene families undergo expansions and contractions in different strains or species (Fabre *et al.*, 2004; Fairhead & Dujon, 2006). The subtelomeric genes are thought to be involved in adaptation to the environment and to promote *de novo* gene creation by recombination inside gene repeats and between different chromosomal arms. Among other genes involved in sugar utilization, the HXT13 and HXT17, HXT15 and HXT16, HXT11 and HXT9 genes are impressive examples of subtelomeric amplifications that could have occurred during extensive human use of *S. cerevisiae* in brewing, baking and winemaking.

The comparison of the SACE laboratory strain S188C with the winery isolate RM11-1a and the clinical strain

Table 2. Hexose transporters in some species of the 'Saccharomycetes complex'

Strain Gene	<i>S. cerevisiae</i> S288C*	<i>S. cerevisiae</i> RM11-1a†	<i>S. cerevisiae</i> YJM789‡	<i>C. glabrata</i> §	<i>S. castellii</i> ¶
HXT1	1	1	1	9	1
HXT2	1	1	1		1
HXT3	1	1	1		1
HXT4	1	1	1		0
HXT5	1	1	1		1
HXT6	1	0	1		1
HXT7	1	1	1		0
HXT8	1	1	0	0	0
HXT9	1	0	1	0	0
HXT10	1	1	1	1	0
HXT11	1	1	1	0	0
HXT13	1	1	1	0	0
HXT14	1	1	1	1	1
HXT15	1	1	1	0	0
HXT16	1	0	0	0	0
HXT17	1	1	1	0	0
GAL2	1	1	1	0	0
Total	17	12	15	11	6

*SGD Project. 'Saccharomyces Genome Database' (2008).

†Saccharomyces cerevisiae RM11-1a Sequencing Project. Broad Institute of Harvard and MIT (<http://www.broad.mit.edu>).

‡Wei *et al.* (2007).

§Sherman *et al.* (2006). *Candida glabrata* possesses nine members highly similar to the first seven hexose transporter genes HXT1-HXT7.

¶Byrne & Wolfe (2005).

^{||}Protein sequence interrupted by Stop codons.

YJM789 isolated from the lungs of an AIDS patient shows some variations of the number of hexose transporter genes (Table 2). *HXT8* and *HXT16* are absent in the pathogenic strain YJM789. *HXT9*, *HXT6* and *HXT16* are absent in the RM11-1a strain and *Hxt4p* and *Hxt8p* sequences are interrupted by stop codons. In contrast, species such as *S. castellii* carries *HXT1*, *HXT2*, *HXT3*, *HXT5*, *HXT7* and *HXT14* orthologues (Byrne & Wolfe, 2005) and *C. glabrata* contains orthologues of *HXT1-7*, *HXT10* and *HXT14* only.

In conclusion, the combination of both phylogeny and neighbourhood analysis allowed us to delineate several new aspects of the evolution of the subfamily of glucose transporters and sensors in *Hemiascomycetes*. It is, however, important to note that genome dynamics sometimes masks the real sequence of events and that for some genes it is not yet possible to identify a clear evolutionary line.

Acknowledgements

We would like to thank Paulo Dias and Professor André Goffeau for helpful discussions, and Professor Isabel Sá-Correia for her support. M.P. is a recipient of a PhD grant (SFRH/BD/27663/2006) from Fundação para a Ciência e a Tecnologia, Portugal. A part of this work was made possible, thanks to privileged access to unpublished annotation data provided by the Génolevures consortium coordinated by Jean-Luc Souciet.

References

- Becker JU & Betz A (1972) Membrane transport as controlling pacemaker of glycolysis in *Saccharomyces carlsbergensis*. *Biochim Biophys Acta* **274**: 584–597.
- Bisson LF, Coons DM, Kruckeberg AL & Lewis DA (1993) Yeast sugar transporters. *Crit Rev Biochem Mol Biol* **28**: 259–308.
- Boles E & Hollenberg CP (1997) The molecular genetics of hexose transporters in yeasts. *FEMS Microbiol Rev* **21**: 85–111.
- Brown CJ, Todd KM & Rosenzweig RF (1998) Multiple duplications of yeast hexose transport genes in response to selection in a glucose. Limited environment. *Mol Biol Evol* **15**: 931–942.
- Byrne KP & Wolfe KH (2005) The yeast gene order browser: combining curated homology and syntenic context reveals gene fate in polyploidy species. *Genome Res* **15**: 1456–1461.
- Dujon B (2005) Hemiascomycetous yeasts at the forefront of comparative genomics. *Curr Opin Genet Dev* **15**: 614–620.
- Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**: 1–19.
- Fabre E, Muller H, Therizols P, Lafontaine I, Dujon B & Fairhead C (2004) Comparative genomics in hemiascomycete yeast: evolution of sex, silencing and subtelomeres. *Mol Biol Evol* **22**: 856–873.
- Fairhead C & Dujon B (2006) Structure of *Kluyveromyces lactis* subtelomeres: duplications and gene content. *FEMS Yeast Res* **6**: 428–441.
- Felsenstein J (2005) *PHYLIP (Phylogeny Inference Package) Version 3.65*. University of Washington, Seattle, WA.
- Gancedo C & Serrano R (1989) Energy-yielding metabolism. *The Yeasts*, Vol. 3 (Rose AH & Harrison JS, eds), pp. 205–259. Academic Press, London.
- Gattiker A, Rischatsch R, Demougin P, Voegeli S, Dietrich FS, Philippsen P & Primig M (2007) *Ashbya* genome database 3.0: a cross-species genome and transcriptome browser for yeast biologists. *BMC Genomics* **8**: 9.
- Huson DH, Richter DC, Rausch C, Dezulian T, Franz M & Rupp R (2007) Dendroscope: an interactive viewer for large phylogenetic trees. *BMC Bioinformatics* **8**: 460.
- Jansen MLA, De Winde JH & Pronk JT (2002) Hxt-carrier-mediated glucose efflux upon exposure of *Saccharomyces cerevisiae* to excess maltose. *Appl Environ Microb* **68**: 4259–4265.
- Kellis M, Patterson N, Endrizzi M, Birren B & Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**: 241–263.
- Ko CH, Liang H & Gaber RF (1993) Roles of multiple glucose transporters in *Saccharomyces cerevisiae*. *Mol Cell Biol* **13**: 638–648.
- Kruckeberg AL (1996) The hexose transporter family of *Saccharomyces cerevisiae*. *Arch Microbiol* **166**: 283–292.
- Kruckeberg AL, Walsh MC & Van Dam K (1998) How do yeast sense glucose? *BioEssays* **20**: 972–976.
- Kurtzman CP & Fell JW (1998) *The Yeasts, a Taxonomic Study*, 4th edn. Elsevier Science B.V., Amsterdam.
- Kurtzman CP & Robnett CJ (2003) Phylogenetic relationships among yeasts of the 'Saccharomyces complex' determined from multigene sequence analyses. *FEMS Yeast Res* **3**: 417–432.
- Lagunas R (1993) Sugar transport in *Saccharomyces*. *FEMS Microbiol Rev* **104**: 229–242.
- Leandro MJ, Fonseca C & Gonçalves P (2009) Hexose and pentose transport in ascomycetous yeasts: an overview. *FEMS Yeast Research* **9**: 511–525.
- Lewis DA & Bisson LF (1991) The HXT1 gene product of *Saccharomyces cerevisiae* is a new member of the family of hexose transporters. *Mol Cell Biol* **11**: 3804–3813.
- Maier A, Volker B, Boles E & Fuhrmann GF (2002) Characterisation of glucose transport in *Saccharomyces cerevisiae* with plasma membrane vesicles (countertransport) and intact cells (initial uptake) with single Hxt1, Hxt2, Hxt3, Hxt4, Hxt7 or Gal2 transporters. *FEMS Yeast Res* **4**: 539–550.
- Naumova ES, Sukhotina NN & Naumov GI (2004) Molecular-genetic differentiation of the dairy yeast *Kluyveromyces lactis* and its closest wild relatives. *FEMS Yeast Res* **5**: 263–269.
- Nelissen B, DeWachter R & Goffeau A (1997) Classification of all putative permeases and other membrane plurispansers of the major facilitator superfamily encoded by the complete genome of *Saccharomyces cerevisiae*. *FEMS Microbiol Rev* **21**: 113–134.

