UNIVERSITE PARIS-SUD XI

Ecole doctorale : Gènes, Génomes, Cellules

THESE

Présentée pour obtenir le grade de

Docteur en Sciences

De l'Université Paris-Sud XI Orsay
Spécialité : Biologie

Par

Guillaume Morel

La levure G*eotrichum candidum* : taxonomie, biodiversité et génome

Soutenue le 20 décembre 2012
Devant le jury composé de :

Pr. Cécile FAIRHEAD,	UMR8621, IGM Université Paris Sud XI	**Présidente du jury**
Dr. Sylvie DEQUIN,	UMR1083-SPO, INRA	**Rapporteur**
Pr. Jean Luc SOUCIET,	UMR7156 Université Louis Pasteur Strasbourg	**Rapporteur**
Pr. Claude GAILLARDIN,	UMR1319 Institut MICALIS, INRA	**Examinateur**
Dr. Joëlle REITZ-AUSSEUR,	Laboratoire SOREDAB	**Examinateur**
Dr. Serge CASAREGOLA,	UMR1319 Institut MICALIS, INRA,	**Directeur de thèse**

AgroParisTech INRA UMR1319
MICALIS – CIRM Levures
Avenue Lucien Brétignières - 78850 Thiverval Grignon

REMERCIEMENTS

Je souhaite tout d'abord remercier le Dr ***Sylvie Dequin****, le Pr* ***Cécile Fairhead****, le Pr* ***Jean-Luc Souciet*** *et le Pr* ***Claude Gaillardin*** *pour m'avoir fait l'honneur de participer à mon jury de thèse et d'avoir accepté d'évaluer mon travail.*

Je tiens à remercier tout particulièrement ***Sylvie Dequin*** *et* ***Jean Luc Souciet*** *d'avoir accepté d'être mes rapporteurs*

Je voudrais ensuite remercier les Dr ***Stéphane Aymerich*** *et Dr* ***Jean-Marie Beckerich*** *pour m'avoir accueilli dans leur unité au sein de laquelle j'ai découvert un environnement de travail de qualité.*

Je tiens à remercier mon directeur de thèse, le Dr ***Serge Casaregola****, pour m'avoir confié ce travail de recherche, ainsi que pour sa disponibilité, son aide et ses précieux conseils au cours de ces années. Merci pour avoir été tout simplement humain dans les moments les plus difficiles.*

Je souhaite remercier vivement Mme ***Choreh Farokh****, chef de service de la direction scientifique du CNIEL, entreprise qui a soutenu financièrement cette thèse CIFRE.*

Les travaux réalisés dans le cadre de ma thèse ont été intégrés au projet ANR «Food-Microbiomes », merci aux Dr ***Pierre Renault****,* ***Joëlle Dupont****,* ***Georges Barbier,*** *pour avoir suivi mes travaux lors des comités de pilotage. Merci à* ***Jeanne*** *et* ***Antoine,*** *collègues thésards du projet, avec qui j'ai eu le plaisir de parler et d'être formé à l'utilisation d'Eugène au cours d'une semaine frappée par la neige.*

Merci à ***Jonathan Kreplak*** *pour m'avoir fourni les éléments nécessaires au bon fonctionnement d'Eugène sur la plateforme de l'URGI.*

Merci au consortium d'annotation de Geotrichum candidum, particulièrement ***Dominique Swennen*** *et* ***Djamila Onésime*** *vos efforts pour notre but commun me touchent énormément.*

J'ai bien évidemment beaucoup de personnes à remercier, pas uniquement pour leurs compétences ou leur implication dans ma thèse, mais tout simplement pour leur soutien, leur gentillesse ou le souvenir impérissable qu'elles m'ont laissé.

Merci aux filles du CIRM, pour leur aide et leur amitié. Merci à ***Noémie****,* ***Christelle*** *et particulièrement* ***Sandrine*** *pour cette collaboration. Travailler à vos côtés a été vraiment agréable.*

*Merci à **Fatima**, ce fut un plaisir de t'encadrer durant ton stage de MASTER2, merci pour cette collaboration fructueuse, pendant que l'annotation et Eugène me donnaient du fil à retordre.*

*Merci à **Stéphane Tribouillet**, tu as toujours été présent et de bon conseil lorsque je débutais sur UNIX*

*Un merci plein d'affection à **Anne** et **Guy** pour votre bonne humeur et votre soutien, sans vous le labo fonctionnerait beaucoup moins bien…**Guy**, tu passeras à la maison pour gouter un petit Bourgogne de derrière les fagots*

*Merci à tous les membres de MICALIS grignonais, être à vos côté a été un véritable plaisir. Merci à **Eliane** pour son oreille attentive et sa bienveillance, merci à **Mathieu**, ton rire de l'autre côté du bâtiment résonne comme une bande son de ces années. Merci à **Brigitte** pour ses discussions dans la navette et le train Grignon-Montparnasse, ces moments m'ont permis d'être aussi un peu en vacances lorsque l'on mentionnait les dentelles de Montmirail et les beautés de la Drôme Provençale.*

*Bon courage à **Olivier**, on a vraiment formé une bonne équipe de « thétards », Bon courage pour la suite et je te donne rendez-vous pour faire une bonne partie…*

*Merci à mes amis, **Yvain**, **Thibaut**, **Fabien**, **Charlotte**, **Alex**, **Morgan** et tous les autres pour votre soutien. Merci d'avoir compris mes absences répétées et mes longs moments de silence radio pendant ces années de thèse. Je me rattraperais, promis…*

*Pour finir, mes intimes remerciements vont à **Céline**, pour son soutien et sa patience, sa compréhension. De même, mille mercis à **Maman** et ma **famille** sans qui évidemment rien de cela n'aurait été possible.*

*Enfin, MERCI à mon **Papou** qui n'aura hélas pas eu le temps de me voir soutenir ma thèse, toutes mes pensées se tourne vers toi. Ton souvenir, ton sourire et ton humour restent gravés dans mon âme. Merci pour tout ce que tu as pu me transmettre pendant 29 ans.*

SOMMAIRE

LISTE DES FIGURES

LISTE DES TABLES

INTRODUCTION GENERALE

Depuis la préhistoire, le fromage est présent dans notre alimentation. Les premières traces de fromage apparaissent dès 8000 ans avant J.C. Cette apparition coïncide à celle de l'élevage datant du Néolithique. Marqueur de civilisation et de développement humain, le fromage répond à la nécessité de conserver le lait, source abondante de protéines.

Selon l'étymologiste Alain Rey, le terme Fromage est apparu vers 1135. Celui-ci est dérivé du bas latin *(caseus) formaticus* (fait dans une forme) issu du terme *forma*. Le terme « Fromage » provient d'une métathèse du mot « formage » attestée en 1180, détachant ainsi le mot de son origine (Rey, 1994).

Le rapport d'information, fait au nom de la commission des Affaires Culturelles sur l'inscription de la gastronomie au patrimoine immatériel de l'UNESCO, rappelle que le fromage recèle un potentiel d'évolution et de variabilité considérable. La diversité des fromages gaulois retient, à ce propos, l'attention du naturaliste Pline l'Ancien (23 av J.C.). Il évoque dans son livre Histoire naturelle, au chapitre XLII « *De diversitate caseorum* » : un fromage au lait de brebis dans la Lozère actuelle (semblable au Roquefort), les meules de fromages du pays Arvernes (Salers) amenées vers Rome par les soldats romain, ainsi que la grande diversité des fromages Alpins. Aujourd'hui encore, l'extraordinaire diversité des fromages français provient de la grande variété des techniques de fabrication utilisées mais également de la richesse des terroirs. En 1962, Charles de Gaulle décrète « qu'on ne peut pas gouverner un pays qui offre 258 variétés de fromages". Quant à Winston Churchill, il déclare pendant l'occupation allemande « qu'un pays capable de donner au monde 360 fromages ne peut pas mourir". En réalité, il existe plus de 1000 variétés de fromages dont 45 bénéficient d'une Appellation d'Origine Contrôlée (A.O.C) (**Figure 1**). Le maintien de cette diversité est dû au savoir-faire séculaire combiné à la variété des terroirs.

La France est réputée comme étant le "pays du fromage" tant au niveau de la production qu'au niveau de la consommation. Selon le Centre National Interprofessionnel de l'Economie Laitière (C.N.I.E.L, 2012), la France avec 1,80 millions de tonnes de fromages produits se situe, en 2011, au troisième rang de la production mondiale de fromages (en volume) après les USA et l'Allemagne. Les Français sont les premiers amateurs de fromages, avec une consommation de 23,7 kg de fromage par habitant et par année (C.N.I.E.L, 2012). Parmi ces fromages, la préférence des français va aux fromages à pâte pressé cuite (Emmental, Comté…). En 2011, les Français ont consommé 220.603 tonnes de fromages à pâte pressée cuite. Ensuite par ordre décroissant, on retrouve les fromages à pâte molle (Camembert, Livarot…), les fromages à pâte pressée non cuite (Reblochon, Salers…), les fromages frais salés (Brousse, Mascarpone,…), les fromages de chèvre (Picodon, Valençay,…), les fromages fondus (Cancoillotte, Vache qui rit,…) et enfin les fromages à pâte persillée (Roquefort, Fourme d'Ambert,…).

Figure 1 : Carte de France des 45 Fromages AOC français

Toujours en 2011, les exportations de fromages représentent une valeur de 2,8 millions d'euros (C.N.I.E.L, 2012). Les pays de l'Union Européenne sont les principaux destinataires de ces exportations (65 %). Les fromages à pâte molle, avec 174.159 tonnes, constituent la plus grande part des exportations fromagères françaises.

Pour conserver ce rang, et même l'améliorer, l'industrie fromagère française doit veiller au respect des normes d'hygiène de plus en plus strictes et au respect de la qualité organoleptique recherchée par les consommateurs. Pour satisfaire ces conditions, il est indispensable de maîtriser la matière

première mais également le processus de transformation du lait en fromage et notamment l'affinage qui constitue l'une des étapes clés du processus de fabrication. L'affinage résulte principalement de l'action d'une grande variété de micro-organismes tels que les bactéries, les levures et les moisissures qui participent à la transformation du caillé en fromage. L'évolution et l'activité de cette flore sont très influencées par les conditions d'affinage ainsi que par les interactions entre les différents micro-organismes.

La flore microbienne est primordiale pour l'établissement des propriétés organoleptiques (arôme, saveur, texture et couleur) mais également sanitaires (effet barrière, activités antimicrobiennes). Pendant 15 ans, avec l'avènement des techniques taxonomiques en biologie moléculaire, les principaux travaux concernant cette flore avaient pour objectif de réaliser un inventaire précis des espèces présentes dans les différentes variétés de fromages. Actuellement, nous avons donc une image plus juste de la biodiversité de cette flore. Elle est composée de bactéries, levures ou moisissure en grande majorité cultivable et est composée de plus d'une vingtaine d'espèces majeures.

Aujourd'hui le fromage est l'aboutissement d'une accumulation de savoirs, de pratiques, d'observations et d'ajustements. De nombreuses pratiques fromagères influent et interfèrent sur la flore microbienne active (température et temps de chauffe du lait, choix du ferment, taille des grains du caillé, intensité du pressage de la pâte, traitement accordé à la croûte, fréquence de retournements, degré d'hygrométrie). Cela se traduit alors par une grande diversité de matrices où moisissures, bactéries et levures évoluent.

C'est ainsi que dans le cadre de l'Union Européenne, l'EFSA (Autorité Européenne de Sécurité des Aliments) organise un débat pour établir une liste préliminaire d'espèces bactériennes et de levures pour lesquelles un statut QPS (Qualified Presumption of Safety) serait recevable. Ces flores complexes sont en grande partie inoculées par le producteur au cours du processus. Cependant, dans de nombreux fromages, une grande proportion de la flore présente sur le produit fini peut provenir de micro-organismes indigènes de l'environnement. Malgré le fait qu'ils sont parfois nécessaires à l'élaboration de ces produits, aucun champignon filamenteux ou microorganismes peu étudiés n'ont été jugé admissible du fait du manque de connaissances permettant leur traçabilité ou concernant leur innocuité.

Regroupant plusieurs laboratoires et industriels, mon sujet de thèse Exploration génomique des levures des fromages s'inscrit dans le projet ANR (Agence Nationale de la Recherche) du programme ALIA (ALimentation et Industries Alimentaires) : Food Microbiomes. Ce projet a pour but de palier

aux manques de connaissances concernant l'identité exacte des flores utilisées dans le fromage, leur histoire d'utilisation et leur métabolisme. Ces lacunes ne leur permettent pas d'avoir à l'heure actuelle un statut «QPS». Ce travail de thèse est basé sur le développement d'une approche innovante obtenue grâce aux nouvelles capacités de séquençage à très haut débit dans le but de construire une base de données des signatures nucléiques du « microbiome » de 40 fromages du patrimoine français et européen qui constituera ainsi une référence pour l'étude des écosystèmes fromagers. De plus, il conviendra au cours de ce programme de développer des outils innovants qui faciliteront l'identification des microorganismes quelle que soit la réglementation.

Tout d'abord, les connaissances actuelles dans la fabrication et l'affinage des fromages ainsi que la microbiologie de l'affinage du fromage seront présentées sous forme d'une synthèse bibliographique. Une deuxième partie abordera les connaissances taxonomiques des levures et leur utilisation dans les fromages. Puis, l'étude de la levure *Geotrichum candidum* sera présentée dans la troisième partie de cette introduction. Ensuite, nous nous intéresserons à l'adaptation des levures aux conditions industrielles et leur évolution due aux utilisations humaines. Enfin dans une dernière partie seront présentées les différentes méthodes de typages effectuées sur les levures et, en particulier, les différentes méthodes de typage existantes sur *Geotrichum candidum*.

La partie résultats aura pour but d'apporter des connaissances académiques sur la levure *G. candidum* et sur les souches industrielles : phylogénie, diversité et premières analyses du génome.

Finalement, une conclusion générale récapitulera les principaux résultats de ce travail et apportera une analyse plus globale en termes d'application dans la stratégie de sélection des ferments d'affinage. Le manuscrit se conclura en présentant les principales perspectives envisagées pour la poursuite de cette thématique de recherche.

En annexe I et II seront présentées deux publications auxquelles j'ai contribué : une revue portant sur les nouvelles perspectives de la taxonomie des levures hémiascomycètes et une publication sur la séquence de la souche *Millerozyma sorbitophila* CBS 7064.

RAPPORT BIBLIOGRAPHIQUE

1 LE FROMAGE, UN ECOSYSTEME MICROBIEN COMPLEXE

1.1 Principes généraux de la technologie fromagère

Les fromages sont issus d'un savoir faire traditionnel, acquis au cours des siècles, qui fait partie du patrimoine culturel de l'humanité. Selon le décret n° 88-1206 du 30 décembre 1988, le terme «fromage» est réservé au produit obtenu à partir de matières d'origine exclusivement laitière. La base essentielle du fromage est le lait. Il peut être de diverses origines, de vache principalement, mais également de brebis, de chèvre, de bufflonne ou d'autres mammifères. Ce produit peut être fermenté ou non, affiné ou non, utilisé seul ou en mélange, coagulé en totalité ou en partie avant égouttage ou après élimination partielle de son eau. Il constitue alors un moyen de conservation alimentaire.

Le processus de la fabrication du fromage passe par quatre grandes étapes de fabrication :

- La première étape est le **caillage**, appelé aussi la coagulation. Il s'effectue grâce à une acidification du lait dans le cas des fromages frais ou par apport d'enzymes coagulantes. Elle conduit à l'obtention d'un gel ou coagulum.

- L'étape qui suit est l'**égouttage**. Elle correspond à la séparation du caillé, la phase solide, et du lactosérum, une phase liquide composée d'eau et des matières solubles (lactose, sels minéraux et protéines solubles). Le salage complète l'étape d'égouttage et contribue à la formation de la croûte.

- Le **salage** agit directement ou par intermédiaire de l'activité de l'eau du fromage sur le développement des microorganismes et sur les activités enzymatiques au cours de l'affinage. Il conduit à la formation de la croûte en créant une zone riche en sel et faible en eau.

- Enfin, l'**affinage** est le stade ultime du processus. Il consiste en une digestion enzymatique, essentiellement d'enzymes microbiennes, du caillé sous l'action des agents coagulants et des microorganismes de la flore, qui complètent et achèvent ce processus de maturation appelé «protéolyse». Il conduit à l'obtention d'un fromage affiné.

La composition et l'évolution de la flore du fromage joue un rôle primordial dans la typicité et la qualité du produit obtenu. La durée de l'affinage varie d'un fromage à un autre, de quelques semaines pour un Camembert à plusieurs mois pour un fromage à pâte. Le fromage évolue donc différemment en fonction du type de pâte pour dégager des saveurs typiques de chaque terroir. Seuls les fromages à pâtes fraîches ne subissent pas cette étape d'affinage.

Selon les paramètres mis en œuvre au niveau des différentes étapes de transformation du lait en fromage, une grande variété de produits peut être obtenue tel que traduit par Fox (2004). Il existe alors une large variété de fromages distincte par le type de la technologie employée (

Figure 2).

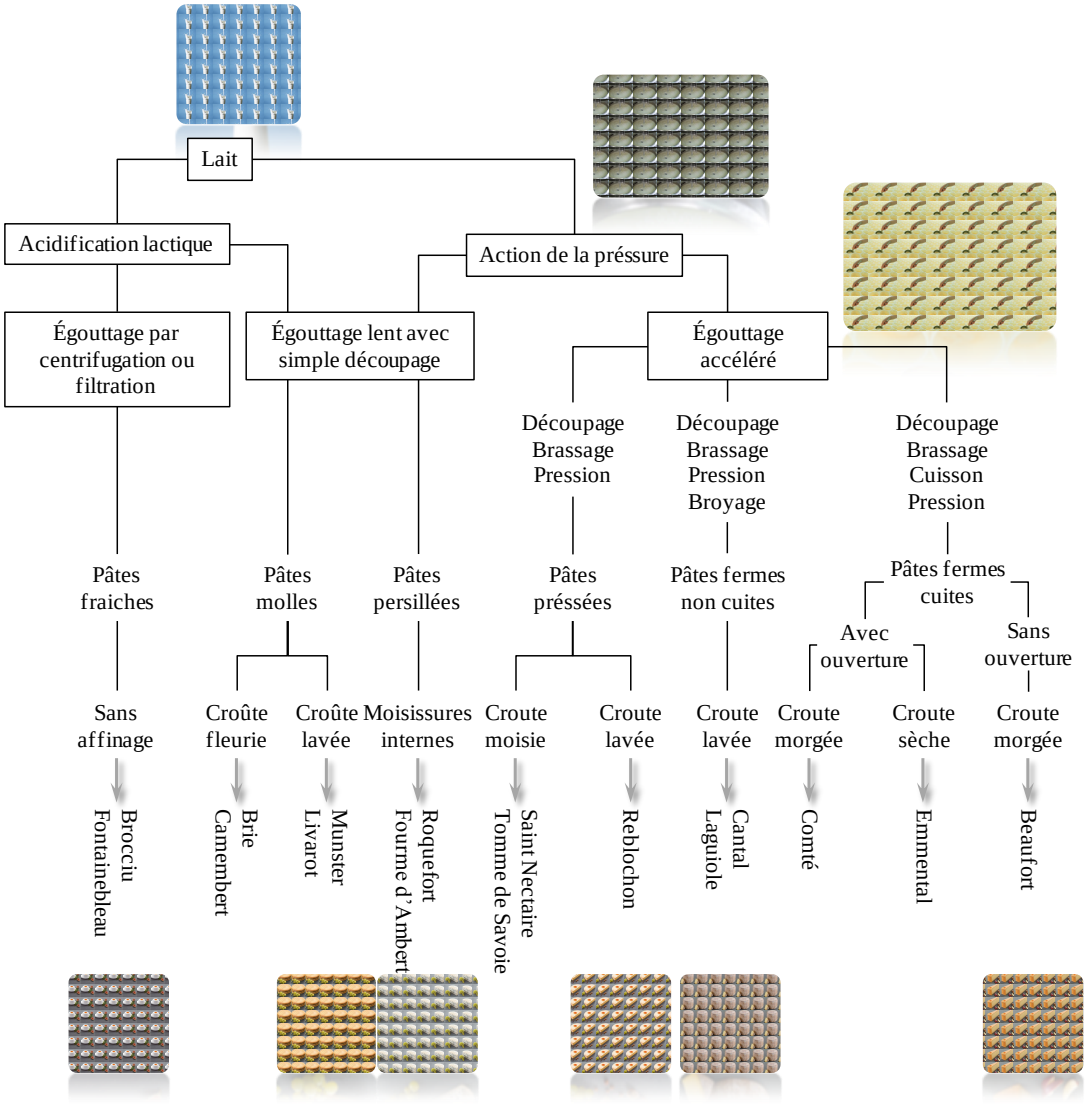


Figure 2 : Diversité de fabrication des spécialités fromagères

1.2 La microbiologie de l'affinage

L'écosystème fromager est une véritable source de biodiversité du fait de la composition de sa flore microbienne. En effet, bactéries, moisissures et levures cohabitent sur la croute et à l'intérieur même du fromage.

1.2.1 Les flores bactériennes

La flore bactérienne des fromages est composée de bactéries Gram négatif qui appartiennent aux familles des *Moraxellaceae*, *Pseudomonadaceae* et des *Enterobacteriaceae*. Elles possèdent des propriétés protéolytiques et lipolytiques impliquées dans le processus d'affinage du fromage.

Elle est aussi composée des bactéries Gram positif, les staphylocoques et les bactéries corynéformes. Les staphylocoques retrouvés dans le fromage sont *Staphylococcus equorum, Staphylococcus vitulinus* et *Staphylococcus xylosus*. Les souches isolées sont halotolérantes et possèdent aussi des propriétés lipolytiques et protéolytiques qui participent au processus d'affinage. Les bactéries corynéformes sont acido-sensibles et le plus souvent aérobies. Le genre le plus étudié est *Brevibacterium*. Ces bactéries possèdent des activités protéolytiques, lipolytiques et estérasiques favorisant leur développement sur la matrice fromagère tout en conduisant à la synthèse de composés d'arômes qui participent aux saveurs du fromage (Goerges et *al.*, 2008).

1.2.2 Les flores fongiques

Outre les espèces emblématiques, *Penicillium camemberti* et *Penicillium roqueforti*, les champignons issus de fromages ne sont pas bien connus. Il a été recensé près de 24 espèces différentes dont *P. camemberti, P. roqueforti* et *Mucor sp.* (Hayaloglu et Kirbag, 2007)**.**

Dans le cadre de ce projet, une taxonomie des champignons filamenteux et des mucors trouvés dans les fromages a été effectué.

Le genre *Mucor* comprend différentes espèces rencontrées dans les fromages. Dans une étude incluant 70 *Mucor sp.,* Hermet et *al.* (2012) recense 6 espèces du genre *Mucor* dans les fromages. Ce sont : *M. circinelloides*, *M. racemosus*, *M. brunneogrisus*, *M. spinosus*, *M. fuscus* et *M. lanceolatus* (Hermet et *al.*, 2012).

Les ascomycètes filamenteux isolés de fromage sont principalement des membres issus de deux classes : les Eurotiomycetes et les Sordariomycètes (Ropars et *al.*, 2012). Au sein des Eurotiomycetes, trois genres sont capables de se développer sur les fromages, *Penicillium*, dont les fameux *P. roqueforti* (fromages bleus, fourmes) et *P. camemberti* (camemberts, bries, coulommiers, etc.). Ou encore le genre *Sporendonema*, auquel appartient *S. casei*, utilisé pour le cantal et le salers. Quant à la classe des Sordariomycètes, on y trouve deux grands genres : les *Fusarium* (*F. domesticum* dans le saint-nectaire et le reblochon) et les *Scopulariopsis* (*S. spp.* dans les tommes des Pyrénées et l'Ossau-Iraty du Pays basque). Les moisissures jouent un rôle déterminant dans la formation des caractéristiques sensorielles des fromages.

Parmi les flores fongiques présentes dans les fromages sont aussi retrouvées les levures.

2 LES LEVURES HEMIASCOMYCETES

Les levures, en raison de leur capacité à fermenter les sucres pour produire l'éthanol et du dioxyde de carbone, sont bien connus pour leur rôle important dans la fabrication de divers aliments et boissons. Ils ont été utilisés depuis des millénaires pour fabriquer des boissons fermentées et des

aliments tels que le vin, le cidre, la bière, le pain et les produits laitiers. Les levures qui nous intéressent ici sont aussi appelées hémiascomycètes et appartiennent aux sous phylum des saccharomycotina (James et *al.*, 2006). Les relations phylogénétiques de 70 espèces types d'hémiascomycète sont présentées dans la **Figure** 3 (Kurtzman et Robnett, 2012).