- Nikolski M & Sherman DJ (2007) Family relationships: should consensus reign?—consensus clustering for protein families. *Bioinformatics* **23**: 71–76.
- Özcan S, Dover J, Rosenwald AG, Wölfl S & Johnston M (1996) Two glucose transporters in *Saccharomyces cerevisiae* are glucose sensors that generate a signal for induction of gene expression. *P Natl Acad Sci USA* **93**: 12428–12432.
- Özcan S, Dover J & Johnston M (1998) Glucose sensing and signaling by two glucose receptors in the yeast *Saccharomyces cerevisiae*. *EMBO J* **17**: 2566–2573.
- Palma M, Goffeau A, Spencer-Martins I & Baret PV (2007) A phylogenetic analysis of the Sugar Porters in hemiascomycetous yeasts. *J Mol Microb Biotech* **12**: 241–248.
- Reifenberger E, Freidel K & Ciriacy M (1995) Identification of novel *HXT* genes in *Saccharomyces cerevisiae* reveals the impact of individual hexose transporters on glycolytic flux. *Mol Microbiol* **16**: 157–167.
- Saier MH, Tran CV & Barabote RD (2006) TCDB: the transporter classification database for membrane transport protein analyses and information. *Nucleic Acids Res* **34**: 181–186.
- Sherman D, Durrens P, Iragne F, Beyne E, Nikolski M & Souciet J-L (2006) Génolevures complete genomes provide data and tools for comparative genomics of hemiascomycetous yeasts. *Nucleic Acids Res* **34**: D432–D435.
- Theodoris G, Fong NM, Coons DM & Bisson LF (1994) High-copy suppression of glucose transport defects by *HXT4* and regulatory elements in the promoters of the *HXT* genes in *Saccharomyces cerevisiae*. *Genetics* **137**: 957–966.
- Wei W, McCusker JH, Hyman RW *et al.* (2007) Genome sequencing and comparative analysis of *Saccharomyces cerevisiae* strain YJM789. *P Natl Acad Sci USA* **104**: 12825–12830.
- Wendell DL & Bisson LF (1993) Physiological characterization of putative high-affinity glucose transport protein Hxt2 of *Saccharomyces cerevisiae* by use of antisynthetic peptide antibodies. *J Bacteriol* **176**: 3730–3737.
- Wieczorke R, Krampe S, Weierstall T, Freidel K, Hollenberg CP & Boles E (1999) Concurrent knock-out of at least 20 transporter genes is required to block uptake of hexoses in *Saccharomyces cerevisiae*. *FEBS Lett* **464**: 123–128.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. Additional information about the hexose transporters and glucose sensors.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

**ANNEXE 3 : EVOLUTION OF THE 12-SPANNER DRUG:H⁺
ANTIporter DHA1 FAMILY IN HEMIASCOMYCETOUS
YEASTS**

Evolution of the 12-Spanner Drug:H⁺ Antiporter DHA1 Family in Hemiascomycetous Yeasts

Paulo Jorge Dias,¹ Marie-Line Seret,² André Goffeau,³ Isabel Sá Correia,¹ and Philippe V. Baret²

Abstract

Frequently, although not exclusively, multidrug resistance (MDR) results from the action of drug-efflux pumps, which are thought to be able to catalyze the active expulsion of several unrelated cytotoxic compounds out of the cell or their intracellular partitioning. The transporters of the major facilitator superfamily (MFS) presumably involved in MDR belong to the 12-spanner drug:H⁺ antiporter DHA1 or to the 14-spanner drug:H⁺ antiporter DHA2 families. The expression of most *Saccharomyces cerevisiae* DHA1 family members was found to confer broad chemoprotection. The evolution of the hemiascomycetous DHA1 proteins, belonging to the Génolevures GL3C007 family, was studied using a combined phylogenetic and gene neighborhood approach. The phylogenetic analysis of 189 DHA1 proteins belonging to the genome of 13 hemiascomycetous species identified 20 clusters. Eleven clusters contained no *S. cerevisiae* members. The phylogenetic clusters were analyzed by the IONS method developed for Identification of Orthologues by Neighborhood and Similarity. This allowed reconstructing the evolutionary history of most DHA1 members within 10 main gene lineages, spanning the whole hemiascomycetes clade, encompassing an evolutionary history of about 350 million years. In addition, five other more species specific lineages, spanning only two hemiascomycetous species, were identified. It is concluded that 57 out of the 143 members of the DHA1 hemiascomycetous members originated from gene duplication events. In half of these duplicates, the two members belong to different phylogenetic clusters, indicating that at least one of them has sufficiently differentiated to provide potential novel functions to this complex family from which most physiological substrates remain unknown.

Introduction

THE SIMULTANEOUS ACQUISITION of resistance to a wide range of structurally and functionally unrelated cytotoxic chemicals occurs in a wide variety of organisms, from bacteria to mammals (Hayes and Wolf, 1997; Higgins, 2007). This phenomenon poses a serious challenge in the treatment of tumors and infectious diseases and in the control of plant pathogens and weeds. Frequently, although not exclusively, multidrug resistance (MDR) results from the action of drug-efflux pumps, which are presumably able to catalyze the active expulsion of several unrelated cytotoxic compounds out of the cell or their intracellular partitioning (Cannon et al., 2009; Fluman and Bibi, 2009; Higgins, 2007; Seeger and van Veen, 2009). In yeast, several MDR transporters have been identified and studied thoroughly, particularly those belonging to the ATP-binding cassette (ABC) superfamily (Jungwirth and Kuchler, 2006; Rogers et al., 2001). A less well-

characterized group of transporters involved in MDR belongs to the major facilitator superfamily (MFS) (Paulsen et al., 1996; Sá-Correia and Tenreiro, 2002; Sá-Correia et al., 2009). Based on both phylogeny and function and according to the transport classification (TC) system, these transporters belong to two families: the 12 transmembrane spanner drug:H⁺ antiporter-1 (DHA1) family (TC #2.A.1.2.) and the 14 transmembrane spanner drug:H⁺ antiporter-2 (DHA2) family (TC #2.A.1.3.) (Saier et al., 2006). The majority of the members of these families were uncovered on release of the *Saccharomyces cerevisiae* genome sequence (Goffeau et al., 1996; Nelissen et al., 1995). According to the drug-pump model, the yeast MFS-MDR transporters characterized so far are thought to actively transport a wide variety of structurally and pharmacologically unrelated drugs and other xenobiotic compounds from the cytosol to the outer medium, thus providing protection from their deleterious action (Sá-Correia et al., 2009). However, this model is currently considered too simplistic, and

¹IBB—Institute for Biotechnology and Bioengineering, Centro de Engenharia Biológica e Química, Instituto Superior Técnico, Technical University of Lisbon, Lisboa, Portugal.

²Genetics, Reproduction, Population—Earth and Life Institute (ELI), Université catholique de Louvain, Louvain-la-Neuve, Belgium.

³Institut des Sciences de la Vie (ISV), Université catholique de Louvain, Louvain-la-Neuve, Belgium.

some evidence supports the notion that a number of these transporters might have a specific physiological substrate (Felder et al., 2002; Sá-Correia et al., 2009; Vargas et al., 2007; Velasco et al., 2004) and that their ability to export multiple drugs may occur fortuitously and be selected opportunistically. In *S. cerevisiae*, the DHA1 family comprises twelve proteins encoded by *QDR1*, *QDR2*, *AQR1*, *DTR1*, *QDR3*, *HOL1*, *FLR1*, *TPO1*, *TPO2*, *TPO3*, *TPO4*, and *YHK8* genes (Nelissen et al., 1997; Sá-Correia and Tenreiro, 2002; Sá-Correia et al., 2009). The expression of the majority of these permeases has been demonstrated to confer resistance to several structurally and functionally unrelated cytotoxic chemicals in *S. cerevisiae* (Sá-Correia and Tenreiro, 2002; Sá-Correia et al., 2009). Compared to the DHA1 family proteins, the biological function of the DHA2 family proteins is, in general, less well characterized (Sá-Correia et al., 2009) and will not be addressed in this study.

S. cerevisiae is classified in the subphylum Saccharomycotina (Hibbett et al., 2007), a taxonomic group more commonly known as Hemiascomycetes. Two major evolutionary events of the hemiascomycetous species characterize the following two clades: (1) the WGD clade, which includes species such as *S. cerevisiae* and *Candida glabrata* that have undergone Whole-Genome Duplication more than 100 million years ago (Wolfe and Shields, 1997); and (2) the CTG clade, which includes species such as *Candida albicans* and *Debaryomyces hansenii* that translate the CUG codon into serine instead of leucine (Dujon, 2006; Wang et al., 2009). A third group, Dipodascaceae, comprises *Yarrowia lipolytica*. The first evolutionary study dedicated to the DHA1 family was focused on 12 *S. cerevisiae* genes, which were grouped in three clusters (Nelissen et al., 1997). More recently, Gbelska et al. (2006) identified 51 DHA1 genes in five hemiascomycetous species and proposed to divide the DHA1 family in nine phylogenetic clusters.

However, the widely used phylogenetic analysis does not reveal the full history of a specific group of genes. The identification of short synteny blocks shared between species can provide a more accurate understanding of gene lineages (Dujon, 2005; Scannell et al., 2007). Gene lineage analysis, defined as "the examination of the descent of specific genes or gene complexes irrespective of their taxonomic affiliations" (Beckstrom-Sternberg et al., 1991), can be used to study both evolutionary pattern and process (Avice, 1989). The identification of gene lineages using a combined phylogenetic and syntenic approach endows the biologist with the capacity of scrutinizing in detail the evolutionary pathway of genes related by common descent (homologues) and to classify them as orthologues, genes related via speciation (vertical descent), or as paralogs, genes related via duplication (Fitch, 1970; Koonin, 2005). More recently, Wolfe (2000) suggested that paralogs arising from a WGD should be called ohnologs in honor of the late Susumu Ohno (1970). Ohnologs are an important legacy of WGD, representing the genes that did not return to single copy in the genome and that form a pool of genetic material from which new functions (neofunctionalization) or specialization of daughter genes (subfunctionalization) can evolve. Using both phylogenetic and syntenic information, two recent studies focusing on the evolution of pleiotropic drug resistance (PDR) transporters of the ABC superfamily (Seret et al., 2009) and of hexose transporters and glucose sensors (Palma et al., 2009) were able to

identify several gene lineages in hemiascomycetous yeasts and identify the orthologue/paralog status of the corresponding genes.

This study establishes a new evolutionary perspective on the hemiascomycetous DHA1 family based on a combined phylogenetic and gene neighborhood approach. For this, a total of 196 aminoacid sequences of DHA1 proteins were identified in the genome of 13 hemiascomycetous species. The nine hemiascomycete yeast genome sequences made available by the Génolevures project (Dujon, 2005; The Génolevures Consortium, 2009) were used to develop a database of the chromosomal neighbors of the DHA1 genes, taking advantage of the high-quality protein annotation of these genomes. The bioinformatics tool "Identification of Orthologues by Neighborhood and Similarity" (IONS) (Seret et al., unpublished data) was used to detect neighborhood evidence linking these DHA1 genes. YGOB database was used to provide neighborhood information on hemiascomycetous species that were not annotated by Génolevures. The synteny relationships between the hemiascomycetous DHA1 genes allowed the identification of 10 lineages encompassing an evolutionary history of about 300 to 400 million years, the time where the hemiascomycete yeasts are believed to have separated from the euascomycete fungi (Berbee and Taylor, 2001).

Materials and Methods

Protein sequences used in this analysis come from different databases: the one of the Génolevures Consortium, the Yeast Gene Order Browser (YGOB), the Saccharomyces Genome Database (SGD), and the Broad Institute.

Identification of the hemiascomycetous DHA1 protein set from the nine species present in the Génolevures database

Protein sequences from *S. cerevisiae*, *C. glabrata*, *Zygosaccharomyces rouxii*, *Kluyveromyces thermotolerans*, *Saccharomyces kluyveri*, *Kluyveromyces lactis*, *Eremothecium gossypii*, *D. hansenii*, and *Y. lipolytica* were retrieved from the Génolevures database in May 2009 (Table 1). Henceforth, the four letters code shown in this table for species abbreviation will be used to designate both yeast genes and species. A bioinformatics analysis based on BLAST and Smith-Waterman alignments, clustering and Condorcet election was applied to the Génolevures dataset by Sherman et al. (2009) in order to define consensus families. All *S. cerevisiae* DHA1 proteins were found to belong to the single GL3C0007 family, which comprises 143 members.

Identification of the hemiascomycetous DHA1 protein set from four other species

Because four hemiascomycetous species, *Kluyveromyces waltii*, *Kluyveromyces polysporus*, *Saccharomyces castellii*, and *Saccharomyces bayanus*, do not have a dedicated database providing high-quality annotation of their genomes, the corresponding DHA1 proteins were manually identified through the *in silico* pipeline described in Supplementary Data Figure S1. Briefly, the genomes of these hemiascomycete yeasts were retrieved from several databases, as indicated in Table 1, comprising a total of 47,090 translated open reading frames

TABLE 1. HEMIASCOMYCETES SPECIES WHOSE GENOMES WERE EXAMINED IN THE PRESENT WORK

Species	Acronym	Clade	Genome database	Number of full-size DHA1 transporters	Number of fragmented proteins
<i>Saccharomyces cerevisiae</i>	SACE	1	Génolevures and SGD ^a	12	0
<i>Saccharomyces bayanus</i>	SABA	1	YGOB and SGD ^b	12	1
<i>Saccharomyces castellii</i>	SACA	3	YGOB and SGD ^b	9	5
<i>Candida glabrata</i>	CAGL	4	Génolevures	10	0
<i>Kluyveromyces polysporus</i>	KLPO	6	YGOB	13	0
<i>Zygosaccharomyces rouxii</i>	ZYRO	7	Génolevures	21	0
<i>Saccharomyces kluyveri</i>	SAKL	10	Génolevures	13	0
<i>Kluyveromyces waltii</i>	KLWA	10	YGOB and Broad Institute	12	1
<i>Kluyveromyces thermotolerans</i>	KLTH	10	Génolevures	15	0
<i>Kluyveromyces lactis</i>	KLLA	11	Génolevures	7	0
<i>Eremothecium gossypii</i>	ERGO	12	Génolevures	8	0
<i>Debaryomyces hansenii</i>	DEHA	—	Génolevures	25	0
<i>Yarrowia lipolytica</i>	YALI	—	Génolevures	32	0
Total				189	7

Number of full-size DHA1 transporters and amino acid fragments identified from the corresponding genome databases are indicated. The clade classification of the hemiascomycete yeast was based on Kurtzman and Robnett (2004).

^aGénolevures, <http://www.genolevures.org/download.html>

^bSGD, http://downloads.yeastgenome.org/sequence/genomic_sequence/orf_protein/ http://downloads.yeastgenome.org/sequence/fungal_genomes/

YGOB, <http://wolfe.gen.tcd.ie/ygob/>

Broad Institute, <http://www.broadinstitute.org/science/projects/fungal-genome-initiative/current-fgi-sequence-projects>

(ORFs). The formatted function of the Blast NCBI package was used to convert these translated ORFs into a hemiascomycete blast database (HBD). Considering the GL3C0007 protein set as a reference standard for DHA1 identification, all members of this protein family were used as query sequences in a blastp search against the HBD. The blastp algorithm used gapped alignment with the following parameter values: expectation value (10^{-30}), open gap (−1), extend gap (−1), threshold for extending hits (11), and word size (3). In order to assess the existence of possible repeated sequences and of non-DHA1 sequences, each of the fragmented sequences identified in the blastp algorithm were scrutinized using (1) blastp against Génolevures database and (2) comparison with the remaining sequences identified by the blastp procedure.

Protein topology and length analysis

The topology of each of the amino acid sequences identified by the blastp algorithm was analyzed using the HMMTOP bioinformatics tool (Tusnády and Simon, 1998). For sequences not showing 12 TMS or showing less than 450 amino acids in length, protein topology was analyzed using TMHMM 2.0 (Sonnhammer et al., 1998), TOPPRED 2 (von Heijne, 1992), and MEMSAT 3 (Jones, 2007) bioinformatic tools as detailed in Supplementary Table S1.