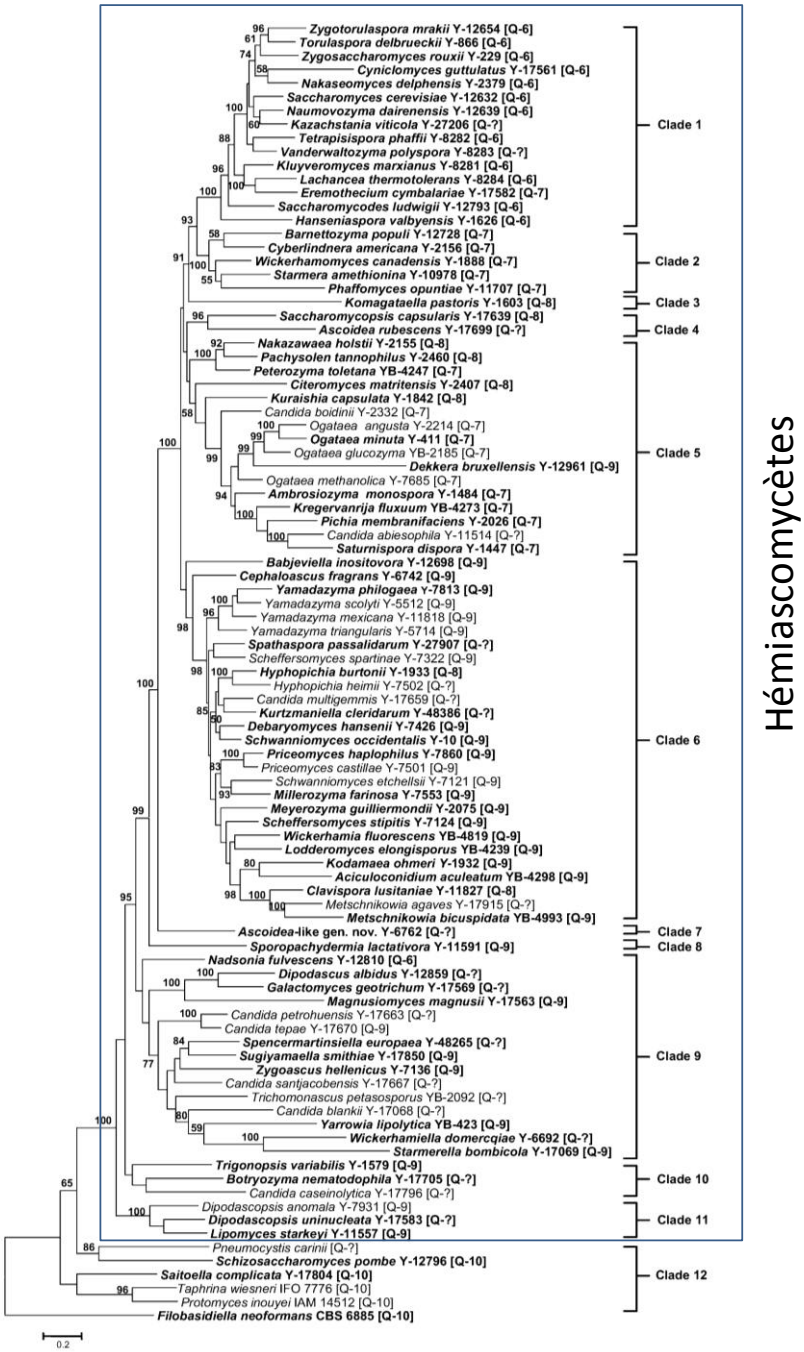


Figure 3 : Relations phylogénétiques entre les levures hémiascomycètes (Kurtzman et Robnett, 2012)

2.1 Taxonomie des levures

La taxonomie consiste essentiellement à : (i) une classification des micro-organismes en groupes taxonomiques selon leur ressemblance et une similarité de séquence, (ii) la catégorisation de ces groupes définis, et (iii) la description de nouveaux organismes. La taxonomie se réfère souvent à l'approche systématique qui établit des différences et des relations entre les organismes afin de les caractériser et de les classer.

2.1.1 Méthodes d'identification biochimique, physiologique et morphologique

Il existe de nombreuses méthodes d'identification des microorganismes basées sur les caractéristiques phénotypiques et physiologiques. La combinaison de ces deux caractéristiques a traditionnellement été utilisée pour les études de taxonomie des levures. Les caractéristiques phénotypiques analysées sont généralement la morphologie avec une observation au microscope pour déterminer le type de l'espèce (levure, champignon ou intermédiaire), la résistance au toucher, pour déterminer si une souche est grasse, poudreuse, libère de l'eau ou se décolle de la gélose, et la croissance (Marcellino et *al.*, 2001). Les caractéristiques physiologiques étudiées sont généralement la fermentation de sucres, l'assimilation de sources de carbone, l'assimilation de huit sources d'azote, la croissance sur cycloheximide à 0,01 %, la croissance sur glucose 50 %, la thermo tolérance *via* test de croissance à quatre températures différentes : 30°C, 35°C, 37°C et 40°C, la production d'uréase sur milieu Christensen, la dégradation de l'arbutine, la dégradation d'amidon (pour certaines espèces), la sporulation, la résistance aux antifongiques ou l'osmotolérance.

Cependant, ces tests ne sont pas satisfaisants pour la délimitation et l'identification des espèces de levures. Les méthodes basées sur les caractéristiques phénotypiques et physiologiques sont de moins en moins utilisées du fait de leur faible reproductibilité et leur faible capacité à discriminer les isolats. Ainsi, les méthodes de typage moléculaire sont largement répandues (Vanhee et *al.*, 2010).

2.1.2 Méthodes d'identification moléculaire

Le séquençage de l'ADN a complètement bouleversé notre vision de la taxinomie des espèces. Les séquences d'ADN ribosomiques ont fait preuve de leur universalité et se montrent appropriées à la délimitation des espèces (Kurtzman, 1992; White et *al.*, 1990). Ainsi dans un premier temps, la séquence D1D2 du 26SrDNA a été majoritairement utilisé (Kurtzman et Robnett, 1998). Utilisées en parallèle pour l'identification des champignons, les séquences ITS sont depuis peu considérées comme un potentiel « barcoding » moléculaire universel pour les espèces fongiques (Schoch et *al.*, 2012) (**Figure 4**).

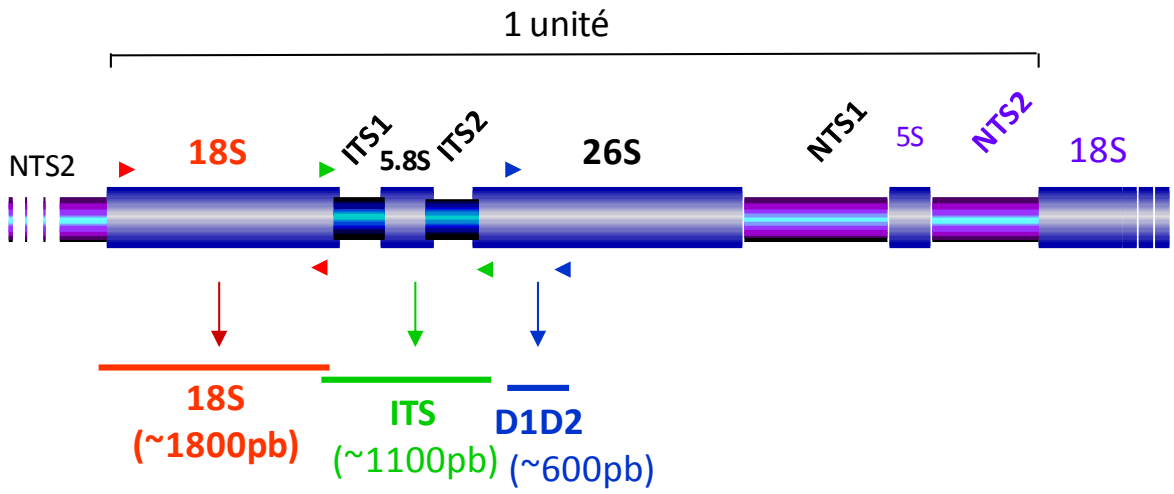


Figure 4 : Schéma représentatif de la grande sous unité de l'ADN ribosomique

Dans la majeure partie des cas, la séquence D1/D2 est considérée comme discriminante pour l'identification des levures. Toutefois, certains exemples montrent que ce n'est pas toujours le cas (Jacques et *al.*, 2009; Kurtzman, 2011).

Cependant, d'un point de vue phylogénétique, il est apparu que les analyses multigéniques sont plus discriminantes et fournissent des arbres plus robustes (Kurtzman et Robnett, 2003). Cela a conduit à une nouvelle classification des principaux clades de levures, celle-ci reste à terminer (Kurtzman et Robnett, 2012).

Plus récemment, la disponibilité d'un nombre croissant de génomes complets a apporté une nouvelle perspective à la taxonomie moléculaire en fournissant un grand nombre de gènes à comparer. Aguileta et *al.* (2008) propose alors un set de 246 gènes performants pour établir une phylogénie robuste. Plus de détails sur la taxonomie des levures sont présentés en **annexe 1** de ce manuscrit.

La classification des espèces est toujours en évolution, il est fort à parier que les clades établis vont être changés devant la recrudescence de nouvelles études écologiques entreprises. Bien qu'encore centrées sur les espèces du genre *Saccharomyces*, ces études écologiques sont effectuées dans de nombreuses régions du monde (Equateur,…) (James et *al.*, 2012) et dans des écosystèmes nouveaux (glaciers, …) (Butinar et *al.*, 2007).

2.2 Rôles des levures dans les fromages

Par leur caractère ubiquitaire, les levures sont fréquemment retrouvées dans les fromages. Certaines espèces sont particulièrement adaptées à ce type de substrat car elles ont la capacité de se développer à basse température, à un faible pH et/ou à de fortes concentrations en sel (Fleet, 1990). Le rôle le plus largement décrit de ces levures dans la fabrication fromagère est la consommation de l'acide lactique qui provoque un phénomène fondamental, la désacidification du caillé. Cela rend possible le développement des bactéries d'affinage acido-sensibles (Gueguen, 1992). Ainsi, l'implantation précoce de *Brevibacterium aurantiacum* et/ou *linens*, ou de la flore corynéforme est favorisée par la croissance de *G. candidum* (Lecocq et *al.*, 1996) ou de *D. hansenii* (Eliskases-Lechner et Ginzinger, 1995). L'activité lipolytique et protéolytique peut conduire à des modifications de texture (Vassal et *al.*, 1986) ou à des meilleures activités enzymatiques. Les levures ajoutées au fromage contribuent à l'affinage et au développement de la flaveur.

Dans une revue, Fleet (1990) liste différentes propriétés des levures sélectionnées pour leur croissance et leur prédominance dans les produits laitiers (**Tableau 1**)

Propriétés des levures sélectionnées pour leur croissance et leur prédominance dans les produits laitiers
1. Fermentation ou assimilation du lactose
2. Production d'enzymes protéolytiques extracelluaires
3. Production d'enzymes lipolytiques extracellulaires
4. Assimilation de l'acide lactique ou lactate
5. Assimilation d'acide citrique
6. Croissance à basse température
7. Tolérance à une concentration élevée en sel

Tableau 1 : Liste des différentes propriétés des levures sélectionnées pour leur croissance et leur prédominance dans les produits laitiers (Fleet, 1990)

L'étude des populations de levures dans le livarot (Gente et *al.*, 2007), montre que les espèces de genre *Kluyveromyces* interviennent en début d'affinage. Alors que *G. candidum* et *D. hansenii* jouent un rôle tout au long de l'affinage et peuvent aussi bien participer à la désacidification du caillé qu'à la production de composés aromatiques. Les principales levures trouvées dans les fromages sont listées dans le **Tableau 2**.

Levures présentes dans les fromages	
Espèces les plus fréquemments décrites[b,c]	Espèces rarement décrites[b,c]
Kluyveromyces marxianus* var. *marxianus	*Saccharomyces unisporus*
Kluyveromyces lactis* var. *lactis	***Saccharomyces exiguus***
Debaryomyces hansenii	*Dipodascus capitatus*
Debaryomyces fabryi	*Pichia fermentans* var. *fermentans*
Debaryomyces tyrocola	*Pichia kluyverii* var. *kluyverii*
Saccharomyces cerevisiae	*Pichia membranifaciens*
Candida zeylanoides	*Pichia pseudocactophila*
Candida catenulata	*Candida rugosa*
Candida intermedia	*Candida sake*
Geotrichum candidum	*Candida tenuis*
Torulaspora delbrueckii	*Pichia jadinii*
Yarrowia lipolytica	*Dipodocus capitatus*
	Candida versatilis
	Issatchenkia occidentalis
	Clavispora lusitaniae
	Zygosaccharomyces rouxii
	Williopsis californica

[b]: L'appartion d'une espèce et son nombre de représentants varient considérablement selon le type de fromage, la laiterie, la localisation géographique, la saison. Seules les espèces les plus dominantes sont indiquées.
[c]: en gras sont présentées les espèces de levures séquencées

Tableau 2 : Liste des principales espèces de levures trouvées dans les fromages

L'action des levures commence dès les premières heures de l'égouttage et se poursuit pendant tout l'affinage. Nous retrouvons alors dans le fromage une succession de flores qui interagissent entre elles au cours d'un processus dynamique. Ainsi, les levures dans les produits laitiers peuvent interagir avec d'autres microorganismes de deux manières différentes : (i) elles peuvent contribuer positivement au processus de fermentation ou de maturation en soutenant la fonction de la culture de départ ; (ii) elles peuvent inhiber ou éliminer les micro-organismes qui ne sont pas souhaités parce qu'ils provoquent des défauts de qualité ou possèdent des caractères pathogènes potentiels.

Le rôle possible des interactions positives entre les levures et ferments lactiques est bien documenté sur les fromages. Ainsi en début d'affinage, *Debaryomyces hansenii* entre en jeu. Cette levure peut tolérer des niveaux de salinité jusqu'à 24 %, alors que la croissance de *Saccharomyces cerevisiae* est inhibée lorsque la salinité atteint 10 %. De *plus, D. hansenii* est capable de se développer à faible température. De ce fait, l'aptitude de certaines souches de *D. hansenii* à utiliser le lactose et l'acide lactique, à croître à basse température, en milieu acide et en présence de sel sont autant de raisons qui permettent à cette espèce de dominer dans les saumures et les fromages (Besancon et *al.*, 1992; Fleet, 1999; Roostita et Fleet, 1996b). Ces propriétés lui confèrent un rôle important dans plusieurs processus agro-alimentaires. *D. hansenii* est l'espèce de levure la plus couramment trouvée dans tous les types de fromage. *D. hansenii* est également fréquent dans les laiteries et en saumure, (Seiler et Busse, 1990). Elle est la levure majoritaire des fromages à pâte molle et à croûte lavée. Certaines souches de *D. hansenii* présentent la particularité de consommer le lactose et l'acide lactique en

même temps. Cette levure désacidifie très rapidement le caillé lors de la fabrication fromagère (consommation de l'acide lactique et production d'ammoniac lors de la dégradation des acides aminés) (Mounier et *al.*, 2008). En compétition avec *D. hansenii*, *Kluyveromyces lactis* et *Kluyveromyces marxianus* sont deux levures retrouvées naturellement dans le fromage. Elles sont aussi utilisées en technologie fromagère par leur ensemencement volontaire dans le lait. Les levures du genre *Kluyveromyces* ont la capacité de consommer le lactate et le lactose. Ainsi les *Kluyveromyces* participent à la désacidification de caillé (Cholet et *al.*, 2007; Kagkli et *al.*, 2006). La désacidification du caillé permet l'implantation ultérieure d'une flore acido-sensible comme les bactéries corynéformes.

Cependant, les souches *Kluyveromyces lactis* sont préférentiellement des agents d'aromatisation en cours d'affinage. En effet, les levures du genre *Kluyveromyces* sont capables de produire des composés d'arôme variés, principalement des esters aux notes fruitées et des alcools. *K. lactis* peut produire des composés soufrés volatils (CSVs) comme le méthanethiol (MTL), le diméthylsulfure (DMS), le diméthyldisulfure (DMDS), le diméthyltrisulfure (DMTS) ou encore le méthythioacétate (MTAc) (Arfi et *al.*, 2002). Il est à noter que *D. hansenii* possède aussi des capacités aromatiques intéressantes, notamment par la production de composés soufrés tels que le DMDS, DMTS et le MTAc (Cholet et *al.*, 2007).

Yarrowia lipolytica est une levure aérobie stricte retrouvée dans les matrices alimentaires riches en lipides et en protéines telles que les fromages et la charcuterie (Gardini et *al.*, 2001; Mounier et *al.*, 2009). Cette levure sécrète des protéases, des lipases, des phosphates et des estérases qui favorisent son développement sur la matrice fromagère (Mansour et *al.*, 2008). Elle n'est pas utilisée comme starter et est parfois considérée comme contaminant.

Saccharomyces cerevisiae est généralement connue pour son implication dans l'élaboration du pain, du vin et de la bière. Elle est utilisée dans les deux derniers cas pour sa contribution lors de la fermentation alcoolique (Fleet, 2003). Cette levure a été, par ailleurs, identifiée dans différent types de fromages (Romano et *al.*, 2001; Roostita et Fleet, 1996b; Viljoen et Greyling, 1995). *S. cerevisiae* n'assimile ni le lactose ni l'acide lactique (Hansen et Jakobsen, 2001; Roostita et Fleet, 1996a). Sa croissance dans les produits laitiers serait donc liée à l'utilisation des acides aminés comme source de carbone et d'azote. L'influence de *S. cerevisiae* sur les qualités organoleptiques fromagères n'a pas encore été caractérisée. Cependant, il a été observé que les souches retrouvées dans le fromage présentent une résistance accrue au sel (Hansen et Jakobsen, 2001).

L'industrie alimentaire étant constamment en recherche de développement de nouveaux produits, les levures présentent de nouvelles pistes d'exploitation. En effet, les levures se développent en coopération avec les bactéries lactiques, et sont souvent à l'origine de la typicité organoleptique des

produits. De plus, les levures peuvent être utilisées comme barrière biologique contre les micro-organismes indésirables et comme agents probiotiques (Fleet, 2007).

Dans ce qui suit, la taxonomie ainsi que les différentes connaissances sur *Geotrichum candidum* décrites seront détaillées.

3 LA LEVURE *GEOTRICHUM CANDIDUM*

3.1 Taxonomie de *Geotrichum candidum*

La taxonomie de *G. candidum* a longtemps été débattue comme l'atteste les nombreux noms qui lui ont été attribué. La levure *Geotrichum candidum* a été décrite par Link en 1809. Depuis, plusieurs synonymes lui ont été attribués tels que : *Botrytis geotricha* (Link 1824), *Oidium lactis* ou *Oospora lactis* (Wouters et *al.*, 2002), *Endomyces geotrichum (Butler et Petersen, 1972), Galactomyces geotrichum (Redhead et Malloch, 1977)*, *Galactomyces candidus* (de Hoog et Smith, 2004) (**Tableau 3**).

	Nom de l'espèce	année de description
Evolution de nom de l'espèce	*Geotrichum candidum*	1809
	Botrytis geotricha	1824
	Oidium lactis	1850
	Dipodascus geotrichum	1972
	Endomyces geotricum	1972
	Galactomyces geotrichum	2000
	Galactomyces candidus	2004
Synonymes retrouvés	*Mycoderma malti juniperini*	1827
	Acrosporium candidum	1827
	Torula geotricha	1829
	Oidium lactis var. luxurians	1854
	Oidium obtusum	1875
	Oidium nubilum	1909
	Oidium humi	1910
	Monila asteroides	1914
	Oidium matalense	1915
	Oidium suaveluens var. minutum	1923
	Oospora fragrans var. minuta	1923
	Oospora lactis var. exuberans	1923
	Geotrichum matalense var. chapmanii	1932
	Geotrichum javanense	1933
	Geotrichum versiforma	1934
	Geotrichum radaelli	1940
	Geotrichum novakii	1966
	Geotrichum silvicola	2005
	Geotrichum bryndzae	2009

Tableau 3 : Evolution du nom de l'espèce *Geotrichum candidum* et différents synonymes retrouvés (www.mycobank.com)

Bien que placé de façon non ambigüe dans les levures hémiascomycètes sur la base de la séquences D1/D2 par Kurtzman et ses collaborateurs (Kurtzman et Robnett, 1995), de nombreux auteurs ont continué à considérer *G. candidum* en tant que champignon filamenteux. Elle est pourtant décrite comme moisissure par Wouters et *al.* (2002), ou encore présentée comme « *filamentous yeast-like fungi* » par de Hoog et Smith (2004). En raison de sa morphologie semblable aux champignons et de sa position taxonomique particulière, beaucoup de confusions ont été associées à la désignation de cette espèce. Ainsi, considéré comme un champignon, la forme imparfaite n'a pas été décrite comme un *Candida* mais comme *Geotrichum*. Ainsi lors de la mise en évidence de sexualité chez cette espèce, le téléomorphe de cette espèce a été proposé comme étant *Endomyces Geotrichum* (Butler et Petersen, 1972). Cette espèce a ensuite été réaffectée au genre *Galactomyces* sous le nom *Galactomyces geotrichum* (Redhead et Malloch, 1977). Cela a conduit à l'actuelle coexistence d'une espèce avec deux noms de genres, *Geotrichum* et *Galactomyces*. *Geotrichum candidum* est donc le synonyme de *Galactomyces candidum*.

En utilisant la méthode d'hybridation de l'ADN, quatre sous-groupes de *G. geotrichum* : *G. geotrichum* stricto sensu, *G. geotrichum* groupe A, B et C ont été distingués dans ce complexe (Smith et *al.*, 1995). Une étude utilisant la méthode classique d'identification et le calcul du % GC a été réalisée pour identifier les taxons des genres : *Geotrichum*, *Galactomyces* et *Dipodascus* (Smith et *al.*, 2000). Plus récemment, une analyse taxonomique basée sur la phylogénie 18S et ITS de *Geotrichum* et des genres apparentés tels que *Dipodascus*, *Magnusomycetes* et *Spharochaete* a suggéré que le complexe *G. geotrichum* / *G. candidum* contenait quatre espèces distinctes (de Hoog et Smith, 2004). L'état de téléomorphe de *G. candidum* a été changé de *Gal. geotrichum* à *Gal. candidus*, depuis cette espèce est distinguée de *Gal. geotrichum* (**Figure 5**).

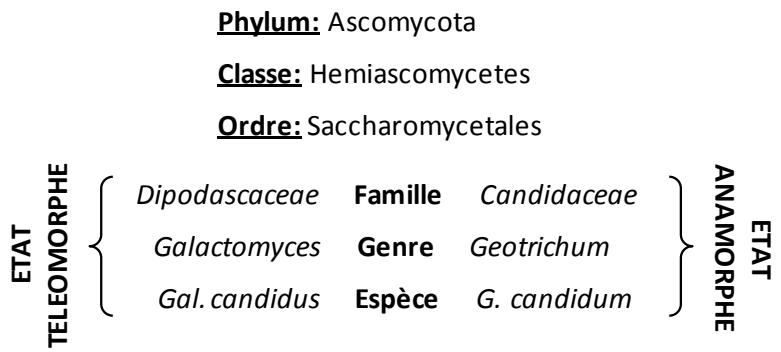


Figure 5 : Nomenclature actuelle des *Galactomyces candidus* / *Geotrichum candidum*.

Etat téléomorphe = forme sexuelle, état anamorphe = forme asexuée. (Pottier et *al.*, 2008)

Ce changement de nom de genre a entrainé depuis de nombreuses confusions entre *Galactomyces geotrichum* et *Geotrichum candidum.* Ainsi, il est fréquent lors de la description de nouvelles espèces que *Gal. geotrichum* soit présent et non *Gal. candidum*.

Les deux noms de genre étant accepté, pour des raisons de praticité, nous avons choisi d'utiliser la dénomination de *G. candidum* dans cette étude car c'est sous cette dénomination que cette espèce est très utilisée dans l'industrie biotechnologique et agro-alimentaire. Ainsi, un changement de nom pourrait introduire des confusions supplémentaires importantes.

La séquence d'ADN ribosomique, bien que très utile pour différencier les espèces n'a pas était en mesure de résoudre l'énigme de cette taxonomie.

Il a par ailleurs été montré sur la base de l'étude de 62 souches, que *Geotrichum candidum* possède un polymorphisme dans sa séquence de rDNA. 32 séquences différentes ont pu être déterminées (Alper et *al.*, 2011). Ce polymorphisme n'est pas exclusif à *G. candidum* et peut être trouvé chez d'autres espèces comme *Clavispora lusitaniae* (Lachance et *al.*, 2003). Il existe 10 variants de la région D1/D2 du rDNA dans les souches de *C. lusitaniae*. Le polymorphisme peut aller jusqu'à plus de 60 paires de base dans la région 18S-ITS1-5,8S-ITS2-26S (**Figure 4**). L'utilisation de ce type de séquences pour l'identification de *Geotrichum candidum* peut-être alors remise en cause. Cependant chez les hémiascomycètes, ces analyses multigéniques ont amélioré la classification des levures ascomycètes en fournissant des taxons circonscrits et des phylogénies robustes (Kurtzman et Robnett, 2003; Suh et Blackwell, 2006). Au moment où j'écris ces lignes, Groenewald et *al.* (2012) vient de publier une étude des *Geotrichum bryndzae, Geotrichum phurueaensis, Geotrichum silvicola* et *Geotrichum vulgare*, en utilisant la concaténation des régions D1/D2 et de gène partiel de l'actine. *G. bryndzae* et *G*. *silvicola* sont à présent considérés comme *Gal. candidum* et *G*. *vulgare* comme *G*. *pseudocandidum*. Par conséquent, l'actuelle classification de *Geotrichum* et des taxons apparentés est uniquement basée sur les séquences 5,8S et la séquence dite D1D2 du LSU RDNA et de l'exon de l'actine (**Figure 6**), (Groenewald et *al.*, 2012).

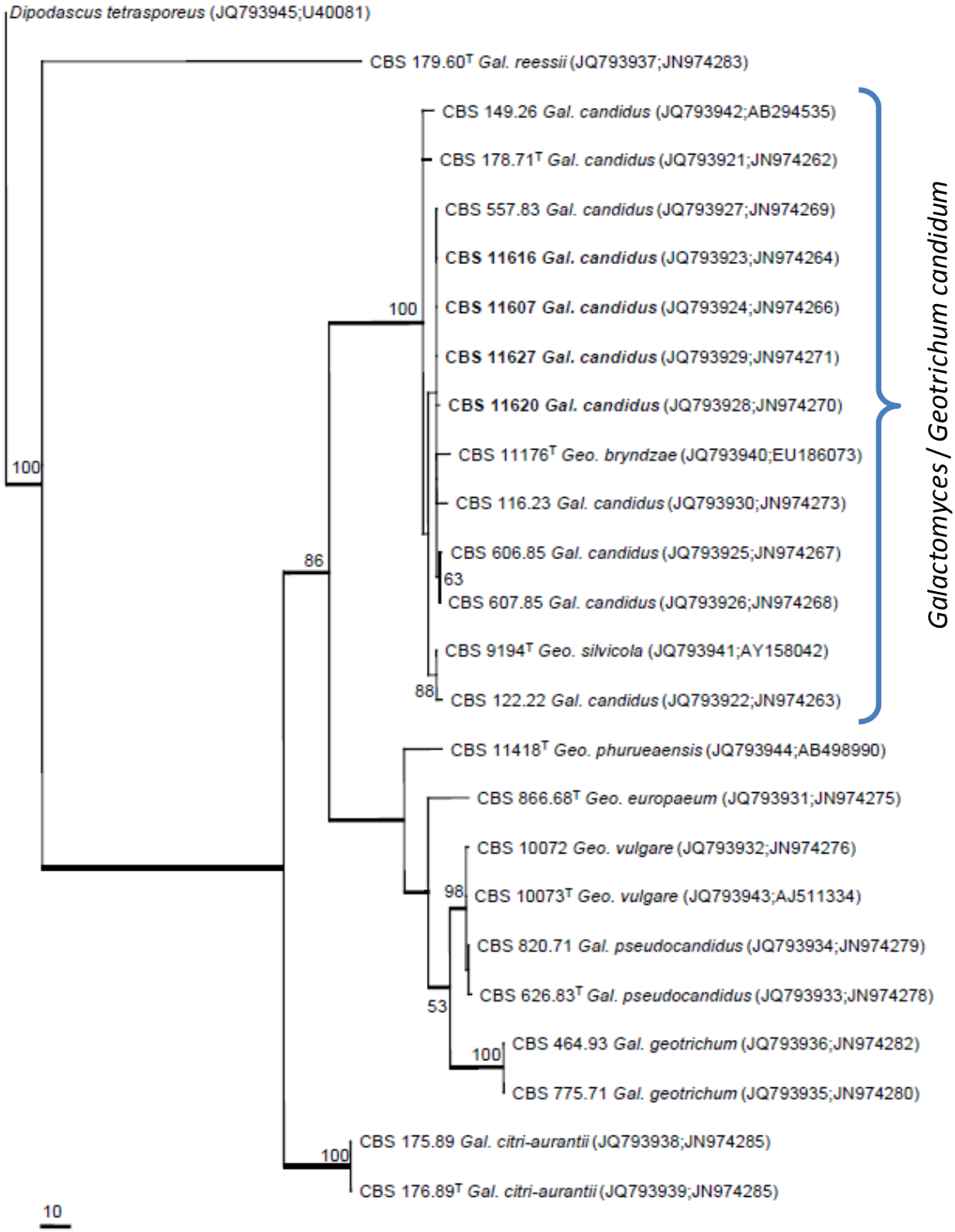
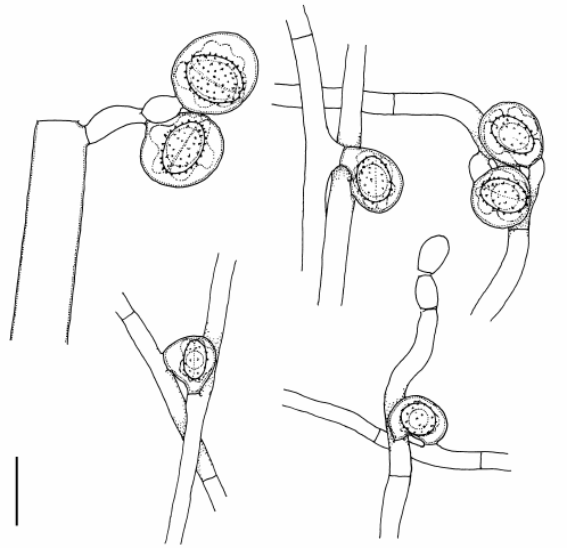


Figure 6 : Délimitation de l'espèce *Geotrichum candidum* d'après Groenewald et *al.* (2012)

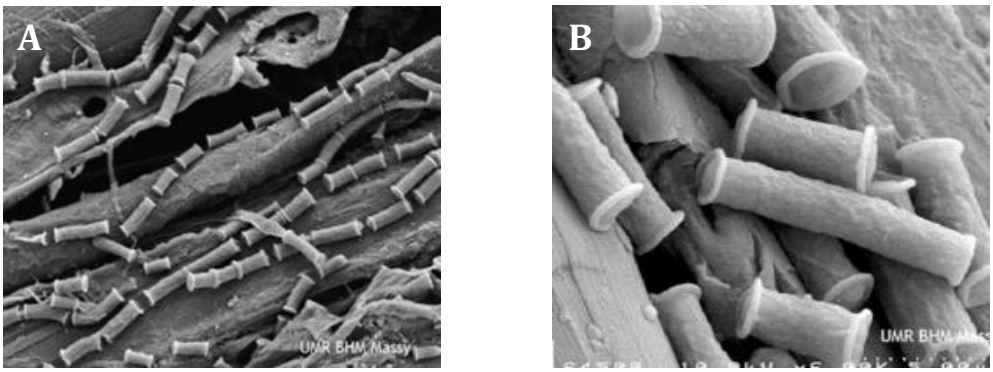
3.2 Morphologie

Selon la description (de Hoog et Smith, 2011), *Geotrichum candidum* produit des asques sur des hyphes indifférenciés sans cloisonnement visible, ou sur le même hyphe avec des gamétanges irrégulières. Un asque peut être formé d'une ascospore. Les ascospores sont globalement ellipsoïdales et ont une taille de 4,0 à 5,5 × 6-8 µm (**Figure 7** et **Figure 8**).



Asques bipodales prédominants avec ascospores matures.
Echelle, barre = 10 µm.

Figure 7 : *Geotrichum candidum*, CBS 180.33 × CBS 557.83. (de Hoog et Smith 2011)



A: *G. candidum* (cylindre) et bactéries sur la surface d'une étagère en bois d'une laiterie x1300
B: *G. candidum* et bactéries sur la surface d'un fromage de chèvre x6000

Figure 8 : Microscopie électronique à balayage de milieu contenant *G. candidum* (Mariani et *al.*, 2007; Mariani et *al.*, 2011)

L'étude des caractères des cultures de *G. candidum* sur différents milieux solides (Gueguen et Jacquet, 1982) montre qu'il est possible de différencier 3 types morphologiques sur l'unique milieu à l'extrait de malt :

- Type 1 : souches de couleur crème, à l'aspect levuriforme, à température optimale située entre 22 et 25°C, à croissance plus réduite à 30°C, à production abondante d'arthrospores, donnant peu de mycélium ; plutôt acidifiantes, à activité protéolytique faible.
- Type 2 : souches dites « intermédiaires ».
- Type 3 : souches bien blanches, plus ou moins feutrées, à température optimale plus élevée (25-30°C), à croissance plus faible à 22°C, sporulant peu, produisant en milieux liquides des mycéliums plus importants que les souches du type 1 ; plutôt alcalinisantes, à activité protéolytique plus marquée.

3.3 Rôles de *Geotrichum candidum* dans les fromages

Outre le fait que l'on trouve des *G. candidum* dans différents habitats comme les plantes, le sol, l'eau, les fourrages, le tractus digestif de l'homme et d'autres mammifères (Pottier et *al.*, 2008), on fait état de présence de *G. candidum* dans la plupart des inventaires de levures des produits laitiers et dans l'ensemble des fromages de production française. La levure d'intérêt apparaît dans les premiers stades de maturation sur fromages à pâte molle comme le Camembert, ou des fromages à pâtes semi-dures (Saint-nectaire, Reblochon) (Marcellino et *al.*, 2001). *G. candidum* est l'espèce de levure dominante dans le Livarot (Larpin et *al.*, 2006), mais aussi dans des fromages italiens comme le Pecorino, la caciotta, la feta, la ricotta fromage au lait de brebis de Sardaigne (Cosentino et *al.*, 2001; Fadda et *al.*, 2001; Fadda et *al.*, 2004). Dans ces derniers fromages, *G. candidum* est l'espèce dominante après *D. hansenii* avec 281 souches identifiées. Dans le fromage polonais : le Rokpol (Wojtatowicz et *al.*, 2001), *G. candidum* est la levure la plus fréquemment identifiée après *D. hansenii*, *C. spherica*, *C. intermedia*.

G. candidum joue de nombreux rôles importants dans la fabrication fromagère, en particulier durant l'étape d'affinage où elle apparaît très tôt. *G. candidum* libère des enzymes comme des lipases et des protéases. L'activité de ces dernières libère des acides gras et des peptides pouvant être métabolisés par les autres populations microbiennes et qui contribuent au développement des saveurs et des autres qualités du fromage.

G. candidum réduit l'amertume des camemberts industriels à travers l'activité de ses aminopeptidases et confère un arôme similaire au camembert traditionnel normand (Mourgues et *al.*, 1983). De nombreuses publications font état de son implication dans la fabrication d'arôme. Ce sont essentiellement des composés soufrés volatiles (CSV) qui sont alors produits, comme le méthanethiol (MTL) (Demarigny et *al.*, 2000). Il a été montré que *G. candidum* possède un fort potentiel à dégrader la L-méthionine. Suite à cette dégradation des composés tels que le MTL, le sulfure de diméthyle (DMS), le disulfure de diméthyle (DMDS), le trisulfure de diméthyle (DMTS) sont alors produits. De plus *G. candidum* est le seul microorganisme à produire le S-méthyl-thioacétate (MTA) (Bonnarme et *al.*, 2001a; Bonnarme et *al.*, 2001b).

Il est alors intéressant de noter l'influence du précurseur de biosynthèse du composé soufré. Ainsi lorsque ce dernier est le S-méthylméthionine, le composé soufré majeur formé est le DMS. Alors que lorsque celui-ci est le L-méthionine, tous les composé cités plus haut sont formés (Spinnler et *al.*, 2001). Arfi et *al.* (2002) décrit la capacité des levures à générer des composés aromatiques volatiles. A l'inverse de *K. lactis* l'apport de MTL exogène ne se traduit pas par une augmentation de production de MTA chez *G. candidum*. Cependant, si *G. candidum* produit peu d'ester, il génère une grande quantité de CSV.

Geotrichum candidum est donc une levure d'intérêt pour la production de CSV dans les fromages. Cependant, comme nous l'avons vu, *G. candidum* est utilisée avec de nombreux microorganismes dans le processus de fabrication des fromages. Nous allons maintenant voir quelles sont les différents bénéfices de ces associations ou compétitions.

Lorsque que l'on fait entrer *G. candidum* et *Y. lipolytica* dans la flore, les fromages se distinguent particulièrement par leur forte intensité aromatique et leurs notes de Munster et d'ammoniac. L'association des deux levures augmente la production de DMTS (Martin et *al.*, 2001).

En fin d'affinage, *G. candidum* prépare la surface du fromage pour la colonisation par les bactéries sensibles à l'acidité comme *Brevibacterium sp*. (Corsetti et *al.*, 2001). Les métabolites produits par *G. candidum* peuvent aussi inhiber les *Listeria monocytogenes* (Dieuleveux et *al.*, 1997; Dieuleveux et *al.*, 1998), et sont capables d'inhiber la croissance ou la sporulation de *Mucor sp*. (Boutrou et Gueguen, 2005). La densité de la population de *G. candidum* à un effet sur la maturité de la croûte, une faible densité facilite les échanges de gaz dans la surface des fromages. *G. candidum* prédominant sur la croûte du fromage aide à déterminer la texture, la cohésion et l'épaisseur de la croûte. Dans certains fromages comme le St. Marcellin, *G. candidum* est responsable de l'apparence du fromage conférant une surface uniforme, blanche et « manteau » de velours (Gueguen, 1992). De manière générale, les souches de *G. candidum* sont utilisées, par les fromagers et l'industrie agro-alimentaire, comme « starter » de fermentation lors de la fabrication du fromage. Cela assure une

bonne maturation du fromage. *G. candidum* entre dans la composition des levains mixtes. Bachmann et *al.* (2003) montre que l'utilisation de *G. candidum* dans les levains permet de combattre le défaut de morge (la croûte orangée de certains fromages devient trop épaisse et visqueuse) (**Figure 9**).

A **B**

A: A gauche, fromage avec bonne surface avec le levain contenant *Geotrichum candidum*; à droite, fromage avec croute visqueuse avec levain de référence
B: Croissance de *Geotrichum candidum* sur croute lavée

Figure 9 : Mise en place d'un levain mixte pour limiter les défauts de morge d'après Bachmann et *al.* (2003)

3.4 *Geotrichum candidum* et santé

L'espèce *G. candidum* n'a pas été listée comme pathogène par l'administration française de sécurité biologique (Journal Officiel de la République Française, 1994) et elle n'apparaît pas non plus dans la liste officielle des agents biologiques publié par Advisory Committee on Dangerous Pathogens (ACDP) (2004). La consommation du fromage est la source majeure d'exposition à cette espèce. Il a été montré que même s'il a été consommé, il n'y a pas de colonisation du tube digestif. Les souches qui ont survécu sont retrouvées dans les selles. Rétrospectivement aucune maladie alimentaire n'a impliqué la consommation de produits contenant *G. candidum* (Pottier et *al.*, 2008). Opportuniste par excellence, on retrouve des souches cliniques *de G. candidum* probablement issus d'infections chez des individus immunodéprimés (Vasei et Imanieh, 1999). Considérée comme une mycose superficielle, les médecins parlent alors de « geotrichose ». On peut tout de fois mettre en doute les identifications comme cela a été le cas pour les cas d'infection où *D. hansenii* avait été identifiée à tort (Desnos-Ollivier et *al.*, 2008). Il est à noter que la plupart des identifications médicales sont effectuées par des non-taxonomistes selon les propriétés physiologiques et morphologiques bien moins performantes que les identifications moléculaires.

3.5 Les autres utilisations de *Geotrichum candidum*

Outre ses propriétés utiles pour les industriels laitiers, *G*. *candidum* a aussi été décrit comme un pathogène de plante. En effet, comme *Botrytis cinerea*, *G*. *candidum* à l'instar de *G. citri aurentii* peut provoquer la dégradation des parois cellulaires végétales par des enzymes pectolytiques produites (Barth et *al.*, 2009). Bien sur, il faut ici être encore une fois prudent quant aux confusions possibles et fréquentes entre *G. candidum* et *Gal. geotrichum*. Cependant, nous allons voir à présent que *G*. *candidum* peut avoir d'autres applications industrielles (production d'enzymes, dépollution…).

3.5.1 Production d'enzymes

Une endo beta xylanase a été purifiée et caractérisée chez *G. candidum* (Rodionova et *al.*, 2000). Les auteurs mettent alors en avant l'utilisation potentielle de cette enzyme dans le processus de biodégradation du bois. En 2000, une polygalacturonase (PG) extracellulaire a été caractérisée (Guessous et *al.*, 2000). Isolée à partir d'agrumes putréfiés, *Geotrichum candidum* montre une activité pectinolytique. La souche de *Geotrichum candidum* synthétise une polygalacturonase extracellulaire active sur la pectine et l'acide polygalacturonique. Récemment, une nouvelle PG a été caractérisée sur une culture sur le marc de raisin comme seul source de carbone (Illková et *al.*, 2012).

Enfin une peroxydase (DyP) a été caractérisée et produite chez *G. candidum* sur des mélasses comme source de carbone (Lee et *al.*, 2000; Sugano et *al.*, 1999). L'expression de cette enzyme hétérologue a ensuite été montrée et produite dans *Aspergillus oryzae* (Sugano et *al.*, 2001; Sugano et *al.*, 2000). Il existerait alors chez la levure *G. candidum* des enzymes de type peroxydase que l'on retrouve habituellement chez les champignons filamenteux. La présence de peroxydases peut avoir des débouchés écologiques, par exemple, dans des opérations de décoloration de mélasses (Kim et Shoda, 1999).

3.5.2 Lipases

Les lipases sont des sérines hydrolases définies comme triacylglycérol acylhydrolases (EC 3.1.1.3). Elles catalysent l'hydrolyse de la liaison ester de tri-, di- et mono-glycérides d'acides gras à longue chaîne en acides gras et glycérol. Elles diffèrent des estérases (EC 3.1.1.1) en raison de leur capacité à hydrolyser des triglycérides à l'interface lipide-eau (Sarda et Desnuelle, 1958). Les lipases appartiennent à la famille structurale super α / ß-hydrolases dont les activités reposent essentiellement sur une triade catalytique habituellement formée par les résidus sérine, histidine et asparagine (Carr et Ollis, 2009). L'hydrolyse du substrat est effectuée par la formation d'un intermédiaire tétraédrique. Dans des conditions thermodynamiques favorables (l'eau par exemple est de faible activité thermodynamique), elles sont capables de catalyser des réactions de synthèse telles que l'estérification ou l'amidation. Ceci est rendu possible grâce à leur résistance aux solvants

organiques. Elles sont très appréciées dans l'industrie parce qu'elles sont capables de catalyser des réactions chimio-sélectives, régio-sélectives et stéréo-sélectives.

Y. lipolytica est la levure qui contient le plus de lipases décrites (Fickers et *al.*, 2011). On dénombre 16 lipases de la famille « Génolevures » GL3R0084 (Sherman et *al.*, 2009) : *LIP2*, *LIP4*, *LIP5*, *LIP7*, *LIP8*, *LIP9*, *LIP10*, *LIP11*, *LIP12*, *LIP13*, *LIP14*, *LIP15*, *LIP16*, *LIP17*, *LIP18* et *LIP19* ainsi que 4 estérases de la famille « Génolevures » GL3C3695 (Sherman et *al.*, 2009) : *LIP1*, *LIP3*, *LIP6* et *LIP20*. Plusieurs lipases de *Geotrichum candidum* ont été décrites dans la littérature. Une lipase de *Geotrichum sp*. a été purifiée (Kamimura et *al.*, 2001). Celle-ci montre sous certaine condition une activité enzymatique stable après 12 heures de fermentation.

De façon plus anecdotique, il a été trouvé sur un compact disc en cours de dégradation une espèce du genre *Geotrichum* non décrite par les auteurs (Garcia-Guinea et *al.*, 2001) (**Figure 10**).

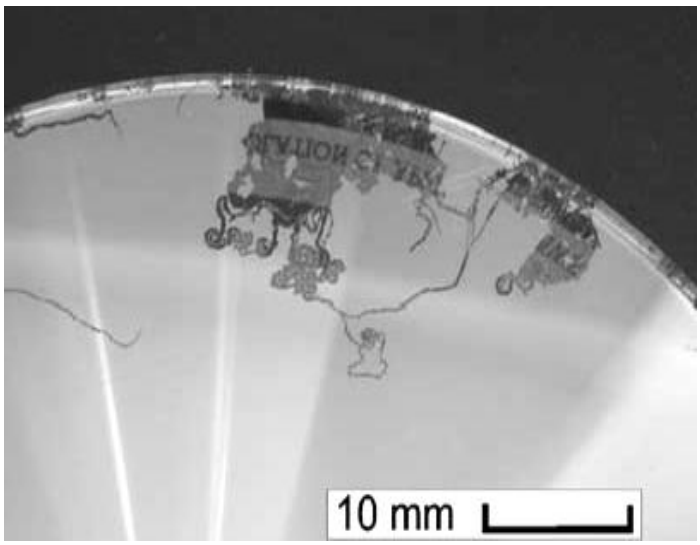


Figure 10 : Compact disque détérioré (Belize, Amérique centrale) (Garcia-Guinea et *al.*, 2001)

4 GENOMIQUE EVOLUTIVE DES LEVURES HEMIASCOMYCETES

En 1996 était publiée la première séquence complète d'un génome eucaryote : celui de *Saccharomyces cerevisiae* S288C (Goffeau et *al.*, 1996). Aujourd'hui, grâce à de nombreux travaux dont ceux effectués par le consortium Génolevures, les levures hémiascomycètes fournissent un ensemble de plus de 30 génomes entièrement ou partiellement séquencés (**Figure 11**) qui ont ouvert la voie aux approches de génomique comparative chez les eucaryotes. Depuis, le séquençage de nombreux autres organismes de ce phylum a pu être réalisé, créant une situation unique pour l'étude de l'évolution des eucaryotes.