Phylogenetic tree and pairwise alignments construction

Multiple alignments of the GL3C0007 proteins were calculated by MUSCLE (Edgar, 2004) and processed using the PHYLIP package (Felsenstein, 1989). In a first step, PROT-DIST and NEIGHBOR packages were used to generate the phylogenetic tree of DHA1 protein family. Because the PROTML package originated a phylogenetic tree with a

similar composition of the different clusters despite some differences in the between-cluster organization, only neighbor joining (NJ) results are presented. The DENDROSCOPE application (Huson et al., 2007) was used to visualize this tree. The sequence identity and similarity shared between pairs of DHA1 proteins was assessed using an all-against-all Needleman-Wunsch alignment approach. This algorithm was run using the NEEDLE package available in the EMBOSS suite (Rice et al., 2000). All needle pairwise alignments made in this work used default values for the gap open and gap extension parameters, 10.0 and 0.5, respectively. It required the development of a diverse set of linux shell scripting. After the chromosome neighborhood analysis of the genes encoding GL3C0007 proteins, all-against-all Needleman-Wunsch alignments were constructed for the members of each DHA1 gene lineage.

Identification of synteny between GL3C0007 family members

The GL3C0007 protein family was analyzed using Identification of Orthologues by Neighborhood and Similarity (IONS) (Seret et al., unpublished). Fifteen neighbor genes were retrieved on each side of the 143 DHA1 encoding genes ("query genes"). The rationale of the IONS method is that two genes of different yeast species whose translation products belong to the same family (homologues by similarity) will be members of the same "Subset of Orthologues by Neighborhood and Similarity" (SONS) if they share at least one pair of neighbors that are also homologous to each other by similarity. The process is reiterated for all possible heterospecific pairwise comparisons of homologues deduced from the protein families. Homologues that do not share a pair of homologous neighbors are divided into two distinct SONS.

Due to the complexity of the DHA1 evolutionary history, the SONS needed manual curation to remove some weak links between chromosomal environments (Supplementary Data Fig. S2). The manual curation procedure involved the conversion of the GL3C0007 neighborhood data into a network. The advantage of this framework is that it allows to scrutinize the synteny relationships established between genes in a simple mathematical context of network topology exploration. Three sources of biological information were used to assess the strength of each neighbor pair connection. (1) Following the rationale of the automated IONS tool, the manual curation procedure valued the closeness of connecting neighbors in relation to the DHA1 query genes. (2) The manual curation also valued the sharing of high sequence similarity between connecting neighbors. The use of sequence percentage identity between two connecting neighbor genes is bounded to be heterogeneous because protein families are not equally conserved. In fact, due to structural reasons, there are proteins that are less variant in evolution than others (i.e., histone or ribosome proteins against kinase proteins, respectively). This criterion demanded the calculation of the sequence identity and similarity percentages between all neighbor pairs linking the DHA1 query genes using Needleman-Wunsch pairwise alignments. Considering that family assignment is based on similarity, the assessment of the sequence identity can be viewed as the adoption of stronger similarity criteria removing some weak links. (3) Using the relative frequency of each protein family in the Génolevures database as the binomial probability of success, the number of members (dimension) of any given protein family comprised in the database of chromosome neighborhood of the GL3C0007 genes was used to assess the significance of each connecting neighbors. The synteny information concerning *K. waltii*, *K. polysporus*, *S. castellii*, and *S. bayanus* species was retrieved from YGOB database (Byrne and Wolfe, 2005).

Results

GL3C0007 protein family: phylogeny and identification of DHA1 gene lineages

A phylogenetic tree was constructed for the 189 non-fragmented DHA1 proteins (Fig. 1). These were distributed in 20 clusters, labeled from A to T, 9 of which were also marked by their *S. cerevisiae* members. Clusters were visually defined from NJ tree topology. Ex post analysis indicates that the average within cluster similarity is always above a 30% threshold and for the 10 largest clusters above 44%. Cluster A comprises the tree root, the DEHA2A14586 (DEHA_13) gene.

The bioinformatics tool developed for Identification of Orthologues by Neighborhood and Similarity, IONS, was used to reconstruct the evolutionary history of each GL3C0007 family member. The automated IONS algorithm identified the synteny links established between DHA1 genes. Manual curation of the neighborhood data made in the CYTOSCAPE environment improved the identification of DHA1 gene lineages when paralogs shared few neighbor pairs. In the phylogenetic tree of Figure 1, the DHA1 genes sharing common neighbors were labeled with the same color. Ten main DHA1 gene lineages, spanning the whole hemiascomycetes clade, plus five other lineages, spanning only two hemiascomycetous species, were identified. Ten gene lineages are described within the nine DHA1 phylogenetic clusters: cluster

E (Dtr1p), cluster F (Qdr1p, Qdr2p, Aqr1p), cluster G (Qdr3p), cluster J (Hol1p), cluster N (Tpo2p, Tpo3p), cluster P (Tpo1p), cluster R (Yhk8p), cluster S (Tpo4p), and cluster T (Flr1p). Cluster N (Tpo2p, Tpo3p) comprises two independent DHA1 gene lineages. Genes belonging to five clusters could not be assigned to lineages. For species not comprised in Génolevures, namely, *K. waltii*, *K. polysporus*, *S. castellii*, and *S. bayanus*, neighborhood data were retrieved from YGOB database. The *in silico* procedure used to identify the DHA1 protein set of these four poorly annotated genomes retrieved a total of 52 DHA1 amino acid sequences (Table 1 and Supplementary Data Fig. S1). Of these, 46 were considered to be non-fragmented sequences, whereas 12 were considered to be fragmented ones. Combining the NEEDLE package with linux shell scripting, we carried out 35,721 pairwise alignments comprising all possible combinations between the DHA1 proteins. In average, the DHA1 proteins share 22.2% (SD 11.1%) of sequence identity and 36.2% (SD 11.9%) of sequence similarity. The identity between proteins included in each of the 10 DHA1 gene lineages ranges from 40.3 to 69.4%, whereas sequence similarity ranges from 55.6 to 79.9%. Because data on chromosome neighborhood of species from the CTG clade is scarce and is only available for *Debaryomyces hansenii*, the combined phylogeny and neighborhood analysis of this Saccharomycetaceae branch was impaired. Although we have focused on the Saccharomycetaceae family, a Dipodascaceae family member, *Yarrowia lipolytica*, was also included in the neighborhood analysis.

Description of the DHA1 gene lineages

Lineage 1 (DTR1) is composed of 11 genes, spanning from *E. gossypii* to *S. cerevisiae* species (Figs. 2 and 3). Contrasting with the remaining post-WGD species that have retained the DTR1 orthologue, *K. polysporus* may have retained the DTR1 ohnolog as suggested by the YGOB. The DTR1 lineage is the only lineage that has no member in the *Y. lipolytica* species and whose evolutionary origin is still unclear.

Lineage 2 (QDR1, QDR2, AQR1) is composed of 25 genes, spanning all species from *Y. lipolytica* to *S. cerevisiae* (Figs. 2 and 3). QDR1 and QDR2 genes constitute a tandem repeat. The WGD event originated two independent gene lineages that culminate in the SACE QDR1/QDR2 and AQR1 ohnologs. All post-WGD species subsequent to the speciation of *C. glabrata* show QDR1 and QDR2 orthologues in tandem repeat.

Lineage 3 (QDR3) is composed of 10 genes and contains a gap comprising *K. polysporus*, *C. glabrata*, and *S. castellii* species which, independently, have lost their QDR3 orthologue (Figs. 2 and 3). We have no trace of any ohnolog, postulating a rapid loss after the WGD.

Lineage 4 (HOL1) is composed of seven genes, belonging to the *E. gossypii*, *K. lactis*, *K. thermotolerans*, *K. waltii*, and *S. kluyveri* species (Figs. 2 and 3), sharing highly conserved chromosome neighborhood. This lineage presents a gap between pre- and post-WGD species. In fact, four successive "Saccharomyces complex" species lack HOL1 homologues and, most importantly, the HOL1 homologues of the *Saccharomyces sensu stricto* species locate in a different chromosome environment, compared with those of pre-WGD species. Two hypotheses can explain this evolutionary pattern: the occurrence of at least five consecutive and independent gene loss events or a horizontal gene transfer between a pre-WGD

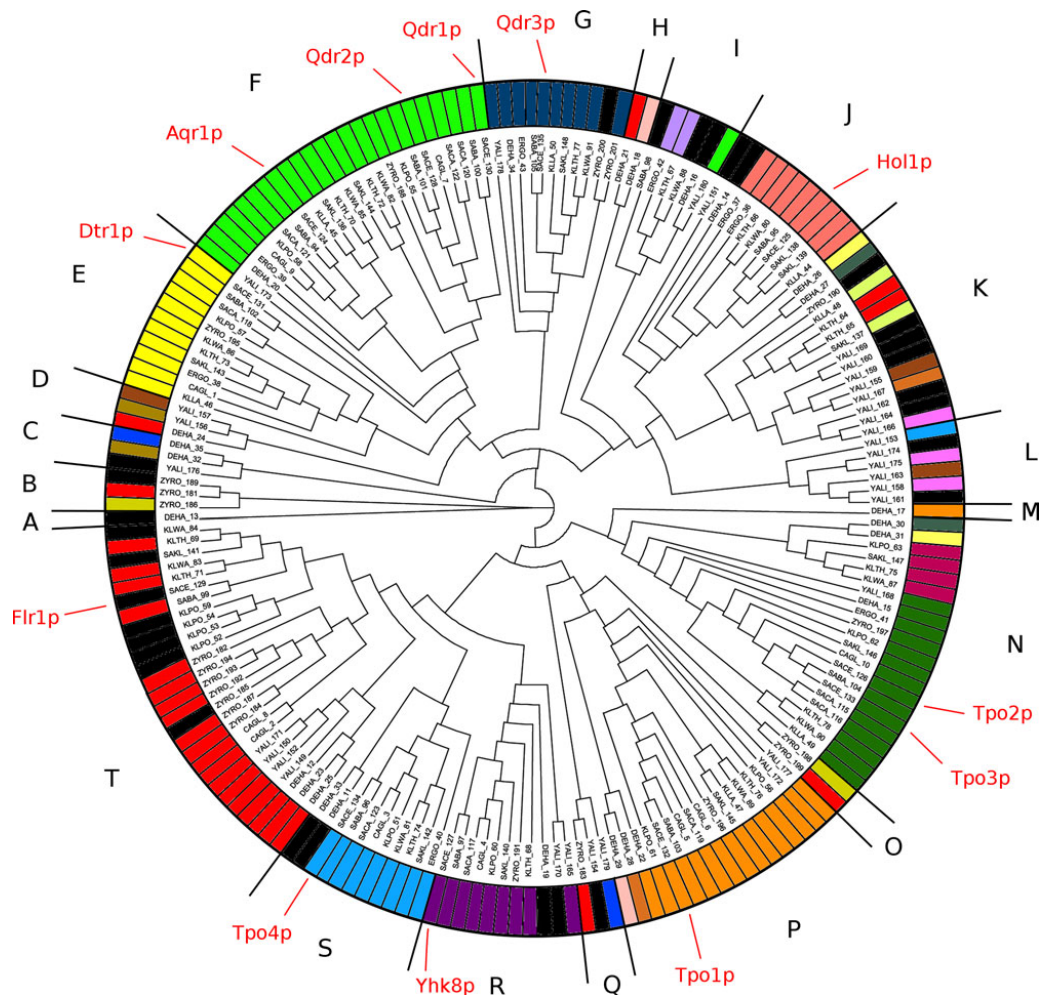


FIG. 1. Phylogenetic tree of the DHA1 protein set of 13 hemiascomycete yeasts. All *S. cerevisiae* DHA1 transporters were found to belong to a single Génolevures protein family, the GL3C0007. The tree shows 20 clusters identified by capital letters as well as by the name of their *S. cerevisiae* members. The colors identify genes belonging to the same lineage. Singleton genes are indicated in black color. Table S1 provides the conversion to the real names of each of these DHA1 proteins.

species and the common ancestor of the *Saccharomyces sensu stricto* group.

Cluster N (Tpo2p, Tpo3p) comprises two independent DHA1 gene lineages, 5 and 6. Lineage 5 (*TPO2*, *TPO3*) is composed of 16 genes (Figs. 2 and 3). *TPO2* and *TPO3* are ohnologs and, with the exception of *K. polysporus* and *C. glabrata*, all post-WGD species have both duplicates. *K. polysporus* species has retained the *TPO3* orthologue while the *C. glabrata* species has retained the *TPO2* orthologue.

Lineage 6 is composed of only four genes, KLTH0B03894g (KLTH_75), orf kwl_SNP:kwl_208-snp.3 (KLWA_87), SAKL0G13376g (SAKL_147), and orf Kpo_2002.86 (KLPO_63), which however, show strong neighborhood conservation

(Figs. 2 and 3). The fact that cluster N comprises two independent DHA1 gene lineages and that lineage 6 is absent in the post-WGD species may suggest that the presence of the two *TPO2* and *TPO3* ohnologs in lineage 5 of the same cluster made redundant the lineage 6 after the WGD.

Lineage 7 (*TPO1*) is composed of 15 genes (Figs. 2 and 3), spanning from YALI0F07062 (YALI_172) to *S. cerevisiae* *TPO1*. Although *K. polysporus* and *C. glabrata* have retained both duplicates produced in the WGD event, all late divergent species have lost one. The duplicate present in the *S. castellii* genome is ohnolog of *S. cerevisiae* *TPO1* while all subsequent species have retained *TPO1* orthologues. The DEHA2C00506 (DEHA_17) gene does not belong to cluster P as it is the single

member of cluster M but shares common neighborhood features with YALIOF07062 (YALI_172), DEHA2C15466 (DEHA_22), and KLLA0F18106 (KLLA_47) genes of cluster P. In consequence, DEHA2C00506 (DEHA_17) gene was assigned to lineage 7.

Lineage 8 (YHK8) is composed of nine genes (Figs. 2 and 3). All post-WGD species have lost one YHK8 homologue. *K. polysporus* and *C. glabrata* species retained different ohno-

logs and all subsequent post-WGD species inherited the *K. polysporus* YHK8 orthologue.

Lineage 9 (TPO4) comprises 10 genes (Figs. 2 and 3) whose chromosome environment is conserved from *E. gossypii* to *S. cerevisiae* species. All post-WGD species have lost one duplicate and the YGOB database indicates that the *K. polysporus* species have retained the TPO4 ohnolog. The neighborhood data allowed to identify YALIOD20064 (YALI_153), a gene