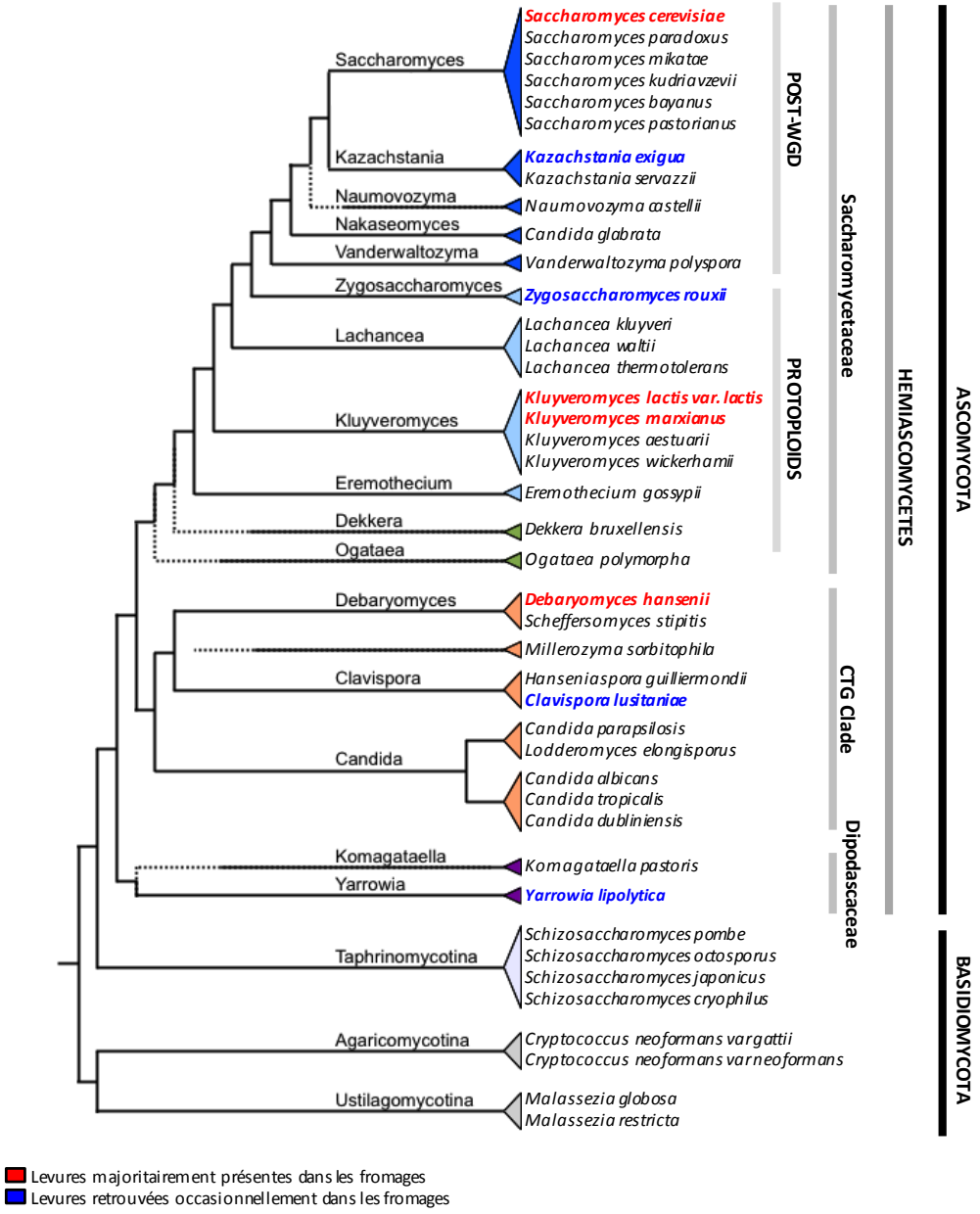


Figure 11 : Cladogramme des levures séquencées adapté de Dujon (2010)

4.1 Génomique comparée des levures hémiascomycètes

Plusieurs génomes de levures présentes dans les fromages ont été séquencés : *Kluyveromyces marxianus, Kluyveromyces lactis, Debaryomyces hansenii, Saccharomyces exiguus, Clavispora lusitaniae, Zygosaccharomyces rouxii, Yarrowia lipolytica* (**Figure 11**). Il était alors volontaire de séquencer des levures impliquées dans la composition des fromages. Une des levures majeures de fromage n'est cependant pas séquencée : *Geotrichum candidum*.

L'importante échelle évolutive couverte, en plus de propriétés génomiques de certains groupes ou certaines espèces, a permis de comprendre des modifications dans le contenu et dans l'architecture des génomes. Nous allons à présent traiter de l'aspect qui m'intéresse dans ce manuscrit, l'évolution des génomes de levure et en particulier leur contenu et leur structure.

4.1.1 Les génomes des Saccharomycotina : Généralités

Les génomes haploïdes des levures séquencées ont une taille qui varie d'environ 9 à 20 méga bases et contiennent un nombre de gènes codant des protéines variant de 4,700 à 6,500 (**Tableau 4**).

Règne	Sous-embranchement	Clade	Espèces	taille du génome (Mb)	Nb de Génes	% de géne intronique
Fungi	**Saccharomycotina**	post-WGD	*S.cerevisiae*	12,4	5762	5%
			C. glabrata	12,3	5205	2,4%
		Protoploïde	*Z. rouxii*	9,7	4996	3,3%
			L. kluyveri	11,3	5321	5,8%
			L. thermotolerans	10,4	5096	5,5%
			K. lactis var lactis	10,6	5078	3,4%
			E. gossypii	9,2	4728	4,6%
			D. bruxellensis	13,4	5600	2%
			H. polymorpha	9,78	5933	1,5%
		CTG	*D. hansenii*	12,2	6396	6,6%
			C. albicans	14,8	6354	6,1%
		Dipodascacae	*P. pastoris*	9,4	5313	11,4%
			Y. lipolytica	20	6588	14,9%
	Pezizomycotina		*P. chrysogenum*	32,2	12943	83,50%
			N. crassa	38	10082	80,30%

Tableau 4 : Exemple de données génomiques d'organismes séquencés (Dujon et *al.*, 2004; Galagan et *al.*, 2003; Mattanovich et *al.*, 2009; Neuveglise et *al.*, 2011; Ramezani-Rad et *al.*, 2003; Souciet et *al.*, 2009; van den Berg et *al.*, 2008; Woolfit et *al.*, 2007)

Chez les levures, les introns sont majoritairement présents en région 5' du gène. Les deux hypothèses résultantes sont alors : soit un gain d'intron en 5' soit une perte d'intron en 3'. Une des explications apportées est l'action d'une reverse transcriptase entrainant une perte des intron en région 3' (Cohen et *al.*, 2012).

Dans le **Tableau 4**, il est intéressant de remarquer que le nombre de gènes avec intron est beaucoup moins important chez les levures que chez les champignons filamenteux. Les levures auraient alors subit, au cours de leur évolution, des mécanismes de pertes massive d'introns.

Nous retrouvons dans l'ensemble des espèces de levures séquencées un nombre limité d'éléments mobiles appartenant à différentes familles, la plupart étant des transposons à LTR de classe I (**Tableau 5**).

Classification		**Hémiascomycètes**									
Ordre	Famille	Post-WGD		Protoploïdes		Clade CTG				Clade Dipodascacae	
		S. cerevisiae		*L. kluyveri*		*D. hansenii*		*C. albicans*		*Y. lipolityca*	
Classe I (retrotransposons)											
LTR	*copia*	*Ty* 1, 4	**(+++)**	*Tsk* 1	**(+)**			*Tca* 1, 4	**(+++)**	Ylt1	**(+++)**
		Ty5	**(+)**			Tdh5	**(+++)**				
		Ty 2	**(+++)**			*Tdh* 2	**(+++)**	Tca2	**(+++)**		
	gypsy	*Ty* 3	**(+)**	*Tsk* 3	**(+)**	*Tdh* 3	**(+)**	*Tca* 3	**(+++)**	*Tyl* 3, 6	**(+)**
LINE								*Zorro* 1, 2, 3	**(+)**	*Ylli*	**(+++)**
Classe II (Transposons à ADN)											
TIR	Tc1-Mariner									*Fotyl*	**(+)**
	hAT			*Rover*	**(+)**						
	Mutator							*Cmut* 1	**(+)**	*Mutyl*	**(+++)**

le nombre de + entre parenthèse représente le nombre de copie présent dans le gènome pour chaque famille: +: 1 copie; ++: 2 copies; +++:plus de 3 copies

Tableau 5 : Eléments transposables décrits chez les levures

Les éléments transposables, grâce à leur capacité à se déplacer dans les génomes et à générer des séquences répétées dispersées ont un rôle dans la structure et la plasticité des génomes de levures. Ils peuvent alors être utilisés dans des méthodes de typages (Typage par PCR inter LTR) dont nous parlerons dans la partie détaillant les différentes méthodes de typages.

L'analyse de divergence de la séquence de l'ensemble des protéines orthologues des différentes levures séquencées a révélé que la distance séparant l'évolution de *S. cerevisiae* et de *Y. lipolytica* est plus grande que celle observée pour l'ensemble phylum des chordés. Cela fait des levures des organismes privilégiés pour étudier la plasticité des génomes eucaryotes sur de longues périodes évolutives.

Nous verrons par la suite qu'il existe une grande variété de processus évolutif.

4.1.2 Sexualité et Mating type

Chez les levures, le croisement est caractérisé par la fusion de cellules haploïdes qui sont similaires en taille et en forme (isogamie), cela produit alors des zygotes (Knop, 2006). La fusion impliquant des cellules diploïdes est également possible, formant alors des cellules polyploïdes (Albertin et *al.*, 2009). Le croisement entre les deux cellules haploïdes est généralement suivi de la caryogamie, qui produit une cellule avec un noyau diploïde unique et donne naissance à un clone diploïde ou entre en méiose donnant alors quatre cellules filles haploïdes. Un retard dans le processus de caryogamie est alors possible, la descendance peut alors recevoir un seul des deux noyaux ayant cohabités un certain temps dans la même cellule ; cette situation offre une opportunité pour les deux parents d'échanger l'ADN mitochondrial mais aussi d'autres éléments génétiques autonomes, tels que des plasmides ou virus, sans échanger l'essentiel de leurs chromosomes.

Selon les espèces, l'accouplement peut se produire (i) entre cellules génétiquement identiques, on parle alors d'homothallisme ; (ii) ou nécessite deux cellules de signes sexuels différents, on parle alors d'hétérothallisme. Chez les levures, le signe sexuel est déterminé par un locus sur un chromosome unique (le locus *MAT* ou *MTL*). Ce locus existe sous deux formes idiomorphes, généralement dénommé *a* et *alpha* ou encore *A* et *B*.

Ce locus contient un ensemble limité de gènes qui contrôlent l'expression de nombreux autres gènes par différents mécanismes moléculaires. Même si ces gènes sont pratiquement les mêmes chez les levures, certains sont parfois absents. On pensait alors qu'il ne pouvait pas y avoir de méiose. Cependant, il a été montré qu'il existait des régulations différentes que chez *S. cerevisiae* (Butler et *al.*, 2004). Des altérations du locus MAT sont fréquemment observées dans les génomes de levures, cela a alors pour effet de bloquer le processus de méiose chez certains organismes (Reedy et *al.*, 2009). La **Figure 12** adaptée deFabre et *al.* (2005) montre que les données génomiques révèlent une grande variété de situations relatives à la sexualité, l'accouplement, le type de commutation, l'homothallisme, et l'évolution de la sexualité.

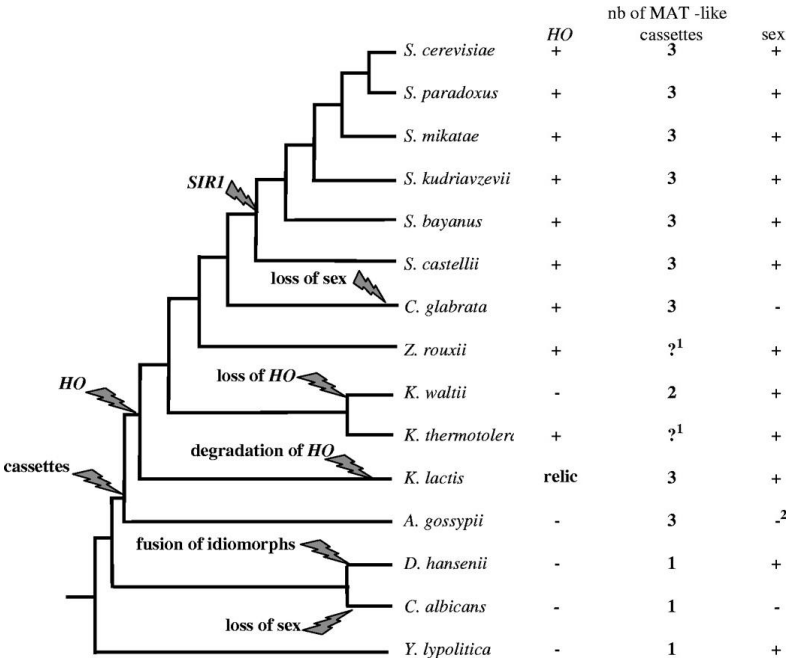


Figure 12 : Gène HO et cassette MAT dans les génomes de levures (Fabre et *al.*, 2005)

Comme nous pouvons le voir dans la **Figure 12** le système a évolué dans un processus en deux étapes dans lequel les cassettes silencieuses HMR / HML apparaissent après la scission de *Y. lipolytica* et le clade CTG, suivi de l'acquisition de l'endonucléase Ho très certainement à partir d'un élément génétique mobile (Butler et *al.*, 2004). Les deux éléments cités précédemment permettent une commutation de signe sexuel.

De plus, l'étude de la synténie du locus *MAT* chez les levures montre qu'il existe une conservation partielle de celle-ci (**Figure 13**). Nous pouvons remarquer la présence du gène *SLA2* jouxtant le locus chez les espèces *K. kluveri*, *K. lactis*, *P. angusta* ou *Y. lipolityca*, *BUD5* chez les espèces du clade WGD.

Dans certains cas, l'absence de sexualité peut être associée au possible pouvoir pathogène. Ainsi, la levure pathogène opportuniste *C. glabrata* semble avoir tous les éléments nécessaires pour que le croisement soit possible ainsi que la commutation de signe sexuel (**Figure 13**) (Butler et *al.*, 2004; Wong et *al.*, 2003).

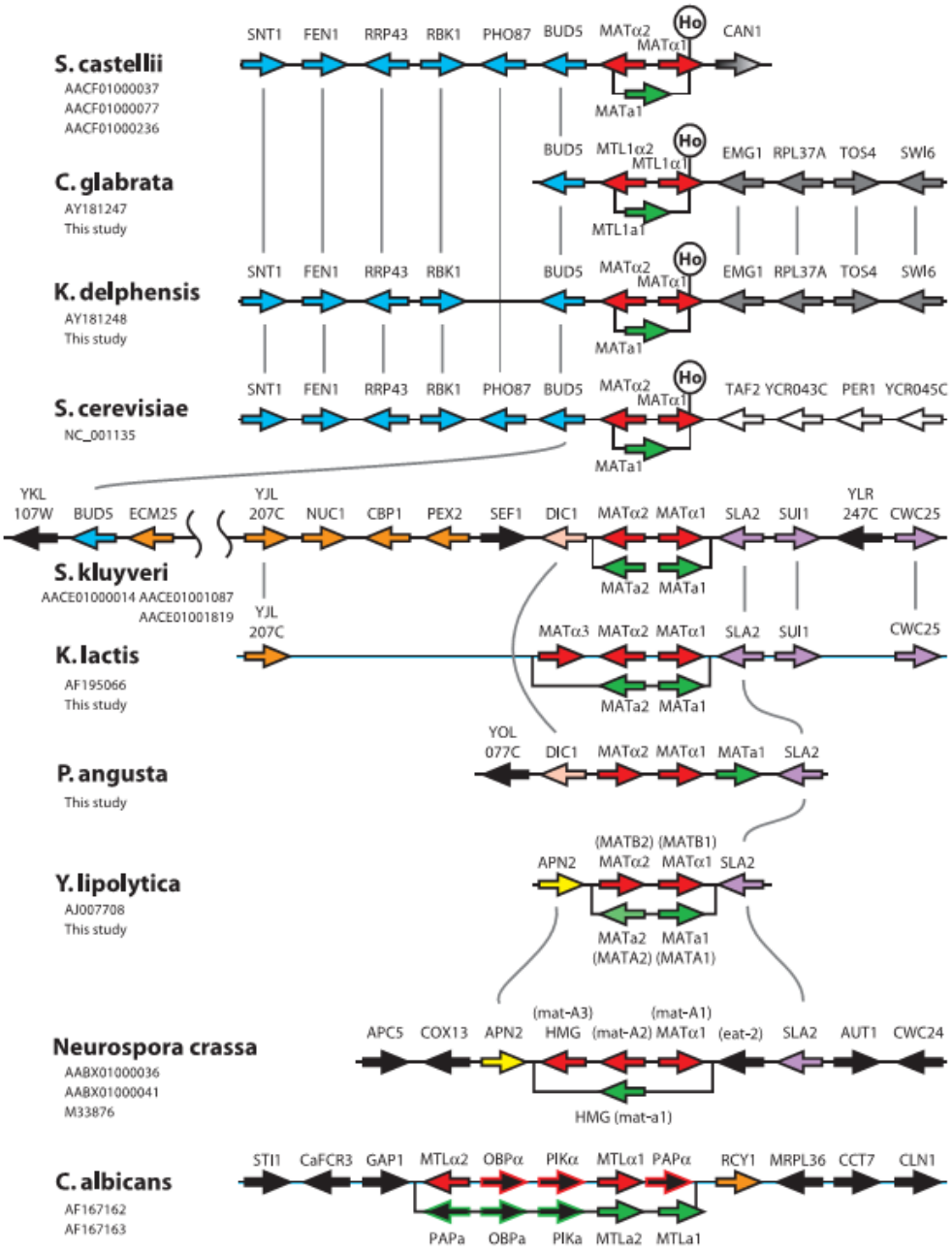


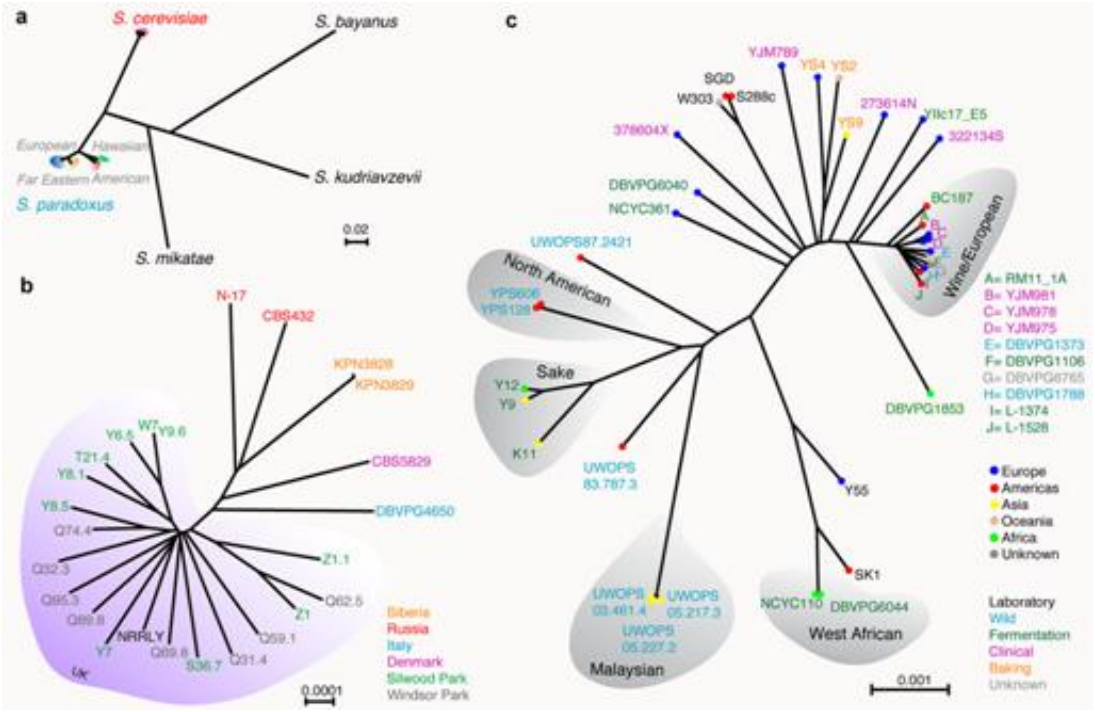
Figure 13 : Organisation des loci MAT dans 9 espèces de levure et du champignon filamenteux *Neurospora crassa* (Butler et *al.*, 2004)

Seules des cellules haploïdes ont été isolées et aucun croisement n'a encore pu être observé, il y a ici un déséquilibre vers un type sexuel. *C. glabrata*, comme beaucoup d'autres pathogènes fongiques, a perdu son mode de reproduction sexuelle, ce qui peut être corrélé avec l'absence d'éléments transposables dans ce génome (Wright et Finnegan, 2001). Cependant le contraire peut être observé chez *C. albicans* qui possède un grand nombre d'éléments transposables.

4.1.3 Génomique des populations

Le séquençage de génomes d'individus d'une même espèce a permis de mieux comprendre les processus évolutifs impliqués dans la variation de leurs séquences. Des études à grande échelle de la variation intraspécifique ont permis d'avoir une meilleure compréhension de la structure des populations mais également de l'histoire évolutive de ces espèces. De plus, les données générées représentent les fondations pour disséquer la relation existant entre génotypes et phénotypes.

La phylogénie basée sur le génome entier des *S. cerevisiae* (Liti et *al.*, 2009) (**Figure 14**) montre qu'il existe cinq lignées (parties grisées). Il s'agit de souches en provenance de Malaisie, d'Afrique de l'Ouest, de souches utilisées pour la fermentation du saké ou associées, d'Amérique du Nord, et un groupe important issu de sources mixtes contenant de nombreuses souches européennes et de souches utilisées pour la fermentation du vin. Les souches restantes sont sur les branches longues. Alors que certaines lignées correspondent à l'origine géographique, comme celles d'Amérique du Nord ou de Malaisie, de nombreuses souches étroitement apparentées sont issus de lieux géographiques très éloignés. Les auteurs estiment que cela pourrait être dû à la circulation humaine et la possibilité d'existence de croisements ultérieurs de ces souches.



a: souches de *S. cerevisiae* et*S. paradoxus séquencées*, *S. bayanus*, *S. mikatae* et *S. kudriazevii* utilisés en *"outgroup"*; **b:** souches de *S. paradoxus*, les souches issues du royaume-uni sont surlignées en violet ; **c:** souches de *S. cerevisiae*

Figure 14 : Phylogénomique des espèces du genre *Saccharomyces* (Liti et *al.*, 2009)

En comparant la population des souches de *S. paradoxus* et *S. cerevisiae*, il a été montré que contrairement aux souches de *S. cerevisiae*, les lignées de *S. paradoxus* sont bien limitées géographiquement.

Les auteurs nous donnent ici une théorie intéressante de l'évolution des levures domestiquées. Au vu de la structure de la population de *S. cerevisiae* qui se compose de quelques groupes bien définis géographiquement et de nombreuses lignées différentes voir issues de croisement entre lignées. L'idée que l'influence humaine a été l'occasion de croisements créant de nouvelles combinaisons de variation préexistantes est préférée à l'existence d'un ou deux évènements de domestication conduisant aux levures de *S. cerevisiae* actuelles.

De plus, il est intéressant de noter que les souches sauvages sont bien plus divergentes que les souches utilisées dans les différentes fermentations. Cela est confirmé par la récente étude sur des *S. cerevisiae* sauvages isolées de forêt primaire, qui sont très divergentes par rapport à toutes les souches étudiées jusqu'à présent (Wang et *al.*, 2012).

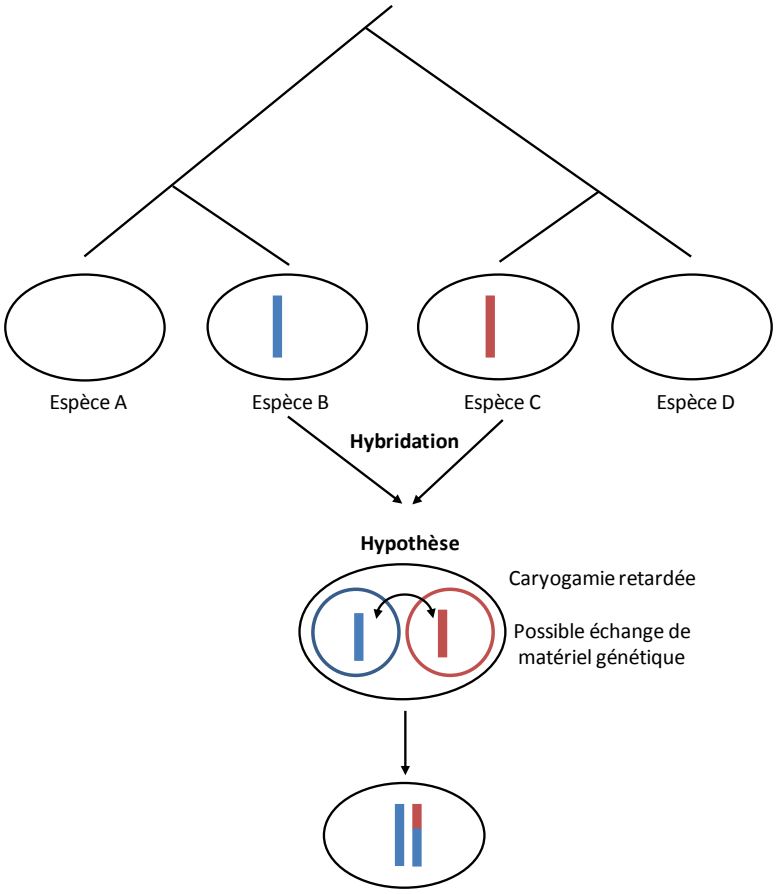
5 GENOMIQUE ET LEVURES INDUSTRIELLES

Les différentes études de comparaison de génomes permettent la mise en évidence d'une importante plasticité du matériel génétique. Celle-ci est traduite par l'existence de réarrangements chromosomiques comme les duplications (génomiques, segmentales, chromosomiques (aneuploïdie)) ou des délétions, inversions, translocations ou insertions. Peuvent également être retrouvées des acquisitions de gènes par le biais de transfert horizontaux ou d'introgressions suivant une hybridation interspécifique. Ces derniers événements peuvent être considérés comme des marqueurs d'adaptation à un environnement dans les levures étudiées par l'acquisition de nouveaux gènes, et par conséquent de nouvelles fonctions. Des discontinuités évolutives sont alors recherchées.

5.1 Hybridations chez les espèces du genre *Saccharomyces*

Le genre *Saccharomyces* est maintenant constitué de 8 espèces: *S. arboricolus, S. eubayanus, S. uvarum, S. cariocanus, S. cerevisiae, S. kudriavzevii, S. mikatae*, et *S. paradoxus* (Kurtzman, 2003; Wang et Bai, 2008) et *S. pastorianus, hybride entre S. eubayanus et S. cerevisiae (Casaregola et al., 2001; Rainieri et al., 2006).*

Dans la **Figure 15**, suite au croisement des souches B et C, la souche hybride va être constituée des génomes de ses deux parents. Chez les plantes, l'introgression est obtenue par hybridation suivi par rétrocroisements ou « *backcross* » successifs. Il est cependant peu probable que ce même processus s'applique chez les levures. Il est plus probable que des échanges de matériels génétiques existent d'un noyau d'un des parents au second, suite à une caryogamie retardée des deux noyaux parentaux. Une perte massive du matériel génétique d'un des parents non soumis à une pression de sélection lors de l'évolution peut alors être observée. La proximité écologique et les conditions extrêmes, haute concentration de sucre, faible concentration d'azote, forte concentration d'éthanol durant la fermentation expliqueraient alors les fréquentes introgressions retrouvées dans les souches industrielles de *S. cerevisiae* (Dujon, 2010). Nous ferons dans ce manuscrit la distinction entre un transfert horizontal direct de gène via une transformation et un transfert horizontal de gène du à un croisement inter spécifique ancestral.



L'arbre représente quatre espèces différentes A, B, C et D.
Les barres bleues représentes le génome de l espèce B; les rouges, celui de l'espèces C.
X signifie l'existence d'un croisement.

Figure 15 : Représentation schématique d'une introgression

5.1.1 Hybridation entre Saccharomyces spp.

Le premier hybride entre *Saccharomyces spp*. découvert est *S. pastorianus* (*S. carlsbergensis*) (Nilsson-Tillgren et *al.*, 1981). Cette levure, utilisée dans le brassage de bière, a été depuis très largement étudiée et séquencée en 2009 (Nakao et *al.*, 2009). Le génome de taille totale 25 Mb contient les deux génomes chromosomiques de *S. cerevisiae* et *S. eubayanus* et le génome mitochondrial de *S. eubayanus*. Les huit chromosomes de *S. pastorianus* présentent des translocations entre les génomes des deux parents. *S. pastorianus* est donc l'hybride des deux levures ayant une activité lors de fermentations industrielles, *S. cerevisiae* et *S. eubayanus.* Il apparait alors clairement l'avantage apporté par la sélection d'hybride. *S. cerevisiae* étant réputé pour ses qualités fermentaires, l'hybride formé est quand à lui capable de fermenter à basse température.

La combinaison de plusieurs méthodes : (i) analyse RFLP (Restriction Length Polymorphisms) de 35 régions géniques, hybridations génomiques sur puces d'ADN *S. cerevisiae*, (ii) analyse de ploïdie par cytométrie de flux, (ii) et détermination quantitative d'expression de gènes par des mesures de PCR quantitatives en temps réel, a permis de confirmer la présence de chromosomes chimériques et de définir les mécanismes impliqués dans leurs origines chez quatre hybrides naturels entre *S. cerevisiae* et *S. kudriavzevii* (Belloch et *al.*, 2009). Après l'hybridation, le génome hybride souffre de réarrangements chromosomiques aléatoires. Dans des conditions à fortes pressions de sélections comme durant la fermentation du vin sont alors sélectionnées les souches résistantes à l'épuisement des nutriments, au stress osmotique, à la température de fermentation et/ou à des niveaux croissants de l'éthanol. Les hybrides étudiés semblent avoir conversés les gènes de deux parents en fonction de la pression de sélection. Les résultats des hybridations sur puces « macro arrays » des souches œnologiques suggèrent que ces hybrides sont généralement diploïdes avec quelques exemples d'aneuploïdie chromosomique. Cependant la souche *S. cerevisiae* VIN7, récemment séquencée, montre que l'existence de souche non diploïde est aussi présente chez les souches œnologiques. En effet, le génome révèle que cette souche est une souche allotriploïde résultant de l'hybridation interspécifique entre *S. cerevisiae* et *S. kudriavzevii* (Borneman et *al.*, 2012). L'existence d'une souche hybride non diploïde montre que le croisement rare entre une souche de *S. cerevisiae* diploïde et une souche *S. kudriavzevii* haploïde est possible naturellement malgré cette difficulté d'obtention en laboratoire (Spencer et Spencer, 1996). L'obtention de triple hybride est possible après perte du chromosome contenant un des deux signes sexuels chez le diploïde, **Figure 16** (Pfliegler et *al.*, 2012).

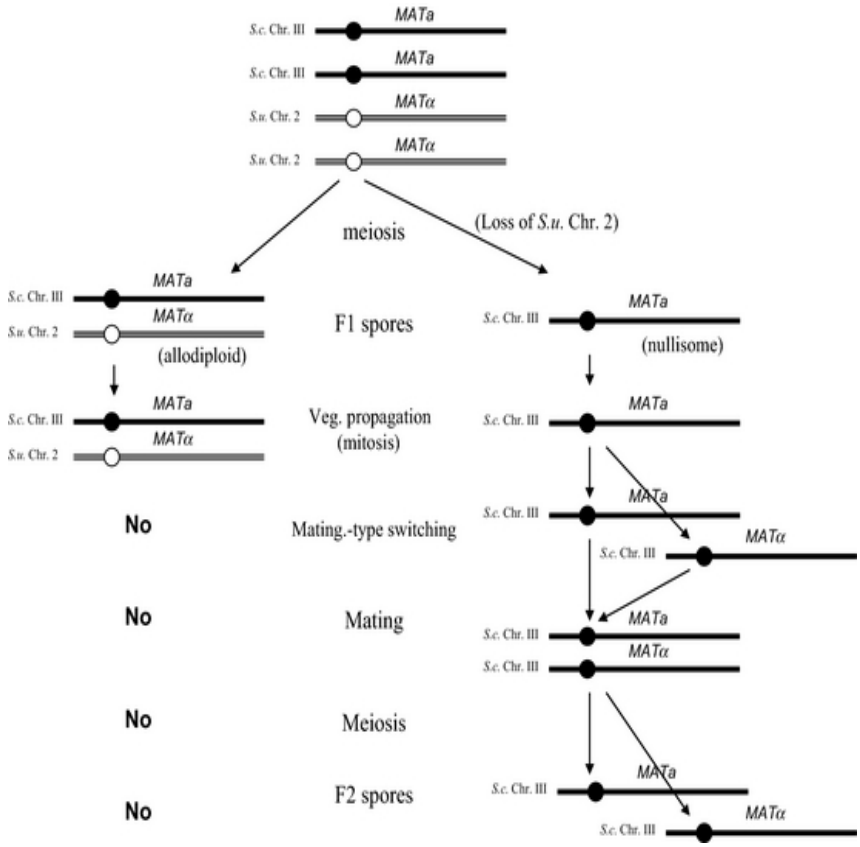


Figure 16 : Perte de la stérilité de l'alloploïde suite à la perte du chromosome contenant le locus *Matα* (Pfliegler et *al.*, 2012)

Les hybridations interspécifiques existent entre diverses espèces du genre *Saccharomyces, S. cerevisiae, S. bayanus, S. kudriavzevii.* Ces hybrides sont spécialisés dans divers processus industriels de fabrication des vins, bières et cidres. En guise d'exemple, les contributeurs de la levure impliquée dans la fermentation de la bière lager *S. pastorianus* sont *S. cerevisiae* et *S. eubayanus.* Dans l'arbre qui suit sont représentées en gris clair les relations phylogénétique entre les espèces du genre *Saccharomyces,* les espèces impliquées dans les processus industriels et / ou hybrides. Les produits de procédés industriels impliquant les hybrides et non-hybrides sont encadrés et les flèches correspondent alors à des hybrides (**Figure 17**) (Dequin et Casaregola, 2011).

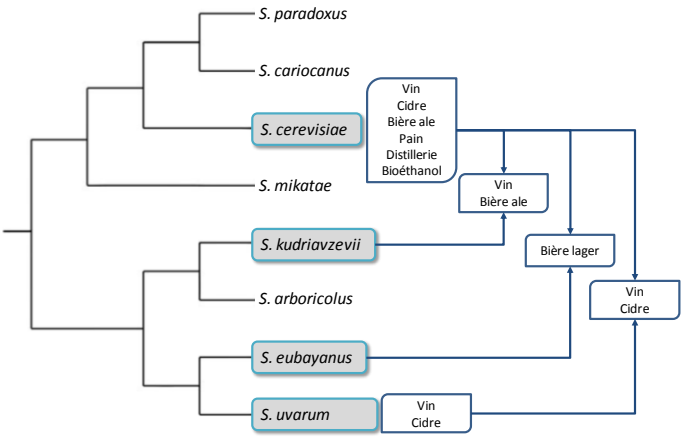


Figure 17 : Représentation schématique des relations phylogénétiques entre les espèces *Saccharomyces* et de leur spécialisation industrielle (Dequin et Casaregola, 2011)

5.1.2 Hybridation avec une espèce non Saccharomyces : Saccharomyces cerevisiae EC1118, Acquisition de gène par introgression

Le séquençage de *S. cerevisiae* EC1118 montre pour la première fois l'existence d'hybridation ancestrale entre *S. cerevisiae* et une espèce non *Saccharomyces*. Chez cette levure diploïde, trois grandes régions du génome (longueur totale : 120 kb) n'ont pas pu pas être alignées avec le génome de référence S288C (**Figure 18**).

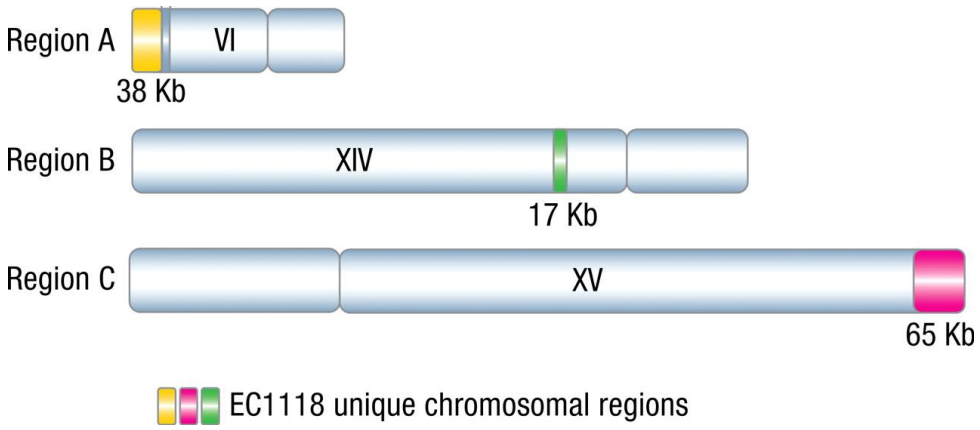


Figure 18 : Distribution chromosomique des 3 régions uniques EC1118 (Novo et *al.*, 2009)

Ces régions comportent 34 gènes dont certains sont impliqués dans des fonctions clés du processus de vinification, tels que le métabolisme du carbone et d'azote, le transport cellulaire, et la réponse au stress. Selon les auteurs, la région C en rose (**Figure 18**) semble être le résultat d'une introgression entre *S. cerevisiae* et une espèce du genre *Saccharomyces* non encore décrite. Il a notamment été montré l'existence d'introgression de *S. cerevisiae* dans *S. paradoxus* (Liti et *al.*, 2006). Les deux autres régions sont plus énigmatiques puisqu'elles prouvent pour la première fois l'existence d'une hybridation de *S. cerevisiae* avec une levure d'une espèce non *Saccharomyces*. La région verte est issue de *Zygosaccharomyces bailii*, contaminant majeur présent au cours de la fermentation. Ce fragment est de plus retrouvé plusieurs fois dans le génome à de point d'intégration différent. Il est aussi retrouvé chez d'autres souches de *S. cerevisiae*. Ce fragment possède une ARS active. Un plasmide est à l'origine de ce transfert horizontal (Galeote et *al.*, 2011). La région jaune pourrait être issue d'une espèce du clade CTG. Ici, les régions acquises sont issues de trois évènements de transferts horizontaux indépendants. Ils peuvent avoir eu lieu par transfert direct ou une nouvelle fois par introgression.

Si l'introgression est probable pour l'acquisition de la région rose, pour les deux autres il peut s'agir de transferts horizontaux.

5.1 Les transferts horizontaux de gènes (HGT) chez les champignons

Les transferts horizontaux sont définis comme un gain d'ADN provenant d'un autre organisme. Chez les procaryotes, trois mécanismes essentiels sont trouvés : la transduction, la conjugaison et la transformation bactérienne ; ils ne seront pas décrits dans ce manuscrit. Chez les eucaryotes les fragments d'ADN exogènes peuvent pénétrer dans la cellule par endocytose ou suite à une endosymbiose (Andersson, 2005; Kidwell, 1993).

Le gène *URA1* de *S. cerevisiae* provient par exemple d'un transfert horizontal de la bactérie *Lactococcus lactis* (Hall et *al.*, 2005). Ainsi, il existe des mécanismes par lesquels l'ancêtre de *S. cerevisiae* aurait pu prendre l'ADN exogène. L'ADN est ainsi retrouvé fréquemment près de télomères, et il est tentant de spéculer quant au rôle de la télomérase dans ce processus.

De nombreux transferts horizontaux ont pu être décrits chez les champignons. Initialement, une grande partie des événements de HGT chez les champignons décrits ont comme donateurs des organismes d'origines bactériennes (**Tableau 6**). Ce phénomène peut être dû au fait que les événements HGT bactériennes sont plus faciles à détecter que les transferts eucaryotes (Fitzpatrick, 2012).

Destinataire	Donneur	Chromosome/ Gène	Références
D. hansenii, *K. lactis*, *Y. lipolytica*	bactéries	14 gènes	(Dujon et *al.*, 2004)
Saccharomyces cerevisiae S288C	Bactéries	13 gènes	(Hall et *al.*, 2005; Hall and Dietrich, 2007)
Candida parapsilosis	Bactéries	Proline racemase et PhzF	(Fitzpatrick et *al.*, 2008)
Aspergillus clavatus	Champignon	Cluster de production de la polyketide synthase ACE1	(Khaldi et *al.*, 2008)
Saccharomycetaceae	Bactéries	11 gènes	(Rolland et *al.*, 2009)
Saccharomyces cerevisiae EC1118	Champignon	34 gènes	(Novo et *al.*, 2009)
60 espèces fongiques	Bactéries	713 gènes	(Marcet-Houben and Gabaldon, 2010)
Sordariomycetes et Saccharomycetes	Bactéries	Urea amidolase	(Strope et *al.*, 2011)
Aspergillus niger	Champignon	Cluster de production de la Fumonisine	(Khaldi and Wolfe, 2011)
S. cerevisiae S288C	*Wickerhamomyces sp.*	*ASP3* catabolisme de l'asparagine	(League et *al.*, 2012)

Tableau 6: Exemples de transfert horizontaux décrits chez les champignons

Selon certains auteurs, il existe des raisons biologiques qui peuvent expliquer pourquoi un transfert de gène bactérien à un champignon est plus probable qu'un transfert eucaryote à eucaryote. Les gènes eucaryotes possèdent des introns à la structure spécifique (Stajich et *al.*, 2007). Il est par ailleurs probable que le nombre et la diversité des populations bactériennes est considérablement plus grande que celle des populations eucaryotes, le nombre de gènes bactériens disponibles dans l'environnement est donc beaucoup plus important (Keeling et Palmer, 2008). Cependant, on peut s'interroger sur ces diverses raisons : le code génétique bactérien est différent, les gènes acquis ne sont plus sous l'influence de promoteurs… Il apparait cependant certain que lors des analyses de génome les transferts horizontaux d'origine bactérienne sont beaucoup plus faciles à trouver, et le manque de données génomiques de champignons peut empêcher l'identification d'un HGT champignon-champignon.

De plus l'architecture des gènes bactériens en opéron peut se traduire par le gain d'une voie métabolique complète par HGT. Si, le transfert d'une voie métabolique complète n'a pas encore été découvert, une analyse récente a indiqué que deux des six gènes (BIO3 et BIO4) de la voie de la biotine chez *S. cerevisiae* ont été acquis par HGT d'une source bactérienne (Hall et Dietrich, 2007).

Des analyses récentes ont commencé à localiser des HGT champignons à champignons (**Tableau 6**). Un certain nombre de ces études a découvert des preuves de transferts horizontaux d'un ensemble de gènes assurant des voies métaboliques fungiques complètes (Khaldi et *al.*, 2008; Khaldi et Wolfe, 2008, 2011; Mallet et *al.*, 2010; Slot et Hibbett, 2007; Slot et Rokas, 2011; Temporini et VanEtten, 2004) .

En guise d'exemple, Khaldi et Wolfe (2011) proposent que le cluster de 25 gènes permettant la production de la mycotoxine, la fumonisine a été transférée horizontalement d'une espèce de champignon Sordariomycete à Aspergillus niger.

Cependant, d'autres explications tout aussi satisfaisantes existent souvent pour résoudre les divergences entre la phylogénie d'un élément génétique en question et le génome de l'hôte (Rosewich et Kistler, 2000)(**Figure 19**). Il convient donc de se montrer vigilant avant toutes conclusions.

En effet, ces hypothèses alternatives incluent l'utilisation de phylogénie des espèces erronée, une comparaison inappropriée entre des séquences paralogues, une rétention sporadique d'un caractère partagé ancestral, des modifications de caractères dans différentes lignées, ou la présence d'une introgression suite à un événement d'hybridation. La discrimination entre HGT et solutions alternatives peut alors se révéler être un défi de taille.

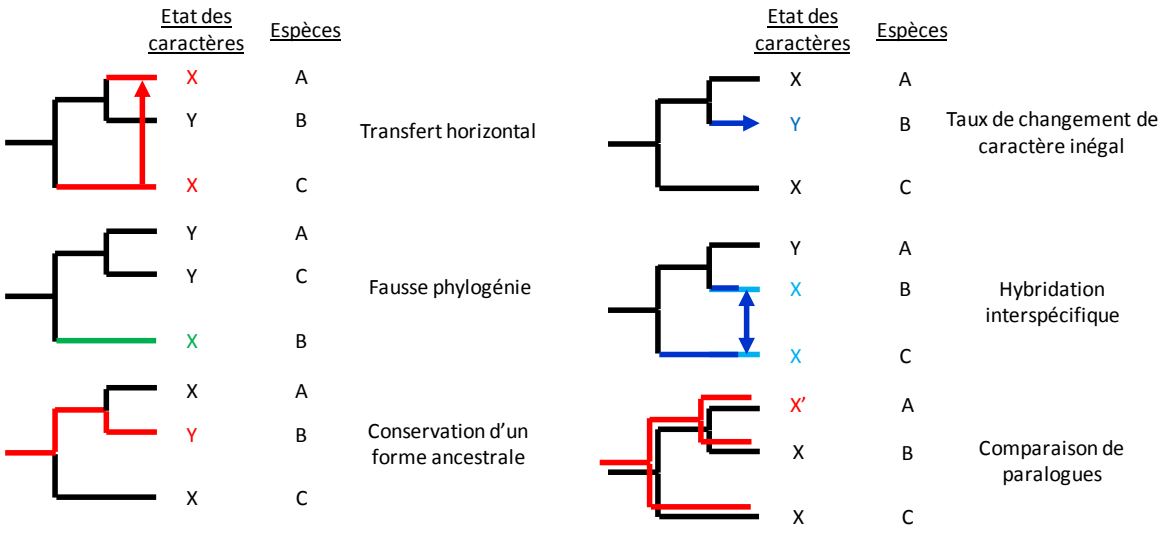


Figure 19 : Représentation schématique d'un transfert horizontal et autres discordances phylogéniques d'après Rosewich et Kistler (2000)

L'existence d'hybridation chez les *Saccharomyces spp.*, nous montre que l'acquisition de gène par introgression est naturelle et les réarrangements chromosomiques dans le génome d'un individu donné peuvent lui conférer un avantage sélectif dans certaines conditions environnementales. Des transferts horizontaux de gènes sont possibles de bactéries à champignons mais aussi de champignons à champignons. Nous allons à présent voir qu'il existe d'autres mécanismes d'adaptation chez la levure, les modifications de régulation transcriptionelle.

5.2 Modification de la régulation transcriptionelle chez les Saccharomyces

Dans une étude comparative des transcriptomes de *S. cerevisiae*, il a été constaté que SSU1 est un gène qui assure la médiation de l'efflux de sulfite chez *S. cerevisiae*. Il confère par conséquent, la résistance au sulfite (Park et Bakalinsky, 2000). L'expression de ce gène est significativement plus élevée dans la souche de levure de vin T73 que dans la souche de laboratoire S288C (Hauser et *al.*, 2001). De plus, une souche hautement résistante aux sulfites présents dans le vin présente une translocation impliquant la région promotrice du gène (allèle SSU1-R) (Goto-Yamamoto et *al.*, 1998).

Les translocations dont définies comme étant des échanges de fragments d'ADN entre deux chromosomes. Il existe deux types de types translocations : les translocations réciproques et les non réciproques. Dans ce dernier cas, une duplication d'un segment de chromosome remplace la délétion d'un fragment d'un autre chromosome.

Il a alors été étudié l'organisation du gène *SSU1* au niveau moléculaire dans différentes souches de levures de vin (Perez-Ortin et *al.*, 2002).

Le diagramme (**Figure 20**) représente la translocation réciproque échangeant le promoteur du gène *SSU1* (noir) avec celui du gène *ECM34* (gris), suivie de duplications segmentales augmentant le nombre d' « enhancer » de *SSU1.* Les souches portant ces séquences sont indiquées dans la **Figure 20**.

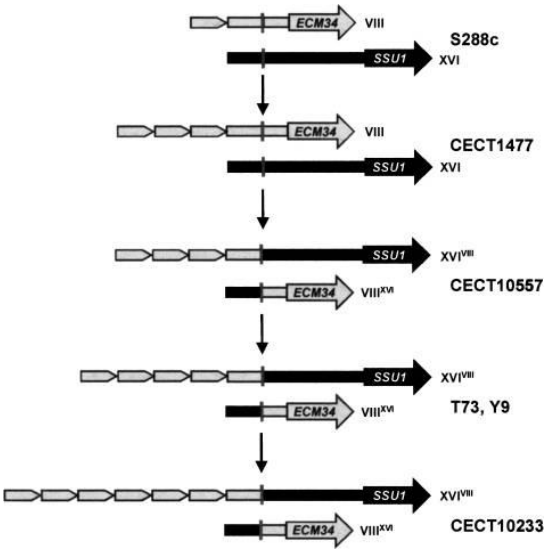


Figure 20 : Diagramme représentant l'organisation génique des gènes SSU1 et ECM34 chez différentes souches de *S. cerevisiae* (Perez-Ortin et *al.*, 2002)

La séquence promotrice de l'allèle *SSU1-R*, chromosome XVI, présente une similitude très élevée avec la séquence promotrice du gène *ECM34*, un gène de fonction inconnue du chromosome VIII. L'accumulation des enhanceurs sur le promoteur de *ECM34* en amont du gène SSU1 chez certaines souches va conférer une meilleure résistance aux sulfites.

Plusieurs translocations ont été montrées chez *S. cerevisiae*, la plupart d'entre elles sont médiées par un élément transposable *Ty* ou par de la recombinaison d'éléments subtélomériques Y (Casaregola et *al.*, 1998; Rachidi et *al.*, 1999). Cependant, peu ont été associées à des modifications fonctionnelles. Nous allons voir dans la partie qui suit que ces événements peuvent être à l'origine de duplication de gènes.

5.3 Duplications de gènes chez *S. cerevisiae*.

Les duplications de gènes jouent un rôle majeur dans l'évolution en fournissant des gènes paralogues qui peuvent acquérir des fonctions spécialisées au cours du temps (Ohno, 1970). Il existe plusieurs possibilités par lesquelles une séquence d'ADN peut se dupliquer dans un génome. La duplication peut être totale, chromosomique, segmentale ou génique.

- La **duplication totale** du génome est également appelée polyploïdisation. Comme nous l'avons vu précédemment, elle peut être le résultat d'une hybridation entre deux espèces différentes (on parle alors d'alloploïde) ou d'une même souche (on parle alors d'autoploïde). Dans ce dernier cas, les gènes dupliqués sont appelés ohnologue.

- La **duplication chromosomique** est le fait qu'un chromosome entier peut se dupliquer dans la cellule. Cette anomalie du nombre de chromosome peut aussi s'appeler aneuploïdie. Cette aberration correspond à un défaut de ségrégation de chromosomes lors de la mitose ou de la méiose.