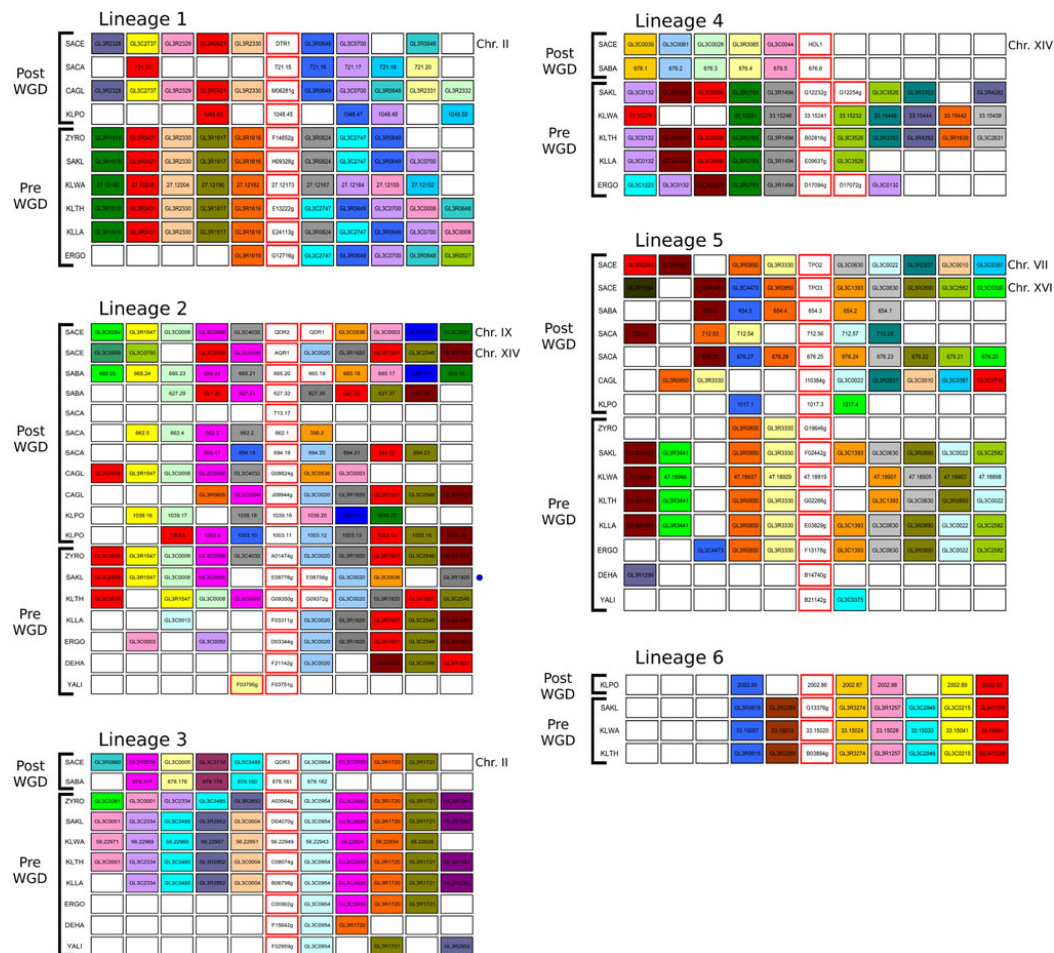


FIG. 2. Manually curated SONS of DHA1 encoding genes. Lineage 1: orthologues of *DTR1* (cluster E). Lineage 2: orthologues of *QDR1*, *QDR2*, and *AQR1* (cluster F). Lineage 3: orthologues of *QDR3* (cluster G). Lineage 4: orthologues of *HOL1* (cluster J). Lineage 5: orthologues of *TPO2* and *TPO3* (subcluster N1). Lineage 6: orthologues of subcluster N2. Lineage 7: orthologues of *TPO1* (cluster P). Lineage 8: orthologues of *YHK8* (cluster R). Lineage 9: orthologues of *TPO4* (cluster S). Lineage 10: orthologues of *FLR1* (cluster T). Each box framed in red represents a DHA1 gene. Adjacent boxes represent the DHA1 gene neighbors. Yellow background represents DHA1 genes not belonging to the phylogenetic cluster associated to the lineage. Homologous neighbors, based on Génolevures families, are highlighted in the same color. A white box represent genes with no homologous neighbors in the neighborhood. The SONS were calculated with 15 neighbors on each side, but this representation was truncated to 5 neighbors for the sake of clarity. A blue dot in the margin is the sign that the corresponding line was issued from a manual curation. Neighborhood data concerning *K. waltii*, *K. polysporus*, *S. castellii*, and *S. bayanus* were retrieved from YGOB database.

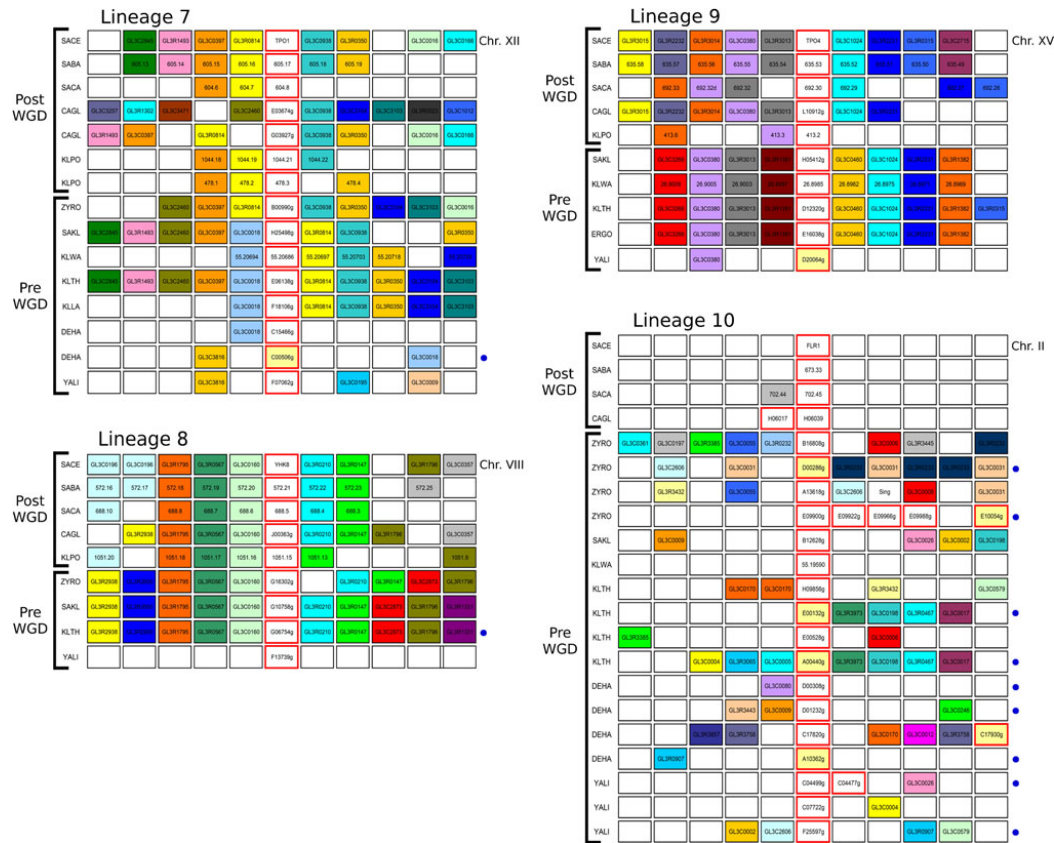


FIG. 2. Continued.

classified in cluster L, as an orthologue of TPO4 as it shares four neighbor pairs with both KLTH0D12320 (KLTH_74) and SAKL0H05412 (SAKL_142) genes while considering 15 neighbors on each side.

Lineage 10 (*FLR1*) is composed of 25 genes (Figs. 2 and 3), all linked by neighbor pairs, although the conservation of the corresponding chromosome environment is weak. This lineage does not exhibit the normal WGD gene pattern, where the ancestral gene originates two ohnolog genes, that can be either subsequently lost or not. The absence of neighborhood conservation of *FLR1* homologues belonging to successive post-WGD species suggests that multiple “expulsion” events (transpositions?) have occurred. Interestingly, the *Z. rouxii* species shows dispersive duplication of seven *FLR1* homologues throughout chromosomes A, B, E, and F (Fig. 3). The neighborhood data shows that ZYRO0B16808 (ZYRO_187) is an orthologue of KLTH0E0528 (KLTH_71) and Scas_702.45 (SACA_114). Moreover, it is likely that ZYRO0B16808 (ZYRO_187) was at the origin of a series of gene duplication events, including a subtelomeric tandem repeat constituted by four genes. Although most of the *FLR1* homologues have retained their cluster T classification, the surplus of genes

observed in this lineage might have decreased the selective pressure over the remaining copies, creating a reservoir for the creation of functional novelty. This hypothetical scenario is consistent with the identification in lineage 10 (*FLR1*) of seven cases of gene duplication (highlighted in yellow on Figs. 2 and 3) followed by sequence divergence of one paralog gene into other phylogenetic cluster. These cases are strongly associated with the subtelomeric localization of the corresponding *FLR1* homologues, in agreement with the properties of subtelomeric genes, thought to be involved in adaptation to the environment and to promote *de novo* gene creation by recombination inside gene repeats and between different chromosome arms (Fairhead and Dujon, 2006).

Discussion

Aiming to obtain a wider perspective on the evolution of the DHA1 protein family in the hemiascomycete yeasts, we have considered in this study eight hemiascomycetous species additional to those examined by Gbelska et al. (2006). These additional species enabled us to follow more accurately the evolution of the chromosomal environments of the DHA1

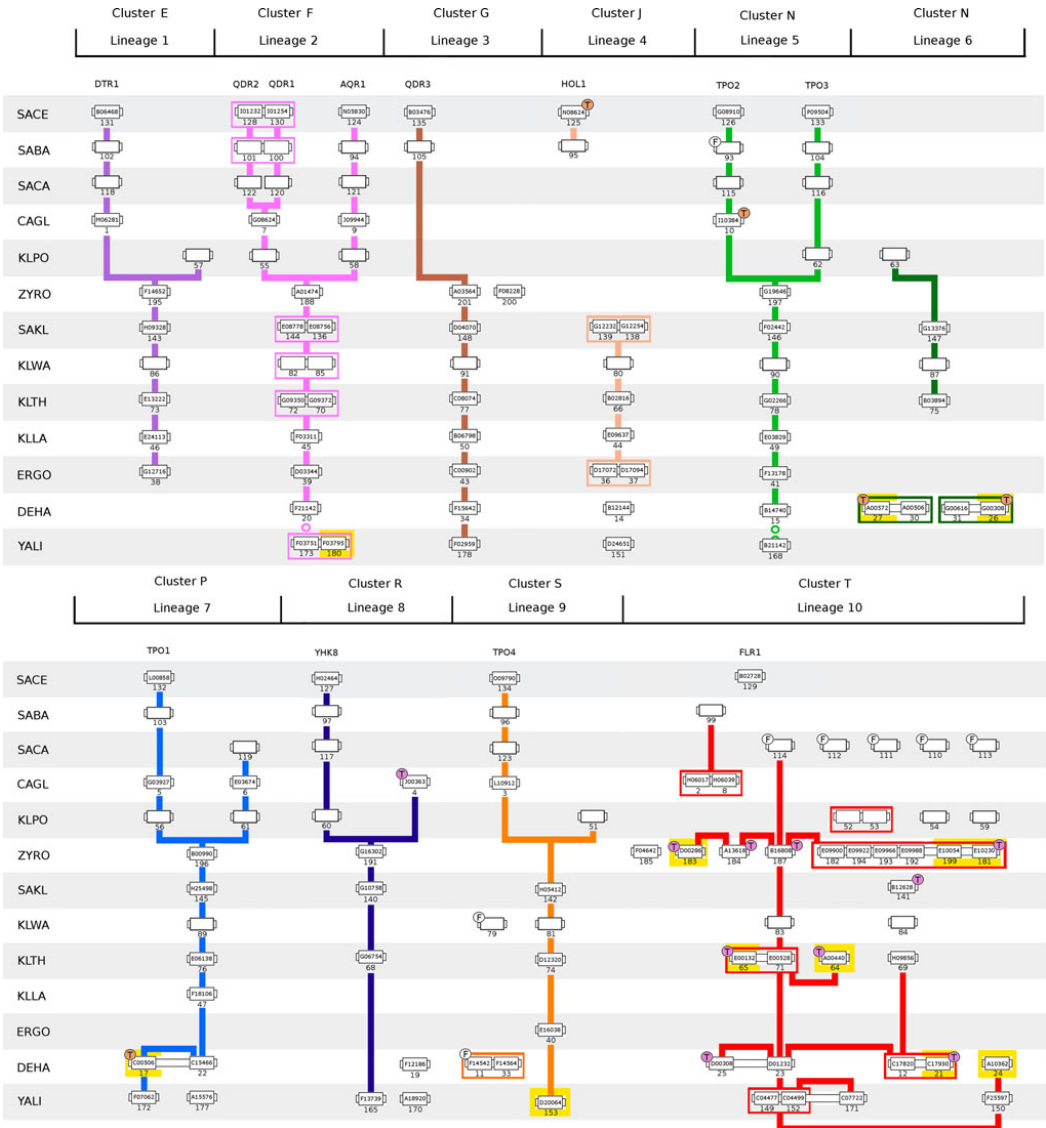


FIG. 3. The 10 main lineages of DHA1 transporters. Each box represents a DHA1 gene identified by a number and, if available, by its name according to Génolevures database. Genes sharing neighborhoods are connected by colored lines. Dotted connectors link genes whose relationship is based on phylogenetic evidence only. Yellow background represents DHA1 genes not belonging to the phylogenetic cluster associated to the lineage. F indicates that the corresponding gene was classified as a DHA1 fragment. T identify that the corresponding gene locates in a subtelomeric region (right or left position indicates the chromosome extremity where the telomer locates). T on a rose background indicates that the corresponding gene distances less than 15 neighbors from extremity while T on an orange background indicates a distance of 16 to 30 neighbors from extremity. Neighborhood data concerning *K. waltii*, *K. polysporus*, *S. castellii*, and *S. bayanus* was retrieved from YGOB database.

genes and to distinguish between orthologous, paralogous, and ohnologous status. The 20 cluster classification of the DHA1 protein family proposed in our study expands considerably the classification proposed by Gbelska et al. (2006). The pairwise comparisons between the DHA1 hemi-

ascmycetes amino acid sequences also showed that the DHA1 gene lineages are homogenous: the 189 nonfragmented DHA1 proteins share, on average, 22.2% of their amino acidic residues, whereas the members of each gene lineage share, on average, 40.3 to 69.4% of sequence identity. Although a few

DHA1 family transporters have been described as recognizing substrates of biological significance (Felder et al., 2002; Vargas et al., 2007; Velasco et al., 2004), most of those known to be required for resistance to several drug and other xenobiotic compounds do not have an assigned biological role yet. The amino acid conservation shown by the phylogenetic clusters suggests that each of these clusters may comprise genes with a specific biological function in common.

Clusters N (Tpo2p/Tpo3p) and F (Qdr1p/Qdr2p/Aqr1p) are the only clusters that comprise genes from all the 13 hemiascomycete yeasts considered in this study, suggesting that the DHA1 proteins belonging to these clusters may play an important biological function in hemiascomycetous species and that their absence would decrease yeast fitness.

Gene duplication is the primary source of new genes and provides essential raw material for the evolution of functional novelty (Ohno, 1970; Taylor and Raes, 2004). Upon duplication, the two gene copies are redundant and, therefore, vulnerable to an inactivating mutation. In order to persist functionally, these gene copies must lose redundancy either through modification of function or, at least one of the duplicates must gain a new function. The occurrence of multiple gene duplication and loss events throughout the entire evolutionary history of hemiascomycete yeasts was recently proposed for two transport protein families: the PDR transporters of the ABC superfamily (Seret et al., 2009) and the hexose transporters (Palma et al., 2009). The synteny and similarity data uncovered in the present study show that gene duplication and gene loss were also major driving forces in the evolution of the DHA1 transporters. The analysis of the GL3C0007 protein family showed that 57 of its 143 members (40%) were originated in gene duplication events. Of these, 28 duplicates (14 pair, 20%) do not have the same cluster classification, indicating that at least one of these duplicated genes (10%) has sufficiently differentiated in sequence to be assigned to a different phylogenetic cluster, potentially providing functional novelty to the corresponding hemiascomycetous species.

The DHA1 encoding genes of the main lineages showed a remarkable conservation of their chromosome environment as well as in their amino acid sequence. In fact, with the exception of *DTR1* and *TPO4*, all *S. cerevisiae* DHA1 genes issued from gene lineages whose origins roots, at least, to the early-diverging hemiascomycete *Y. lipolytica*. Nevertheless, considering that five contemporary DHA1 gene lineages still maintain the propensity to undergo gene duplication and the significant level of identity shared by DHA1 members, we suggest that this protein family had a common ancestral origin, from a first primordial DHA1 gene. Based on this assumption it is hypothesized that many phylogenetic clusters identified in this study were originated in duplication events followed by sequence divergence of the resulting DHA1 genes. A similar study to the one carried out in this article by combining phylogeny and neighborhood analysis on well-annotated genomes will be required to identify the DHA1 gene lineages in the other CTG clade species.

Acknowledgments

This research was supported by FEDER and "Fundação para a Ciência e Tecnologia" (contracts PTDC/BIO/72063/2006 and PTDC/BIA-MIC/72577/2006) and PhD scholarship (SFRH/BD/23331/2005) to P.J. Dias. We thank the Genole-

vures consortium, coordinated by Jean-Luc Souciet, for access to the database of protein sequences from *Zygosaccharomyces rouxii*, *Kluyveromyces thermotolerans*, *Saccharomyces kluyveri*, and *Eremothecium gossypii*. We thank two anonymous reviewers for helpful comments.

Author Disclosure Statement

The authors declare that no conflicting financial interests exist.