- La **duplication segmentale** correspond à la duplication d'une série de gènes contigus. Elle peut être intrachromosomale ou interchromosomale. Dans le premier cas, on peut retrouver des duplications en tandem.

- Enfin, la **duplication génique**, où les éléments dupliqués peuvent se retrouver l'un à coté de l'autre, on parle alors de duplications en tandem. Ces duplications peuvent aussi être dispersées au sein du génome.

L'ancêtre de *S. cerevisiae* a subit une duplication ancestrale de son génome entier suivie d'une perte de gène (Kellis et *al.*, 2004; Wolfe et Shields, 1997). On retrouve dans son génome des événements de duplication génique conférant à *S. cerevisiae* une adaptation accrue à son environnement.

Des gènes présents en région subtélomérique montrent de fréquentes duplications au sein du génome de *S. cerevisiae*. Chez la levure, l'évolution a eu pour conséquence que la région est utilisée pour amplifier les gènes impliqués dans l'utilisation de sources de carbone. On retrouve par exemple : (i) trois gènes *MAL* formant un opéron codent pour une maltose perméase, une maltase et un trans-facteur ; (ii) le gène *MEL* qui confère la capacité à fermenter le mélibiose (Pryde et *al.*, 1997).

Ces souches peuvent être isolées à partir d'environnements très différents (Naumov et *al.*, 1995) et la plasticité de la région subtélomérique permet l'amplification de gènes utiles et la génération de variation entre ces gènes. Il est à noter que les régions télomériques ne sont pas toujours accessibles dans les différents génomes séquencés. On sait aujourd'hui que les génomes de *Y. lipolytica* et *D. hansenii* contiennent des éléments tel que les DnaK-like… Et que *K. lactis*, contient des gènes dupliqués en position subtélomèrique (Fairhead et Dujon, 2006).

S. cerevisiae possède une famille de gènes de 24 membres, les gènes *PAU*. Ici encore, ces gènes sont situés en position subtélomérique. Étant donné que *S. cerevisiae* est la principale espèce naturelle utilisée pour la fermentation du moût de raisin et la production de vin et que l'expression des gènes *PAU* est induite au cours de la fermentation du vin, il est proposé que l'expansion de familles des gènes *PAU* pourrait contribuer à une adaptation de *S. cerevisiae* au stress subi lors de la fermentation du vin (Luo et van Vuuren, 2009). Les analyses de la localisation chromosomique et de la synténie ont révélé que les gènes *PAU* auraient pu être amplifiés par des duplications segmentales, par des mécanismes de translocation chromosomique mais aussi par des recombinaisons associées au rétrotransposon *Ty*.

Par l'exemple des gènes *MAL*, il a été étudié l'évolution et la divergence fonctionnelle des familles de gènes subtélomériques dans les lignées de levure (Brown et *al.*, 2010). Cette étude montre que les familles de gènes en position subtélomérique ont une évolution et une expansion beaucoup plus rapide que les familles qui ne contiennent pas de gènes subtélomériques. Les événements de duplication sont fréquents et suivis par des allèles fonctionnels divergents permettant le métabolisme des glucides différents.

Dans la **Figure 21** sont présentés les principaux gènes décrits en position subtélomérique chez les espèces du genre *Saccharomyces.*

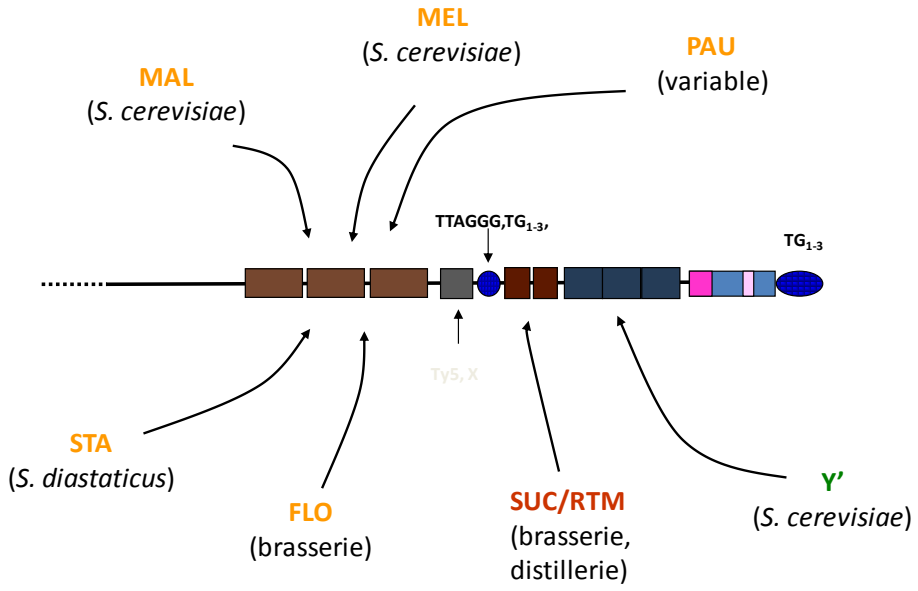


Figure 21 : Principaux gènes subtélomériques dupliqués chez les espèces du genre *Saccharomyces*

Le gène dupliqué est donc présent en deux copies. Généralement, l'une des copies est préservée tandis que l'autre est libre d'évoluer et d'accumuler des mutations. Ces mutations peuvent alors conduire à ce que l'on appelle une néofonctionnalisation du gène. De plus, il arrive que cet ensemble de mutation conduise à une mutation délétère du gène. La copie est alors inactive et l'on parle de pseudogénisation (Hughes, 1994; Lynch et Conery, 2000).

Toutefois, il arrive que la copie dupliquée ne mute pas et conserve la fonction originale. Ce processus nécessite alors certainement un intérêt sélectif pour la cellule. Il s'agit d'un avantage quantitatif qui découle du fait que la quantité d'ARN messager ou de protéine codée par ce gène augmente. On parle d'une « augmentation du dosage génique ». En plus des différents exemples cités précédemment, on retrouve des gènes ribosomiques dont la synthèse protéique est une des activités

importantes de la cellule. Chez *S. cerevisiae* sont présents les paralogues *RPL20A* et *RPL20B* (Koszul et *al.*, 2004).

Un des facteurs expliquant les duplications est la présence d'éléments transposables actifs (Friedman et Hughes, 2001). Il a été montré que ces duplications spontanées peuvent survenir chez *S. cerevisiae* par l'intermédiaire d'un mécanisme de rétroposition *via* le transposon LTR *Ty*1 (Schacherer et *al.*, 2004). Dispersés dans les génomes, les éléments répétés comme les retrotransposons de classe I (*Ty*) permettent d'introduire des duplications dépendamment de *RAD52* impliquées dans le mécanisme de BIR (*Break-Induced Replication*). Les duplications créées sont alors subtélomériques et peuvent être issues de translocation inter-chromosomique.

6 METHODES DE TYPAGES

6.1 Enjeux des méthodes de typage

Les études de typages permettent un suivi moléculaire d'un inoculum lors d'un processus industriel, une mise en collection raisonnée des souches, la gestion de propriétés intellectuelles, des études épidémiologiques, et d'obtenir une idée sur la biodiversité d'une espèce. En effet, l'industrie à tendance à standardiser les inocula et les conditions de maturité qui peuvent mener à la perte de cette dernière.

Le typage consiste en la caractérisation d'isolats par son analyse phénotypique et génotypique. Il peut être employé pour étudier l'évolution des espèces, la distribution des populations microbiennes et pour différencier les souches d'une espèce.

Un schéma de typage doit permettre de différencier deux souches génétiquement indépendantes et de déterminer avec certitude que des souches typées dans différents laboratoires sont similaires ou distinctes. Un typage idéal doit être une méthode très discriminante, reproductible, standardisée, universelle, facile d'application et applicable à toutes les souches d'une espèce.

6.2 Principales méthodes de génotypages

Il existe de nombreuses méthodes moléculaires dont les plus connues sont la PFGE, la RAPD, le Microsatellite typing ou la MLST. Elles permettent de typer les populations microbiennes. Ces méthodes sont très utilisées en épidémiologie et plusieurs études ont montré qu'elles permettent

d'identifier les souches bactériennes responsables d'infections à partir de marqueurs génétiques déterminés. Nous allons voir qu'elles sont aussi fréquemment utilisées chez les levures.

6.2.1 Pulsed Field Gel Electrophoresis (PFGE)

La PFGE a été développée par Schwartz et Cantor en 1984 afin de séparer les grandes molécules d'ADN (> 50 kb) que l'électrophorèse classique en gel d'agarose ne permet pas de résoudre (même en diminuant au maximum la concentration d'agarose).

Le principe de l'électrophorèse en champ pulsé consiste à alterner l'orientation du champ électrique au cours du temps. Chaque changement de champ électrique réoriente la molécule d'ADN dans le gel augmentant ainsi la probabilité que la molécule d'ADN soit orientée de façon à passer à travers les mailles du gel. Cette probabilité dépend de la taille de la molécule et la vitesse de migration d'un fragment d'ADN dans le gel varie dans le sens inverse de sa taille. L'électrophorèse en champ pulsé permet ainsi de séparer des fragments d'ADN d'une taille allant de moins de 1 kb à une dizaine de mégabases. Cette méthode permet d'analyser le caryotype et d'estimer la taille du génome.

Cette méthode est une méthode de choix pour l'étude des souches de *S. cerevisiae* issues du vin. Chez *G. candidum*, il a été mis en évidence une variabilité du nombre de chromosomes et un degré de polymorphisme élevé au sein des souches (Gente et *al.*, 2002a). Les profils PFGE ont montré un haut degré de polymorphisme ce qui indique, selon les auteurs, une grande variabilité entre les souches (**Figure** 22). Les auteurs remarquent alors que la taille du génome et la présence de grands chromosomes semble être corrélées avec le morphotype. Les souches présentant un morphotype moisissure ou intermédiaire ont tendance à avoir de plus grands génomes que les souches avec un morphotype de type levure. Les souches étudiées possèdent un nombre variable de 5 à 8 chromosomes. La taille du génome a été estimée entre 11 et 19 Mb.

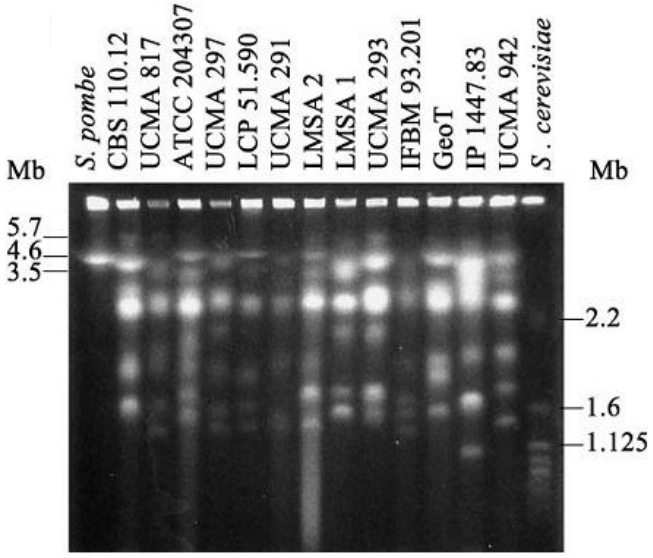


Figure 22 : Gel d'électrophorèse en champ pulsé de 13 souches de *G. candidum* (Gente et *al.*, 2002a)

Le polymorphisme des tailles de chromosomes n'est pas rare, Corredor et *al.* (2003) montre qu'il existe aussi un polymorphisme du aux réarrangements chromosomiques chez *D. hansenii.* Chez *G. candidum*, ce polymorphisme du nombre de chromosomes peut s'expliquer par l'existence de souches diploïdes hétérozygotes. Un nombre variable de chromosomes n'a jamais été montré chez les levures séquencées.

6.2.2 Random Amplified Polymorphic DNA (RAPD)

La technique RAPD-PCR a été mise au point en 1990 par Williams, Welsh et Mc Clelland (Welsh et McClelland, 1990; Williams et *al.*, 1990). Elle est basée sur la réaction d'amplification en chaîne (PCR) avec une amorce d'ADN génomique unique choisie au hasard. Elle constitue un moyen rapide pour réaliser des « screening » en génétique moléculaire. Cette méthode permet de différencier les variétés, les sérotypes et les types moléculaires des espèces et des genres. Depuis les années 1990, la RAPD a été utilisée pour caractériser les profils génétiques de plusieurs microorganismes impliqués dans des pathologies et pour des analyses de la diversité génétique de populations microbiennes (Sidrim et *al.*, 2010). Cette technique a été utilisée pour identifier les souches de levures dans les produits laitiers (Andrighetto et *al.*, 2000; Lopandic et *al.*, 2006). En l'utilisant, Andrighetto et *al.* (2000) est parvenu à identifier 42 des 46 isolats de levures issues du milieu fromager. Ainsi ont été identifiées les espèces *S. cerevisiae*, *K. marxianus, K. lactis, D. hansenii, Y. lipolytica* et *T. delbrueckii.*

Cette technique a aussi été utilisée pour typer un certain nombre de levures. Parmi elles, on retrouve *S. cerevisiae* (Baleiras Couto et *al.*, 1996), *Candida albicans* (Gyanchandani et *al.*, 1998), *Candida zeylanoides et D. hansenii* (Romano et *al.*, 1996) et *G. candidum* (Gente et *al.*, 2002b; Marcellino et *al.*, 2001). Marcellino et *al.* (2001) et Gente et *al.* (2002b) ont observé une très grande diversité des profils chez les souches de *G. candidum*. Dans le premier cas, elle semble être reliée au lieu d'isolation. Dans le second la diversité semble liée au lieu d'isolation ainsi qu'à la morphologie de *G. candidum*.

6.2.3 Typage par PCR inter LTR

Nous avons vu précédemment que les insertions d'éléments transposables (TE) varient en type et en nombre de copies selon les espèces. De plus, il a été montré que le contenu et la localisation des TE varient également au sein de souches de la même espèce. Ce polymorphisme procure une donnée supplémentaire sur les connaissances de la variabilité des souches au sein d'une même espèce. Cela

a été démontré chez *S. cerevisiae (Fink, 1986; Warmington et al., 1987)* et a permis l'établissement d'une nouvelle méthode de typage, la PCR interdelta (Ness et *al.*, 1993).

C'est la méthode de prédilection des industrielles pour typer les *S. cerevisiae* et est couramment utilisée.

Cette méthode a permis par exemple de suivre l'évolution des populations de *S. cerevisiae* durant la fermentation pour la fabrication de vin (Xufre et *al.*, 2011) mais aussi d'évaluer la diversité génétique de deux levures majeures des fromages, *Debaryomyces hansenii* et *Kluyveromyces marxianus* (Sohier et *al.*, 2009). La méthode appliquée sur 56 souches de *D. hansenii* et 61 souches de *K. marxianus* a confirmé la grande diversité génétique déjà observée pour *D. hansenii* et a révélé une grande diversité chez *K. marxianus*. La méthode a ainsi pu montrer la grande variabilité intraspécifique des espèces présentes dans le fromage, sans qu'il soit possible de corréler le typage avec l'origine géographique des isolats ou le type de fromage.

Cette méthode apporte un outil rapide et robuste pour étudier la biodiversité des espèces au sein d'un écosystème complexe mais peut aussi être utilisée pour contrôler l'apparition ou la disparition des souches au cours de la fermentation ou de l'affinage d'un fromage.

6.2.4 Microsatellite typing

Le microsatellite typing (MLP) repose sur la variabilité de la taille des séquences microsatellites. Ce sont des répétitions en tandem de 2 à 5 nucléotides. Comme le MLP teste la présence d'allèles différents à un locus donné, l'étude sur des hétérozygotes est possible. Plusieurs études ont déjà rapporté l'utilisation de cette technique pour le génotypage d'espèce fongique. Cette méthode a été utilisée pour typer les souches de *S. cerevisiae* (Legras et *al.*, 2005), elle révèle douze loci hypervariables permettant de discriminer les souches et montre les interactions génétiques entre elles. Legras et *al.* (2007) *via* une étude des 12 loci sur 651 souches. La **Figure 23** est l'arbre consensus obtenu à partir de l'étude microsatellite. 95 % des levures isolées du vin ont pour ancêtre commun une souche proche des souches libanaises. Les auteurs suggèrent une migration de la Mésopotamie vers l'Europe des souches de *S. cerevisiae* en corrélation avec la domestication de la vigne et la migration de la vigne.

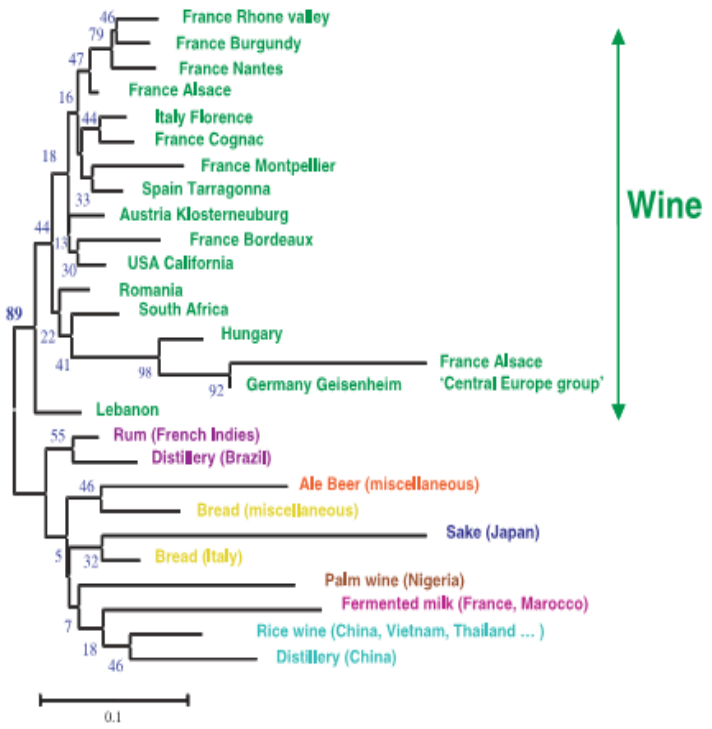


Figure 23 : Arbre phylogénétique consensus des populations de *S. cerevisiae* (Legras et *al.*, 2007)

Cette étude suggère alors, que la population de levures *S. cerevisiae* a été fortement influencée par la technologie humaine à travers l'histoire, la dérive génétique naturelle et la migration. Cela conduit à des populations progressivement différenciées.

6.2.5 Multilocus sequence typing

Le *MultiLocus Sequence Typing* (MLST) est une technique de typage moléculaire qui a été d'abord décrite pour une bactérie pathogène *Nesseria meningitidis* par (Maiden et *al.*, 1998). Cette approche a été développée dans le but de permettre l'identification précise des souches bactériennes ou fongiques pathogènes ou non au profit de la surveillance épidémiologique et dans l'intérêt de la santé publique et d'évaluer la biodiversité des souches. Elle est basée sur l'analyse directe de la séquence de plusieurs gènes de ménages (*housekeeping genes*). Ainsi un schéma MLST comprend 400 à 600 pb d'un groupe de cinq à dix gènes présents en une seule copie dans le génome. Les gènes de ménages sont des gènes dits constitutifs, très conservés durant l'évolution, qui assurent les fonctions indispensables à la vie de tous les types de cellules comme, par exemple, des gènes qui codent les protéines du cytosquelette (β-actine, α-tubuline), des enzymes de la voie du glucose (la glycéraldéhyde-3-phosphate déshydrogénase), etc. La technique est aujourd'hui très utilisée dans de nombreux domaines tels que cliniques, épidémiologiques ou agro alimentaires.

La MLST est la technique de typage la plus utilisée pour l'ensemble des études épidémiologiques et de structuration de population (Meyer et *al.*, 2009), elle a prouvé son pouvoir discriminant pour un grand nombre de pathogène fongique comme *Candida albicans* (Bougnoux et *al.*, 2002), *C. galbrata* (Dodgson et *al.*, 2003), *C. tropicalis* (Tavanti et *al.*, 2005), *C. dubliniensi* (McManus et *al.*, 2008), *C. krusei* (Jacobsen et *al.*, 2007), *Aspergillus fumigatus* (Bain et *al.*, 2007), et *Cryptococcus neoformans* (Meyer et *al.*, 2009), *C. gattii* (Feng et *al.*, 2008) mais aussi *Saccharomyces cerevisiae* (Ayoub et *al.*, 2006; Munoz et *al.*, 2009). Dans ces derniers cas, l'analyse MLST offre un une méthode de typage pour les souches de *S. cerevisiae* et permet d'étudier les relations génétiques existantes entre elles.

De plus, Ayoub et *al.* (2006) montre que la MLST permet de différencier les souches selon leur origine géographique, en regroupant les souches asiatiques d'une part et les souches industrielles ou libanaises d'autre part. Le schéma décrit ici, semble moins discriminant que le typage Microsatellite (92-97 % contre 99 %). Cela sera confirmé par l'analyse de génome complet (Liti et *al.*, 2006) et l'analyse Microsatellite (Legras et *al.*, 2007).

Enfin, le typage avec la technique MLST des souches de *C. galbrata* causant des infections nosocomiales mené par (Lott et *al.*, 2010) a mis en évidence des profils alléliques spécifiques de la localité géographique et de la période d'isolation des souches. Cela a conduit à la détermination de la structure des populations clonales et à la connaissance de l'évolution au cours du temps d'isolats pathogènes.

Les points forts de la méthode sont sa capacité à fournir des données solides pour étudier l'évolution et les distances génétiques. C'est une méthode reproductible et reconnue pour sa portabilité. En effet, la technique reproduite par des laboratoires différents fournit les même résultats et la plupart des schémas MLST publiés sont développés comme outils pour la large communauté scientifique en étant disponible en ligne dans des bases de données internationales http://www.mlst.net/ et http://pubmlst.org/.

Les méthodes de typage moléculaire ont ainsi apporté aux scientifiques, avec plus ou moins de succès, la possibilité de différencier les souches microbiennes pathogènes des souches environnementales et industrielles, ce qui permet la mise en collection raisonnée des souches.

RESULTATS

Chapitre 1

A multi-gene phylogeny of the genus *Galactomyces*/*Geotrichum* and the related genera *Dipodascus* and *Magnusiomyces*. Reinstatement of the genus *Geotrichum* Link

Guillaume Morel[1,2], Sandrine Mallet[1,2], Serge Casaregola[1,2]

[1]INRA UMR1319, Micalis Institute, CIRM-Levures, 78850 F-Thiverval-Grignon, France.

[2]AgroParisTech UMR1319, Micalis Institute, 78850 F-Thiverval-Grignon, France.

Corresponding author:serge.casaregola@grignon.inra.fr

Keywords: *Geotrichum candidum*, Saccharomycotina, Taxonomy, Arthroconidial yeast,

Abbreviations: Horizontal Gene Transfer (HGT)

The GenBank/EMBL/DDBJ accession numbers for sequences described in this article have been deposited to EMBL/GenBank and are listed in Table 1

Abstract

The taxonomy of the genera, *Geotrichum/Galactomyces*, *Dipodascus*, *Magnusomycete* and *Saprochaete* has undergone many changes since the description of *Geotrichum candidum* by Link in 1809, especially because the anamorphs and teleomorphs of the same species were associated to different taxa. Though the use of molecular methods in preference to biohemical and morphological methods has greatly improved yeast taxonomy in recent years, structural differences in rDNA and significant intra-specific variability of rDNA sequences have hampered the establishment of a robust taxonomy for this group of species. We chose to compare the sequences of protein coding genes, and in this way avoided the inconsistency met with ribosomal DNA analysis. Our multi-genic analysis of 41 species covering the entire yeast tree with four housekeeping genes, ACT1, RPB1, RPB2 and TEF1 led to a robust phylogeny with three distinct taxa: *Geotrichum/Galactomyces*, Dipodascus and *Magnusomycete/Saprochaete*, with Yarrowia lipolytica and *Arxula adeninivorans* having a basal position in the Saccharomycotina part of the tree. Whereas the genus Magnusiomyces was not affected by our study, the species *Galactomyces reessii* and *Geotrichum klebahnii* were found to belong to the *Dipodascus* clade.. All the other *Geotrichum* and *Galactomyces* species analyzed here grouped in a specific clade. Whereas the type strains *Galactomyces pseudocandidum* CBS 820.71^T and *Geotrichum vulgare* CBS 10073^T were found to have nearly identical sequences for the three markers, *ACT1*, *RPB1* and *TEF1*, the *RPB2* marker displayed a 29 bp divergence out of 635 bp (96.5 % identity) between the two stains, suggesting that the *RPB2* gene of one of the two strains originates from another uncharacterized species through a horizontal gene transfer event or that one of the strains analyzed is an inter-specific hybrid. In accordance with the recently accepted rule “one fungus, one name” of the amended article 59 of the International Code of Botanical Nomenclature, we propose to reinstate the genus *Geotrichum* Link in order to eliminate the confusion between the different names of the anamorphs and teleomorphs associated with this genus and to maintain the widely used name *Geotrichum candidum* for this important food and biotechnological species.