References

- Avise, J.C. (1989). Gene trees and organismal histories: a phylogenetic approach to population biology. *Evolution* 43, 1192–1208.
- Beckstrom-Sternberg, S.M., Rieseberg, L.H., and Doan, K. (1991). Gene lineage analysis of populations of *Helianthus niveus* and *H. petiolaris*. *Plant System Evolut* 175, 125–138.
- Berbee, M.L., and Taylor, J.W. (2001). Fungal molecular evolution: gene trees and geologic time. In: *The Mycota, Vol. VIIIB, Systematics and Evolution*. D.J. McLaughlin, E. McLaughlin, and P.A. Lemke, eds. (Springer, Berlin), pp. 229–246.
- Byrne, K.P., and Wolfe, K.H. (2005). The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res* 15, 1456–1461.
- Cannon, R.D., Lamping, E., Holmes, A.R., Niimi, K., Baret, P.V., Keniya, M.V., et al. (2009). Efflux-mediated antifungal drug resistance. *Clin Microbiol Rev* 22, 291–321.
- Dujon, B. (2005). Hemiascomycetous yeasts at the forefront of comparative genomics. *Curr Opin Genet Dev* 15, 614–620.
- Dujon, B. (2006). Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. *Trends Genet* 22, 375–387.
- Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32, 1792–1797.
- Fairhead, C., and Dujon, B. (2006). Structure of *Kluyveromyces lactis* subtelomeres: duplications and gene content. *FEMS Yeast Res* 6, 428–441.
- Felder, T., Bogengruber, E., Tenreiro, S., Ellinger, A., Sá-Correia, I., and Briza, P. (2002). Dtr1p, a multidrug resistance transporter of the major facilitator superfamily, plays an essential role in spore wall maturation in *Saccharomyces cerevisiae*. *Eukaryot Cell* 1, 799–810.
- Felsenstein, J. (1989). PHYLIP—Phylogeny Inference Package (Version 3.2). *Cladistics* 5, 164–166.
- Fitch, W.M. (1970). Distinguishing homologous from analogous proteins. *Syst Zool* 19, 99–113.
- Fluman, N., and Bibi, E. (2009). Bacterial multidrug transport through the lens of the major facilitator superfamily. *Biochim Biophys Acta* 1794, 738–747.
- Gbelska, Y., Krijger, J.J., and Breunig, K.D. (2006). Evolution of gene families: the multidrug resistance transporter genes in five related yeast species. *FEMS Yeast Res* 6, 345–355.
- Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., et al. (1996). Life with 6000 genes. *Science* 274, 546, 563–567.
- Hayes, J.D., and Wolf, C.R. (1997). *Molecular Genetics of Drug Resistance*. (Harwood Academic, Newark, NJ).
- Hibbett, D.S., Binder, M., Bischoff, J.F., Blackwell, M., Cannon, P.F., Eriksson, O.E., et al. (2007). A higher-level phylogenetic classification of the fungi. *Mycol Res* 111, 509–547.
- Higgins, C.F. (2007). Multiple molecular mechanisms for multidrug resistance transporters. *Nature* 446, 749–757.

- Huson, D.H., Richter, D.C., Rausch, C., DeZulian, T., Franz, M., and Rupp, R. (2007). Dendroscope: an interactive viewer for large phylogenetic trees. *BMC Bioinformatics* 8, 460.
- Jones, D.T. (2007). Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics* 23, 538–544.
- Jungwirth, H., and Kuchler, K. (2006). Yeast ABC transporters—a tale of sex, stress, drugs and aging. *FEBS Lett* 580, 1131–1138.
- Koonin, E.V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* 39, 309–338.
- Kurtzman, C.P., and Robnett, C.J. (2003). Phylogenetic relationships among yeasts of the “*Saccharomyces* complex” determined from multigene sequence analyses. *FEMS Yeast Res* 3, 417–432.
- Nelissen, B., Mordant, P., Jonniaux, J.L., De Wachter, R., and Goffeau, A. (1995). Phylogenetic classification of the major superfamily of membrane transport facilitators, as deduced from yeast genome sequencing. *FEBS Lett* 377, 232–236.
- Nelissen, B., De Wachter, R., and Goffeau, A. (1997). Classification of all putative permeases and other membrane plurispansers of the major facilitator superfamily encoded by the complete genome of *Saccharomyces cerevisiae*. *FEMS Microbiol Rev* 21, 113–134.
- Ohno, S. (1970). *Evolution by Gene Duplication*. (Springer, New York).
- Palma, M., Seret, M.L., and Baret, P.V. (2009). Combined phylogenetic and neighbourhood analysis of the hexose transporters and glucose sensors in yeasts. *FEMS Yeast Res* 9, 526–534.
- Paulsen, I.T., Brown, M.H., and Skurray, R.A. (1996). Proton-dependent multidrug efflux systems. *Microbiol Rev* 60, 575–608.
- Rice, P., Longden, I., and Bleasby, A. (2000). EMBOS: the European Molecular Biology Open Software Suite. *Trends Genet* 16, 276–277.
- Rogers, B., Decottignies, A., Kolaczowski, M., Carvajal, E., Balzi, E., and Goffeau, A. (2001). The pleiotropic drug ABC transporters from *Saccharomyces cerevisiae*. *J Mol Microbiol Biotechnol* 3, 207–214.
- Sá-Correia, I., and Tenreiro, S. (2002). The multidrug resistance transporters of the major facilitator superfamily, 6 years after disclosure of *Saccharomyces cerevisiae* genome sequence. *J Biotechnol* 98, 215–226.
- Sá-Correia, I., Dos Santos, S.C., Teixeira, M.C., Cabrito, T.R., and Mira, N.P. (2009). Drug:H⁺ antiporters in chemical stress response in yeast. *Trends Microbiol* 17, 22–31.
- Saier, M.H., JR., Tran, C.V., and Barabote, R.D. (2006). TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Res* 34, D181–D186.
- Scannell, D.R., Frank, A.C., Conant, G.C., Byrne, K.P., Woolfit, M., and Wolfe, K.H. (2007). Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. *Proc Natl Acad Sci USA* 104, 8397–8402.
- Seeger, M.A., and Van Veen, H.W. (2009). Molecular basis of multidrug transport by ABC transporters. *Biochim Biophys Acta* 1794, 725–737.
- Seret, M.L., Diffels, J.F., Goffeau, A., and Baret, P.V. (2009). Combined phylogeny and neighborhood analysis of the evolution of the ABC transporters conferring multiple drug resistance in hemiascomycete yeasts. *BMC Genomics* 10, 459.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13, 2498–2504.
- Sherman, D.J., Martin, T., Nikolski, M., Cayla, C., Souciet, J.L., and Durrens, P. (2009). Génolevures: protein families and synteny among complete hemiascomycetous yeast proteomes and genomes. *Nucleic Acids Res* 37, D550–D554.
- Sonnhammer, E.L., Von Heijne, G., and Krogh, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol* 6, 175–182.
- Taylor, J.S., and Raes, J. (2004). Duplication and divergence: the evolution of new genes and old ideas. *Annu Rev Genet* 38, 615–643.
- The Génolevures Consortium. (2009). Comparative genomics of protoploid Saccharomycetaceae. *Genome Res* 19, 1696–1709.
- Tusnády, G.E., and Simon, I. (1998). Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J Mol Biol* 283, 489–506.
- Vargas, R.C., Garcia-Salcedo, R., Tenreiro, S., Teixeira, M.C., Fernandes, A.R., Ramos, J., et al. (2007). *Saccharomyces cerevisiae* multidrug resistance transporter Qdr2 is implicated in potassium uptake, providing a physiological advantage to quinidine-stressed cells. *Eukaryot Cell* 6, 134–142.
- Velasco, I., Tenreiro, S., Calderon, I.L., and Andre, B. (2004). *Saccharomyces cerevisiae* Aqr1 is an internal-membrane transporter involved in excretion of amino acids. *Eukaryot Cell* 3, 1492–1503.
- von Heijne, G. (1992). Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J Mol Biol* 225, 487–494.
- Wang, H., Xu, Z., Gao, L., and Hao, B. (2009). A fungal phylogeny based on 82 complete genomes using the composition vector method. *BMC Evol Biol* 9, 195.
- Wolfe, K.H., and Shields, D.C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387, 708–713.
- Wolfe, K. (2000). Robustness—it’s not where you think it is. *Nat Genet* 25, 3–4.

Address correspondence to:

Philippe V. Baret
Earth and Life Institute
Université catholique de Louvain
Croix du Sud 2 bte 14
1348 Louvain-la-Neuve, Belgium

E-mail: philippe.baret@uclouvain.be