Introduction

Geotrichum candidum (teleomorph *Galactomyces candidus*), is a filamentous fungus-like yeast displaying high morphologic variability and high phenotypic diversity. It is an ubiquitous species, which can be found in a wide range of habitats such as plant tissues, silage, soil, milk, cheese, air and water (Kurtzman et *al.*, 2011; Pottier et *al.*, 2008). *G. candidum* is mainly known because it is an important component of the microflora of soft cheeses such as Camembert and semi-fresh goat's and ewe's milk cheese and it has been used as a starter (see for review Boutrou and Gueguen (2005)). Members of the genus *Galactomyces* are commonly isolated from soil, air, water, milk, silage, plant, tissues, digestive tract in humans and other mammals (Pottier et *al.*, 2008).

The taxonomic position of *G. candidum* has been heavily debated. Because of its fungal-like morphology and its peculiar taxonomic position, much confusion was associated to the naming of this species. Although It was classified as yeast by Kurtzman and Robnett (1995) and by the two major monographs of Kurtzman and Fell (1998) and of Barnett et *al.* (2000), it was classified as moulds (Wouters et *al.*, 2002). *G. candidum* and related species were considered as filamentous yeast-like fungi by de Hoog and Smith (2004). The genus *Geotrichum* was created by Link in 1809 (Link, 1809) for the only species of the genus *G. candidum.* Since, several synonyms have been attributed to this species: *Botrytis geotricha* (Link 1824), *Oidium lactis* or *Oospora lactis* (Wouters et *al.*, 2002), *Endomyces geotrichum* (Butler and Petersen, 1972), *Galactomyces geotrichum* (Smith et *al.*, 2000) *Dipodascus geotrichum* (Butler and Petersen, 1972), *Galactomyces candidus* (de Hoog and Smith, 2004). Although *G. candidum* was an imperfect yeast, it was given the genus name *Geotrichum,* instead of *Candida.* A teleomorphic genus name, *Endomyces,* for the anamorph *G. candidum* was proposed by Butler and Petersen, (1972). This species was further placed into the genus *Galactomyces* under the name *Galactomyces geotrichum* (Redhead and Malloch, 1977). This led to today's co-existence of a species with two genus names, *Geotrichum* and *Galactomyces*, like it is the case in ascomycetous and basidiomycetous fungi, including the confusion between the two distinct taxa *Geotrichum candidum* and *Galactomyces geotrichum* (Butler and Petersen, 1972). de Hoog et *al.* (1986) used a combination of phenotypic and nuclear genome analyses to revise the genus *Geotrichum* into an anmorphic *Geotrichum*, two teleomorphic genera *Galactomyces* and *Dipodascus*. Using D1/D2 sequence analysis, Kurtzman and Robnett (1995) have shown that *G. candidum* was associated to one of the two clades of *Dipodascus* Using DNA hybridization, four subgroups of *G. geotrichum*: *G. geotrichum* sensu stricto, *G. geotrichum* group A, B and were distinguished by Smith et *al.* (1995). A key using classical identification and GC% was presented to identify taxa of the genera: *Geotrichum*, *Galactomyces* and *Dipodascus* (Smith et *al.*, 2000). In

parallel, 18S ribosomal DNA was shown to be largely distinct between two groups of the genera *Dipodascus*/*Geotrichum*, analysis led to to a polyphyly of these two genera (Ueda-Nishimura and Mikata, 2000). DNA/DNA hybridization separated Dipodascus species into two groups plus additional species, namely, two isolates from *Dipodascus aggregatus* and one from *Dipodascus ovetensis*, which could not be satisfactorily classified (Smith and Poot, 2003). More recently, a taxonomic analysis of *Geotrichum* and related genera such as *Dipodascus, Magnusomycete* and *Spharochaete* based on rDNA 18S and ITS phylogeny proposed that the *G. geotrichum*/*G. candidum* complex contained four separate species (de Hoog and Smith, 2004): *Gal. geotrichum*, *Gal. candidus*, *Gal. pseudocandidum* and *G. europaeum*. In particular, the teleomorphic state of *G. candidum* was changed from *Gal. geotrichum* to *Gal. candidus,* since *Gal. geotrichum* species was distinguished from *G. geotrichum*. This work also reassigned a number of species to different genera, such as *Dipodascus sp*. which have been reclassified as *Saprochaete* or *Magnusiomyces sp*. The SSU analysis led to an incongruous sepatartion of these ensemble of species in two groups of species.

Recent work has shown that within the species *G. candidum*, ribosomal DNA sequence was highly variable between and even within isolates (Alper et *al.*, 2011), suggesting that this species carried distinct rDNA units. Recent work by B. Stielow and his collaborators suggested that, variability of ITS rDNA within and between species may be critical for the delineation of some species (Stielow et al, poster at the ICY 2012, Madison). Although very useful for delineating species, ribosomal DNA sequences, are not always able to resolve by and large inter-generic relationships, especially species belonging to complexes. Therefore, the current classification of *Geotrichum* and related taxa is only based on SSU and ITS, whereas in the Saccharomycotina subphylum, multi-gene analysis was shown to improve yeast classification by providing well-circumscribed taxa and robust phylogenies (Kurtzman and Robnett, 2003; Suh and Blackwell, 2006). Very recently and despite the sequence variability associated to ribosomal sequences, Groenewald et *al.* (2012) used a combination of 26S and *ACT1* cioding gene sequences to delineate *G. candidum* and closely related species. This work inferred that *Geotrichum sivilvicola* and *Geotrichum brindzae* were synonyms of *G. candidum* and *Geotrichum vulgare* was conspecific with *Gal. pseudocandidus*.

Since ribosomal DNA was bringing a number of inconsistencies on the phylogenetic position *G. candidum* with a number of of its related neighbours,we took advantage from the on-going sequencing of *G. candidum* to facilitate the amplification and the sequencing of widely used phylogenetic markers, anf thus to avoid the problems inherent to ribosomal DNA. We performed a multi-gene analysis on 41 species, which placed unambiguously *G. candidum* in the ascomycetous yeast tree. We further analyzed the phylogenetic relationship between several representatives of the *Geotrichum*, *Galatomyces*, *Dipodascus, Magnusomyces* and *Saprochaete* genera and showed that

previous phylogeny either based on phenotypic characteristics and morphology or based on short ITS sequence led to polyphyletic genera. We propose a reclassification of several species belonging to *Galatomyces*, *Dipodascus, Magnusomycetes* and *Saprochaete* genera.

Materials and methods

Strains used in this study

All strains used in this study are listed in Table 1. Cells were routinely grown in YPD medium (1% yeast extracts, 1% peptone, 1% glucose) at 28°C with shaking.

Genomic DNA extraction

Cells grown in 15 ml YPD medium overnight at 28°C were centrifuged at 2500 x g for 3 min, and the resulting pellet was washed twice in 750µl H_2O. The resulting pellet was resuspended in 200µl of lysis buffer (1% sodium dodecylsulfate, 2% triton, 100mM NaCl, 50mM EDTA, 50mM Tris, pH 8), 200µl chloroform phenol (pH=8) and 300mg glass beads. The suspension was supplemented with 200µl TE buffer and centrifuged for 5 min at 12000g. The aqueous portion was transferred to a new tube, and two chloroform extractions were carried out. DNA was precipitated with 1.5 vol. ethanol 100% and centrifuged for 4 min at 12000g. The pellet was rinsed with 400µl 70% ethanol, solubilised in 400µL TE and with 3µl RNAse (10mg/ml) for 30 min at 37 °C. After a second precipitation with Ethanol 100%, the pellet was dried and resuspended in 50µl TE buffer. The DNA concentration was quantified on agarose gel.

PCR amplification

Routinely, 2µl DNA (between 25 and 50ng) was added to 48µl PCR Reaction mix containing 0.8mM dNTP mixture, 0.1µM forward and reverse primers in the recommended buffer and 1 unit of TaKaRa Ex Taq®. Reactions were run on a 2720 Thermal cycler (Applied Biosystems). ITS1 and ITS4 region DNA was amplified with 600 nM primer ITS1 (5'-TCCGTAGGTGAACCTGCGG-3') and 600 nM ITS4 primers (5'-TCCTCCGCTTATTGATATGC-3')(Lott et *al.*, 1993), and the following thermocycler parameters 94°C for 5 min, followed by 30 cycles at 94°C for 30 s, 50°C for 30 s, and 72°C for 30 s, followed by one final extension at 72°C for 5 min. Similarly, 600 nM (each) NL-1 (5'-GCATATCAATAAGCGGAGGAAAAG-3') and NL-4 (5'-TCCGTGTTTCAAGACGG-3') (Kurtzman and Robnett, 1998) was used to amplify the

D1/D2 variable domain of the 26S ribosomal DNA (rDNA) gene, Rpb2_6F (5'-TGGGGKWTGGTYTGYCCTGC-3') and Rpb2_7R (5'-CCCATWGCYTGCTTMCCCAT-3') 2400 nM (each) was used to amplified the RNA polymerase II gene (Liu et *al.*, 1999). CA14 (5'-AACTGGGATGACATGGAGAAGATCTGGC-3') and CA5R (3'- GTGAACAATGGATGGACCAGATTCGTCG-5') 600 nM (each) were used to amplified the *ACT1* gene exon designed by Daniel and Meyer (2003). Finally, to amplify the gene encoding the first subunit of RNA polymerase II large subunit (Matheny et *al.*, 2002), the RPB1 2400 nM primers gRPB1-Afor (5'-GAKTGTCCKGGWCATTTTGG-3') and fRPB1-Crev (5'-CNGCDATNTCRTTRTCCATRTA-3') concentration have been used as followed: First denaturation 94°C 5min, amplification with 35 cycles 94°C 30 s, 50°C 40 s and 72°C 1min. A final elongation of 5 min at 72°C was added.

DNA sequence determination and phylogenetic analysis

PCR fragments were sequenced on both strands by Eurofins MWG Operon (Ebersberg, Germany) using primers that served for the PCR amplification. Sequences were assembled with the phred/phrap/consed package. Sequences were analyzed with the EMBOSS package and various programs in the EMBOSS environment (Rice et *al.*, 2000), including BLAST and FASTA. Sequence alignments were generated by using MUSCLE (Edgar, 2004) version 3.7 implemented in phylogeny.fr and were manually adjusted with Genedoc (http://www.psc.edu/biomed/genedoc).Phylogenetic trees were reconstructed with Phyml (Guindon and Gascuel, 2003) version 3.0 and neighbor-joining implemented in MEGA5 (Tamura et *al.*, 2011) Phylogenetic trees were visualized with NJplot (Perriere and Gouy, 1996).

Results

Use of the *ACT1* coding sequence to delineate the species *G. candidum*

As stated in the introduction, placement of the species studied here has changed so many times that, with the exception of *Geotrichum candidum,* the names appearing in Kurtzman et *al.* (2011) were used throughout this work. Several studies have pointed towards an important intra-specific variability of ribosomal DNA (Alper et *al.*, 2011; Groenewald et *al.*, 2012), therefore we have chosen to use only coding sequence. Considering the confusion concerning the *G. candidum* species, we first analyzed a total of 121 strains thought to belong to this species for their *ACT1* coding sequence to assess a possible intra-specific variability. A total of 51 strains were from the CIRM-Levures and the

rest, 68 strains, were provided by various cheese producers or starter producers. The *ACT1* sequences from these strains were compared to the sequence of the type strains of *G. candidus* CBS 178.71^{T} and *G. candidum* CBS 615.84^{T}. All strains presented a sequence varying by 4 bp at most with that of the type strain of *G. candidum* CBS 615.84^{T} indicating that they all belong to the species *G. geotrichum* (data not shown). These figures are well below the threshold to delineate species using the *ACT1* coding sequence defined by Daniel and Meyer (2003). Three of these strains showed an identical sequence to that of *G. sylvicola* type strain CBS 9194T. Groenewald et *al.* (2012) recently have shown that *G. candidum* and *G. sylvicola* were conspecific on the basis of the lack of divergence in the *ACT1* coding gene and the D1/D2 sequences.

Several studies pointed towards an important intra-specific variability at the sequence level using the random amplified micro-satellites technique and RAPD-PCR, (Gente et *al.*, 2002b) at the level of the chromosome length polymorphism (Gente et *al.*, 2002a) or at the level of ribosomal sequence (Alper et *al.*, 2011) To increase the sensitivity of our analysis, we sequenced the 5' end of the gene encoding the B-tubulin, which contains the more variable intronic sequences. Hardly any variability was observed in a total of 77 strains. Our conclusion from the analysis of the *ACT1* coding sequence and of the B-tubulin gene introns is in contradiction with the above quoted studies.

Our results are in sharp contrast with that observed by de Hoog and Smith (2004) and Alper et *al.* (2011). Both studies concluded in an important variation in the rDNA ITS and in the various rDNA part like D1D2, 26S and ITS, respectively. The rDNA variability could be attributed to the presence of various type of rDNA units with divergent sequences as shown for *Yarrowia lipolytica* (Fournier et *al.*, 1986). More recently, Stielow et al extended this results to the ITS of *G. candidum* as well as other species related to *Geotrichum candidum* (Stielow et al, in the press).

Phylogenetic relationship between *Geotrichum/Galactomyces*, *Dipodascus* and *Magnusiomyces* species

The phylogenetic relationship of a number of species of the *Geotrichum*/*Galactomyces*, *Dipodascus, Saprochaete* and *Magnusiomyces* was analyzed using the partial sequence of three housekeeping genes, the exon2 of the *ACT1* gene (Daniel and Meyer, 2003), the genes encoding RNA polymerase first largest subunit (*RPB1*) (Matheny et *al.*, 2002), RNA polymerase second largest subunit (*RPB2*) (Liu et *al.*, 1999) and the translation factor 1alpha (*TEF1*) in 21 species belonging the following genera *Galactomyces, Geotrichum, Dipodascascus, Magnusiomyces* and *Saprochaete*. The accession numbers of the four sequences are listed in Table 1. Several other markers like *MCM7* and *mtCOXII* were assayed, but could not be amplified for the totality of the strains analyzed here.

Interestingly, analysis of the sequences of *Gal. candidus* CBS 178.71^{T} revealed Single Polymorphism Nucleotides (SNPs), amounting to 3 bp and 8 bp out of 713 bp in *ACT1* exon 2 and 635 bp in *RPB2*. In addition, our analysis of genetic variability within the species extended to other markers confirmed this result (data not shown). This is consistent with this strain being diploid heterozygote, as suggested by Groenewald et *al.* (2012), who found that it was auto-fertile.

In order to better define the phylogenetic relationship between these genera we performed a phylogenetic analysis with 21 species of the above mentioned genera as well as 13 Saccharomycotina species, seven euascomycotes, an archeascomycete and a basidyomycete. The three concatenated sequences amounted to 2362 positions. *Coprinopsis cinerea sequences* were used as outgroup. The figure 1 presents the tree established with phyML algorithm with 100 replicates. Most of the branches of the tree shown in Figure 1 were well supported by high bootstrap values. However, we removed from the tree *D. albidus* and *D. geniculatus*, whose sequences affected the robustness of the tree. Unlike previous studies (de Hoog and Smith, 2004), all the species from the genera *Dipodascus*, *Magnusiomyces*, Sapprochete and *Galactomyces/Geotrichum* that we analyzed grouped together. The tree also indicated that *Y. lipolytica* and *B. adeninivorans* have a basal position in the Saccharomycotina subphylum. This was not observed in a recent phylogeny of the type species of 70 described genus (Kurtzman and Robnett, 2012), but these authors used a different set of genes and of species. However, they found that the type species of the genera *Dipodascus*, *Galactomyces* and *Magnusomyces* grouped together, in agreement with our analysis. The position of *D. albidus*, the type strain of the genus *Dipodascus* was unfortunately too weakly supported to be included in the tree, but other *Dipodascus* species were analyzed.

de Hoog and Smith (2004) deduced from a phylogenetic reconstruction with rDNA SSU that the species belonging to the genera of interest were split in two parts separated by the CTG clade, the *Saccharomyces* clade and other Saccharomycotina. Using coding sequences, we clearly show here that it is not the case. Our analysis distinguished three distinct clades (Figure 1). The three genera *Dipodascus*, *Magnusiomyces*/*Sapprochete* and *Galactomyces/Geotrichum* defined using SSU and ITS by de Hoog and Smith (2004) are all polyphyletic. This was also the case when ITS was used to generate a phylogeny of the species belonging to these genera as separated entities (de Hoog and Smith, 2011a, b, c, d). As expected we have a mixture of anamorphic and teleomorphic names in the same clade; this is true for *Geotrichum*/*Galactomyces* in the clade I and *Saprochate*/*Magnusomyces* in the clade III. Apart from the *Geotrichum sensu stricto* species comprising *G. candidum*, *G. geotrichum*, *G. citri-aurentii*, *G. pseudocandidum*, *G. vulgare*, *G. silvicola*, and *G. europeanum*, which all grouped together, all the other clades contained species attributed to *Geotrichum/Galactomyces*. This is the case of clade II in which *Gal. reesii* and *G. klebahnii* can be found. However the name

changes have been so frequent in this part of the Saccharomycota subphylum that, in our tree, each species has synonym belonging to any of the studied genera.

The clade 1 is well circumscribed with most branchings well supported by high bootstrap values and it contains all the *Geotrichum sensu stricto* species. Results are similar to that obtained recently by Groenewald et *al.* (2012) on the basis of the rDNA D12D2 and *ACT1* sequence analysis. *G. europaeum* is closely related to *G. candidum*. *Galactomyces citri-autrenti* has a basal position in the *Geotrichum/Galactomyces* clade. However these authors considered that *G. pseudocandidum* and *G. vulgare* were conspecific. In agreement with thee authors, we found that both species had nearly identical sequences for *ACT1*, *RPB1* and *TEF1* with 1 bp, 4 bp and 3 bp, respectively. However, we found that *RPB2* from both strains diverged by 29 bp (95% identity). This divergence is similar to that seen between well circumscribed species like *Galactomyces. geotrichum* CBS 772.71T and *G. candidum* CLB 918 which amounted to 32 bp.

Interestingly, the two species *Gal. pseudocandidus* and *G. vulgare* were introduced by de Hoog and Smith (2004) on the basis of DNA/DNA hybridization relatedness. However the relatedness between the two species was rather low as it amounted to 58%. A similar figure was obtained when *Saccharomyces pastorianus* was tested against other *Saccharomyces* species (Vaughan-Martini and Kurtzman, 1985). It was further shown that *S. pastorianus* was a hybrid and this could explain high DNA/DNA reassociation relatedness. The close to or complete identity between *ACT1, RPB1* and *TEF1*, or the high divergence between *RPB2* may be due to Horizontal Gene Transfer (HGT). These, and especially introgressions, have been found at high frequency in the genus *Saccharomyces* (Liti et *al.*, 2006; Novo et *al.*, 2009). *Gal. pseudocandidus* and *G. vulgare* are closely related and it is not unreasonable to think that they may have mated, and the extant strains analyzed here is the result of one or several rounds of hybridizations as observed in the genus *Saccharomyces* (Novo et *al.*, 2009) (Novo et al, 2012) or in the genus *Millerozyma* (Mallet et *al.*, 2012). The analysis of other strains belonging to these species may shed light on this discrepancy. So far, introgression and HGT between different yeast species were shown to involve several contiguous genes (Liti et al, 2006; Novo et al, 2009). We attempted to evidence similar sequence divergence in genes surrounding *RPB2* on the basis of a conserved syntheny between *G. candidum* and both species. We were able to amplify genes supposedly surrounding RPB2 but we found they were very similar in sequence (data not shown). This may indicate that RPB2 is the only gene transferred horizontally or that synteny is not conserved between G. candidum and the couple *Gal. pseudocandidus*/*G. vulgare.*

The second clade is also well circumscribed with high bootstrap values. This clade contains three *Dipodascus* species (*Dipodascus aggregatus*, *Dipodascus tetrasporeus* and *Dipodascus australiensis*), *Gal. reesii* and *G. klebahnii.* On the basis of rDNA SSU sequence comparison, these species belonged to the same group, but in our study, they are clearly distinct from the group *Geotrichum/Galactomyces*, which also belong to the ITS group I (de hoog and Smith, 2004). In a study analyzing rDNA SSU and D1/D2, *D. tetrasporeus* was placed near some *Geotrichum* species. This species was more loosely associated to *D. aggregatus* (Nagahama et *al.*, 2008). In the same study, another clade was formed by several *Geotrichum* species and *D. australiensis,* itself related to *Gal. geotrichum* and *Gal. citri--aurantii*.

Our data also place *Gal. reesii* in the genus *Dipodascus*, which was not observed by de Hoog and Smith (2004) on the basis of the short ITS sequence comparison. The latter analysis splits the *Dipodascus* species in two clades, separated by the genus Geotrichum. *G. klebhani* was associated to one of these two *Dipodascus* clades. Our study places together *G. klebhani*, *Gal . reesii* and *D. australiensis*.

Overall, results obtained by rDNA and protein coding sequence are not similar, since our data clearly separated *D. australiensis* from the *Geotrichum* clade. This discrepancy could be due to the peculiar sequence of rDNA in this group of species, as already discussed.

The third clade was entirely made of *Magnusomyces*/*Saprochete* species. This clade is quite distinct from the other two clades. In this case, similar results were observed with protein coding sequences and rDNA sequences, including the work by Nagahama et *al.* (2008) and the work by de Hoog and Smith (2004).

Interestingly, our data do not contradict completely the results obtained with rDNA. What gene coding sequence brings is the clear separation between the *Geotrichum* clade and the Dipodascus clade, which is accompagnied by the more robust placement of some species like *Gal. reesii* and *G. klebhani* in the *Dipodascus* clade.

Reinstatement of the genus *Geotrichum*

Geotrichum was initially proposed as the genus name catering for *G. candidum* by Link in 1809. There are many reasons for which we propose to reinstate this genus.

This species is very much used in biotechnology and agro-food industry under the name *Geotrichum candidum*. Thus, the name change introduced by has introduced an important confusion. In addition, the genus *Galactomyces* has been described as teleomorph names of *Dipodascus* species and

Geotrichum species (de Hoog and Smith, 2004). But for some species the two names are still used, increasing the misunderstanding surrounding. At the moment, the species name *Galactomyces candidus* is not recognized as a species by NCBI; in addition, *Galactomyces geotrichum* is considered as the teleomorph of *Geotrichum candidum* (*Taxonomy ID:* 27317), whereas they are quite distinct species (Fig. 1). As a consequence of which, for instance, the species was decribed, because this study did not include the type strain of *G. candidum*/*Gal. candidus* (CBS 178.71T). It was recently demonstrated that *G. bryndzae* CBS 11176T this species is conspecific with *G. candidum*. (Groenewald et *al.*, 2012).

The 18th International Botanical Congress, held in July 2011, adopted an amendment to Article 59 of the International Code of Botanical Nomenclature which abandons the dual nomenclature traditionally used in mycology to designate the sexual and asexual forms of a fungal species (Norvell, 2011). The "one fungus , one name" rules impose that one name has to be chosen for the genus *Geotrichum*/Galactomyces. The choice of the most recent genus described is recommended by the code; however, in the case of *Geotrichum*/*Galactomyces* the various reasons presented above argue strongly in favor of *Geotrichum*.

Proposed new species combinations for *Geotrichum*

Type strain of the genus: *Geotrichum candidum* Link.

1 *Geotrichum pseudocandidum* Morel and Casaregola comb. nov.

Basionym: Galactomyces pseudocandidus de Hoog & M.Th. Smith, Studies in Mycol. 50: 503, 2004 319

This species has for synonym: *Geotrichum vulgare* Wuczkowski, Bond & Prillinger, *Int.J.Syst.Evol.Microbiol.* 56:302 *2006*

2. Geotrichum pseudogalactomyces. Morel and Casaregola comb. nov.

Basionym: Endomyces geotrichum E.E. Butler & L.J. Petersen, Mycologia 64: 367. 1972.

Proposed new species combinations for *Dipodascus*

1. Reinstatement of *Dipodascus reessii* (van der Walt) von Arx

Basionym: Endomyces reessii Walt J.P.van der 1959 Antonie van Leeuwenhoek 25

This species has for synonym: *Galactomyces reessii* Redhead S.A. et *al.* 1977 Canad. J. Bot. 55(13)

2. *Dipodascus klebahnii* (Stautz) Morel and Casaregola, comb. nov.

Basionym: *Oospora klebahnii* Morenz J. 1963 Mykol. Schriftenr. 30

This species has for synonym: *Geotrichum klebahnii* (Stautz) Morenz, Mykologische Schriftenreihe, 30, 1960

Magnusomyces,

Folowing the discovery of sexuality in some species of the genus Saprochaete, the name *Magnusiomyces* was given to the teleomorphic species of the genus *Saprochaete* by de Hoog and Smith (2004). Our analysis as well as other analysis from others (de Hoog and Smith, 2004, 2011a, b, c, d; Nagahama et *al.*, 2008) indicates that *Magnusomyces* and *Saprochaete* species form a well circumscribed clade.

The 18th International Botanical Congress, held in July 2011, adopted an amendment to Article 59 of the International Code of Botanical Nomenclature, that the concept of sexuality was not valid for the definition of species. We therefore propose that the species belonging to the genus *Saprochaete* studied here be now part of the genus *Magnusomyces*.

Proposed new species combinations for *Magnusomyces*

1. *Magnusomyces* fragrans (Boekhout) Morel and Casaregola, comb. nov.

Basionym : *Oospora fragrans* Berkhout, C.M. 1923. De schimmelgeslachten Monilia, Oidium, Oospora en Torula. :1-71

This species has for synonym *Oidium suaveolens* Krzemecki 1913 Zentbl. Bakt. ParasitKde Abt. 2 *Geotrichum fragrans* Morenz J. 1963 Mykol. Schriftenr. 30, *Saprochaete suaveolens* (Krzemecki) de Hoog & M.T. Sm., Studies in Mycology, 50(2):508, 2004

Type : CBS 152.25 (CLB 1387)

2. *Magnusomyces suaveolens* (Krzemecki) Morel and Casaregola, comb. nov.

Basionym*:. Oidium suaveolens* Krzemecki, Zentbl. Bakt. ParasitKde, Abt. 2, 38: 577. 1913

This species has for synonym: *Saprochaete clavata* (de Hoog, M.Th. Smith & Guého) De Hoog and Smith, Studies in mycology 50: 489–515. 2004

Type: CBS 425.71 (ex-holotype) (=CLIB XXX), human lung tissue,U.S.A.

3. *Magnusomyces vini* (van der Walt & van Kerken) Morel and Casaregola comb. nov.

Basionym: *Candida ingens* van der Walt & van Kerken. Antonie van Leeuwenhoek 27: 285. 1961

This species has the following synonyms: *Geotrichum ingens* (van der Walt & van Kerken) de Hoog, M.Th. Smith & Guého, Mycotaxon 63: 346. 1997; *Pichia humboldtii* Rodrigues de Miranda & Török, Antonie van Leeuwenhoek 42: 343. 1976 and *Saprochaete ingens* (van der Walt &

van Kerken) De Hoog and Smith, Studies in mycology 50: 489–515. 2004

Type: CBS 517.90 (ex-holotype), wine cellar, South Africa.

Acknowledgements

We thank Christelle Louis-Mondésir for expert technical assistance. We also thank the *Centre National Interprofessionnel de l'Economie Laitière* (CNIEL) and the *Syndicat Professionnel des Producteurs d'Auxiliaires pour l'Industrie Laitière* (SPPAIL) for providing us with strains isolated from the cheese environment. This work received funding from the Agence National pour la Recherche grant “Food Microbiomes” (ANR-08-ALIA-007-02). GM was supported by a CIFRE fellowship with CNIEL.

References

Alper, I., Frenette, M., and Labrie, S. (2011). Ribosomal DNA polymorphisms in the yeast *Geotrichum candidum*. Fungal biology *115*, 1259-1269.

Barnett, J.A., Payne, R.W., Yarrow, D., and Barnett, L. (2000). Yeasts: Characteristics and Identification (Cambridge University Press).

Boutrou, R., and Gueguen, M. (2005). Interests in *Geotrichum candidum* for cheese technology. International journal of food microbiology *102*, 1-20.

Butler, E.E., and Petersen, L.J. (1972). *Endomyces geotrichum* a perfect state of *Geotrichum candidum*. Mycologia *64*, 365-374.

Daniel, H.M., and Meyer, W. (2003). Evaluation of ribosomal RNA and actin gene sequences for the identification of ascomycetous yeasts. International journal of food microbiology *86*, 61-78.

de Hoog, G.S., and Smith, M.T. (2004). Ribosomal gene phylogeny and species delimitation in *Geotrichum* and its teleomorphs. Studies in Mycology *50*, 489–516.

de Hoog, G.S., and Smith, M.T. (2011a). Chapter 27 - *Dipodascus*. In The Yeasts (Fifth Edition) (London: Elsevier), pp. 385-392.

de Hoog, G.S., and Smith, M.T. (2011b). Chapter 31 - *Galactomyces* Redhead & Malloch (1977). In The Yeasts (Fifth Edition) (London: Elsevier), pp. 413-420.

de Hoog, G.S., and Smith, M.T. (2011c). Chapter 45 - *Magnusiomyces* Zender (1977). In The Yeasts (Fifth Edition) (London: Elsevier), pp. 565-574.

de Hoog, G.S., and Smith, M.T. (2011d). Chapter 91 - *Geotrichum* Link. In The Yeasts (Fifth Edition) (London: Elsevier), pp. 1279-1286.

de Hoog, G.S., Smith, T., and Guého, E. (1986). A Revision of the Genus *Geotrichum* and Its Telemorphs (Centraalbureau voar Schimmelcultures).

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

Fournier, P., Gaillardin, C., Persuy, M.-A., Klootwijk, J., and Heerikhuizen, H.v. (1986). Heterogeneity in the ribosomal family of the yeast *Yarrowia lipolytica*: genomic organization and segregation studies. Gene *42*, 273-282.

Gente, S., Desmasures, N., Jacopin, C., Plessis, G., Beliard, M., Panoff, J.M., and Gueguen, M. (2002a). Intra-species chromosome-length polymorphism in *Geotrichum candidum* revealed by pulsed field gel electrophoresis. International journal of food microbiology *76*, 127-134.

Gente, S., Desmasures, N., Panoff, J.M., and Gueguen, M. (2002b). Genetic diversity among *Geotrichum candidum* strains from various substrates studied using RAM and RAPD-PCR. Journal of applied microbiology *92*, 491-501.

Groenewald, M., Coutinho, T., Smith, M.T., and van der Walt, J.P. (2012). Species reassignment of *Geotrichum bryndzae*, *Geotrichum phurueaensis*, *Geotrichum silvicola* and *Geotrichum vulgare* based on phylogenetic analyses and mating compatibility. International journal of systematic and evolutionary microbiology.

Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic biology *52*, 696-704.

Kurtzman, C.P., and Fell, J.W. (1998). The Yeasts - A Taxonomic Study (Elsevier Science).

Kurtzman, C.P., Fell, J.W., and Boekhout, T. (2011). The Yeasts: A Taxonomic Study (Elsevier).

Kurtzman, C.P., and Robnett, C.J. (1995). Molecular relationships among hyphal ascomycetous yeasts and yeastlike taxa. Canadian Journal of Botany *73*, 824-830.

Kurtzman, C.P., and Robnett, C.J. (1998). Identification and phylogeny of ascomycetous yeasts from analysis of nuclear large subunit (26S) ribosomal DNA partial sequences. Antonie van Leeuwenhoek *73*, 331-371.

Kurtzman, C.P., and Robnett, C.J. (2003). Phylogenetic relationships among yeasts of the '*Saccharomyces* complex' determined from multigene sequence analyses. FEMS yeast research *3*, 417-432.

Kurtzman, C.P., and Robnett, C.J. (2012). Relationships Among Genera of the Saccharomycotina (Ascomycota) from Multigene Phylogenetic Analysis of Type Species. FEMS yeast research *17*, 1567-1364.

Link, H.F. (1809). *Observationes in ordines plantarum naturales*. Dissertatio prima Mag Ges *Naturf. Freunde, Berlin.*, 3: 3-42**.**

Liti, G., Barton, D.B., and Louis, E.J. (2006). Sequence diversity, reproductive isolation and species concepts in *Saccharomyces*. Genetics *174*, 839-850.

Liu, Y.J., Whelen, S., and Hall, B.D. (1999). Phylogenetic relationships among ascomycetes: evidence from an RNA polymerse II subunit. Molecular biology and evolution *16*, 1799-1808.

Lott, T.J., Kuykendall, R.J., and Reiss, E. (1993). Nucleotide sequence analysis of the 5.8S rDNA and adjacent ITS2 region of *Candida albicans* and related species. Yeast *9*, 1199-1206.

Mallet, S., Weiss, S., Jacques, N., Leh-Louis, V., Sacerdot, C., and Casaregola, S. (2012). Insights into the life cycle of yeasts from the CTG clade revealed by the analysis of the *Millerozyma* (*Pichia*) *farinosa* species complex. PloS one *7*, e35842.

Matheny, P.B., Liu, Y.J., Ammirati, J.F., and Hall, B.D. (2002). Using RPB1 sequences to improve phylogenetic inference among mushrooms (Inocybe, Agaricales). Am J Bot *89*, 688-698.

Nagahama, T., Abdel-Wahab, M.A., Nogi, Y., Miyazaki, M., Uematsu, K., Hamamoto, M., and Horikoshi, K. (2008). *Dipodascus tetrasporeus* sp. nov., an ascosporogenous yeast isolated from deep-sea sediments in the Japan Trench. International journal of systematic and evolutionary microbiology *58*, 1040-1046.

Norvell, L.L. (2011). Fungal nomenclature. 1. Melbourne approves a new Code. Mycotaxon *116*, 481-490.

Novo, M., Bigey, F., Beyne, E., Galeote, V., Gavory, F., Mallet, S., Cambon, B., Legras, J.L., Wincker, P., Casaregola, S.*, et al.* (2009). Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast *Saccharomyces cerevisiae* EC1118. Proceedings of the National Academy of Sciences of the United States of America *106*, 16333-16338.

Perriere, G., and Gouy, M. (1996). WWW-query: an on-line retrieval system for biological sequence banks. Biochimie *78*, 364-369.

Pottier, I., Gente, S., Vernoux, J.P., and Gueguen, M. (2008). Safety assessment of dairy microorganisms: *Geotrichum candidum*. International journal of food microbiology *126*, 327-332.

Redhead, S.A., and Malloch, D.W. (1977). The Endomycetaceae: new concepts, new taxa. Canadian Journal of Botany *55*, 1701-1711.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet *16*, 276-277**.**

Smith, M.T., de Cock, A.W., Poot, G.A., and Steensma, H.Y. (1995). Genome comparisons in the yeastlike fungal genus *Galactomyces* Redhead et Malloch. International journal of systematic bacteriology *45*, 826-831.

Smith, M.T., and Poot, G.A. (2003). Genome comparisons in the genus *Dipodascus* de Lagerheim. FEMS yeast research *3*, 301-311.

Smith, M.T., Poot, G.A., and de Cock, A.W. (2000). Re-examination of some species of the genus *Geotrichum* Link: Fr. Antonie van Leeuwenhoek *77*, 71-81.

Suh, S.O., and Blackwell, M. (2006). Three new asexual arthroconidial yeasts, *Geotrichum carabidarum sp*. *nov*., *Geotrichum histeridarum sp*. *nov*., and *Geotrichum cucujoidarum sp*. *nov*., isolated from the gut of insects. Mycological research *110*, 220-228.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular biology and evolution *28*, 2731-2739.

Ueda-Nishimura, K., and Mikata, K. (2000). Two distinct 18S rRNA secondary structures in *Dipodascus* (Hemiascomycetes). Microbiology *146 (Pt 5)*, 1045-1051.

Vaughan-Martini, A., and Kurtzman, C.P. (1985). Deoxyribonucleic Acid Relatedness among Species of the Genus *Saccharomyces* Sensu Stricto. International journal of systematic bacteriology *35*, 508-511.

Wouters, J.T.M., Ayad, E.H.E., Hugenholtz, J., and Smit, G. (2002). Microbes from raw milk for fermented dairy products. International Dairy Journal *12*, 91-109.

Figure legends

Figure 1 **Molecular Phylogenetic analysis:** Phylogenetic tree reconstructed with Maximum Likelihood and neighbor-joining on the concatenation *ACT1*, *RPB1*, *RPB2* and *TEF1* coding sequences amounting to 2362 positions. Bootstrap values were from 100 replicates. Bootstraps over 60 are indicated; the first number is PhyML bootstrap, the second is neighbor joining one

Figure 1

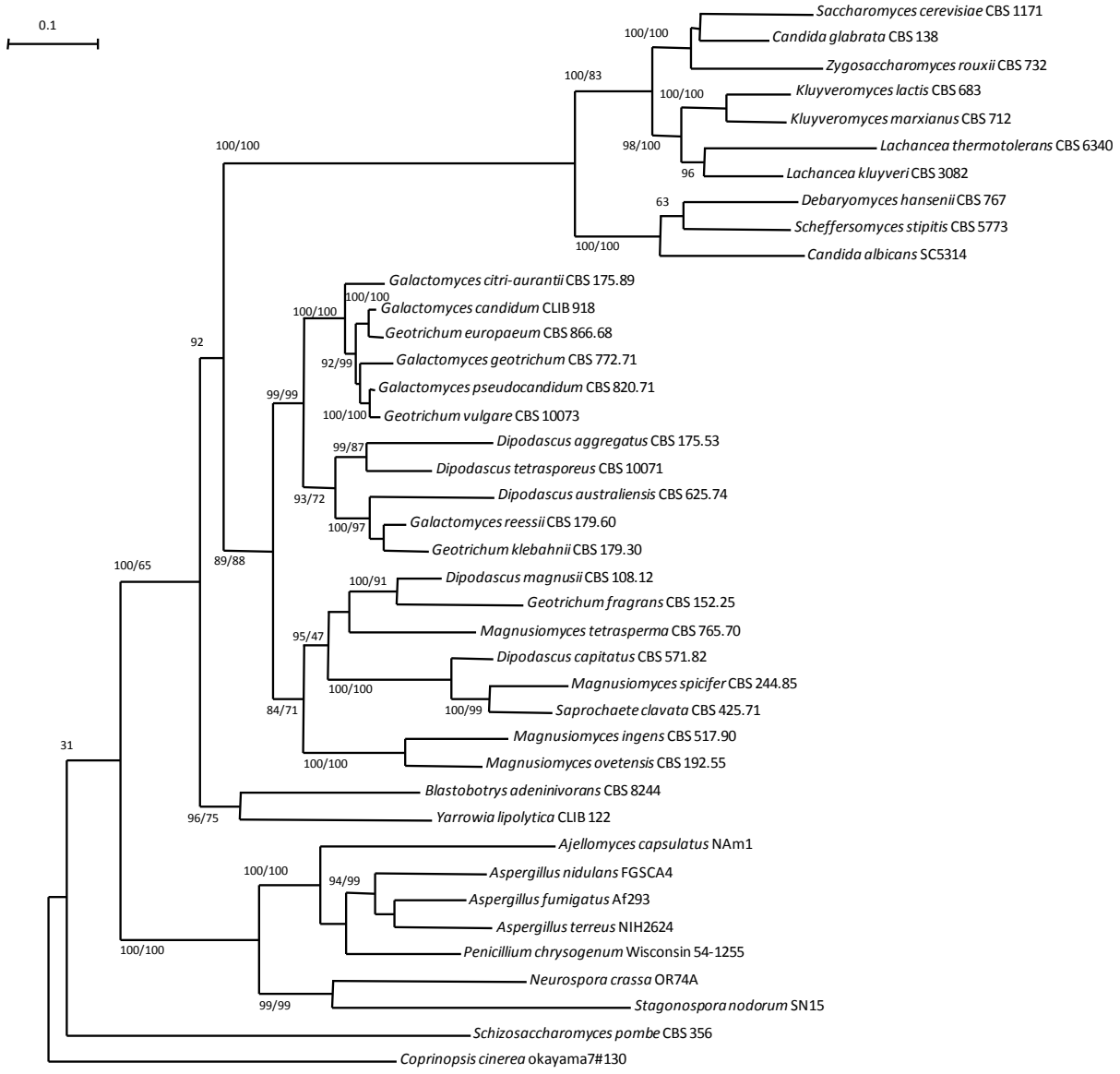


Table 1 Yeast strains compared in this study

Genus	Species	New species name proposed in this study	Strain designation[a]		GenBank accession numbers			
			CBS	CIRM	ACT1	RPB2	RPB1	TEF1α
Geotrichum								
	Galactomyces geotrichum	*Geotrichum pseudocandidum*	CBS 772.71	CLIB 1363	HE984444	HE984463	HE984486	n.a
	Galactomyces pseudocandidum	*Geotrichum pseudocandidum*	CBS 820.71	CLIB 1375	HE984445	HE984464	HE984487	n.a
	Geotrichum vulgare		CBS 10073	CLIB 1379	HE984446	HE984465	HE984488	n.a
	Galactomyces candidum	reinst. *Geotrichum candidum*	CBS 11176	CLB 1366	n.a	HE984474	HE984497	n.a
	Galactomyces candidum	reinst. *Geotrichum candidum*		CLIB 918	HE984447	HE984466	HE984489	n.a
	Galactomyces citri aurentii	*Geotrichum citri-aurentii*	CBS 175.89	CLIB 1362	n.a	HE984462	HE984485	n.a
	Galactomyces candidum	reinst. *Geotrichum candidum*	CBS 9194	CLIB 1378	n.a	HE984467	HE984490	n.a
	Geotrichum europaeum		CBS 866.68	CLIB 1371	n.a	HE984468	HE984491	n.a
Dipodascus								
	Dipodascus aggregatus		CBS 175.53	CLIB 1356	HE984440	HE984458	HE984481	n.a
	Dipodascus albidus		CBS 766.85	CLIB 1357	HE984441	HE984459	HE984482	n.a
	Dipodascus australensis		CBS 625.74	CLIB 1358	HE984442	HE984460	HE984483	n.a
	Dipodascus capitatus		CBS 571.82	CLIB 1384	HE984450	HE984471	HE984494	n.a
	Dipodascus geniculatus		CBS 184.80	CLIB 1359	HE984452	HE984475	HE984498	n.a
	Dipodascus tetrasporus		CBS 10071	CLIB 1360	n.a	HE984473	HE984496	n.a
	Geotrichum klebahnii	*Dipodascus klebahnii*	CBS 179.30	CLIB 1364	HE984438	HE984455	HE984478	n.a
	Galactomyces reessii	*Dipodascus reessii*	CBS 179.60	CLIB 1365	n.a	HE984456	HE984479	n.a
Magnusiomyces								
	Magnusiomyces tetrasperma		CBS 765.70	CLIB 1383	HE984439	HE984457	HE984480	n.a
	Magnusiomyces spicifer		CBS 244.85	CLIB 1382	HE984449	HE984470	HE984493	n.a
	Magnusiomyces ovetensis		CBS 192.55	CLIB 1381	HE984453	HE984476	HE984499	n.a
	Magnusiomyces ingens		CBS 517.90	CLIB 1386	HE984454	HE984477	HE984500	n.a
	Saprochaete clavata	*Magnusiomyces clavata*	CBS 425.71	CLIB 1385	HE984448	HE984469	HE984492	n.a
	Geotrichum fragrans	*Magnusiomyces fragrans*	CBS 152.25	CLIB 1387	HE984451	HE984472	HE984495	n.a
	Magnusiomyces magnusii		CBS 108.12	CLIB 1380	HE984443	HE984461	HE984484	n.a

[a]Source of strains: CBS, Centralbureau voor Schimmelcultures, Utrecht, the Netherlands; CLIB : CIRM-Levures, Thiverval-Grignon, France

Chapitre 2

Specialization of the cheese isolates of the species *Geotrichum candidum* revealed by MLST

G. Morel, F. Laaghouiti, S. Mallet, N. Jacques and S. Casaregola

INRA UMR1319, Micalis Institute, CIRM-Levures, 78850 F-Thiverval-Grignon, France.

AgroParisTech UMR1319, Micalis Institute, 78850 F-Thiverval-Grignon, France.

Correspondence: Serge Casaregola, serge.casaregola@grignon.inra.fr

Key words: Saccharomycotina yeast, *Galactomyces candidum*, *Geotrichum silvicola*, genetic diversity, adaptation

Running title: *Geotrichum candidum* MLST

Abbreviations: MLST, Multi Locus Sequence Typing; LTR, Long Terminal Repeat; CLP, Chromosome Length Polymorphism

Abstract

Geotrichum candidum is an ubiquitous yeast and is essential for cheese making during which it provides various biochemical activities affecting texture and aroma. Previous characterizations of this species only addressed cheese isolates. Here, we report a study of the genetic diversity of dairy and environmental strains of *G. candidum*. We developed a Multi Locus Sequence Typing (MLST) scheme, and the analysis of 19 housekeeping genes identified a set of five loci sufficiently divergent to yield a total of 30 Sequence Types (STs) among 55 *G. candidum* strains, 37 of them isolated from dairy sources or cheese, the rest from the environment. Overall sequence variability was low: 0.9%. Phylogenetic analyses clearly differentiated two main clades comprising 50 of the 55 isolates. One clade included mainly clearly distinct environmental isolates and few cheese strains. The second clade was constituted of 39 strains (14 STs), all but one isolated from cheese. This suggests a degree of adaptation to the dairy ecosystems by a group of specialized *G. candidum* strains. In addition to the MLST scheme, we developed a fast and reproducible RAPD-like method for *G. candidum*, PCR inter LTR; this method could type the dairy strains which could not be distinguished by MLST. Thus, l all of the strains, except a group of five potentially identical strains from Haute-Savoie, could be individualized. Our findings indicate that the genetic diversity of *G. candidum* is low and that the diversity observed using RAPD may be linked to genomic plasticity. Finally, by analyzing the distribution of the mating types in the strains studied, we found that genetic exchanges are frequent in *G. candidum*. This may explain the low genetic diversity observed in this species.

Introduction

Geotrichum candidum (teleomorph *Galactomyces candidum*) is commonly found in foodstuffs, either as an integral part of their normal constitution or as a contaminant. It is naturally present in raw milk (Desmasures et *al.*, 1997) and is marketed as a starter for cheese making, because of its proteolytic activities, its aromatic properties and its covering properties: it is desirable on the surface of semi-hard cheeses, and mould-ripened or smeared soft cheeses (Marcellino et *al.*, 2001).

G. candidum was first described by Link in 1809. Since then, several other names have been given to *G. candidum* (Kurtzman et *al.*, 2011). For a long time, it was considered to be a filamentous fungus. It now classified in the Saccharomicotina subphylum (de Hoog and Smith, 2004; Groenewald et *al.*, 2012; Kurtzman and Fell, 1998); Morel et *al.*, (in preparation).

G. candidum displays substantial morphological diversity and a wide ecological distribution. The genetic basis of this variability has been investigated: work based on RAM-PCR (Marcellino et *al.*, 2001) and RAPD-PCR (Gente et *al.*, 2002b) observed a high degree of genetic diversity of *G. candidum*, in some way correlated to its various ecological niches and morphological polymorphism. PFGE profiling of 13 strains revealed very diverse numbers of chromosomes, and therefore genome sizes. This led to the suggestion that genome size correlated with strain morphology (Gente et *al.*, 2002a).

In addition to various typing methods based on RAPD and chromosome length polymorphism, there are two main typing methods for yeasts. Multi-Locus Sequence Typing (MLST) is based on comparisons of single housekeeping gene sequences and it is reliable and reproducible. This method can be used to create a database, which can be continuously updated and enlarged by independent laboratories (Bougnoux et *al.*, 2004; Taylor and Fisher, 2003). It is the most popular typing method for yeast and has been applied to more than 10 species, including *Candida albicans* (Bougnoux et *al.*, 2002), *Candida glabrata* (Dodgson et *al.*, 2003), *Candida tropicalis* (Tavanti et *al.*, 2005), *Candida dubliniensis* (McManus et *al.*, 2008), *Candida krusei* (Jacobsen et *al.*, 2007), *Saccharomyces cerevisiae* (Munoz et *al.*, 2009) and *Cryptococcus gattii* (Feng et *al.*, 2008). The second method, microsatellite typing, is based on the variability of small repeats of two or three nucleotides, has proven to be valuable (Garcia-Hermoso et *al.*, 2007; Legras et *al.*, 2007; Legras et *al.*, 2005; Pan et *al.*, 2012). It is particularly suitable for isolates that diverged recently (Haasl and Payseur, 2011; Klaassen, 2009). Although inconsistent results have been obtained with MLST and microsatellite analyses (Vanhee et *al.*, 2009) studies in yeast, and in particular with *C. albicans* have shown that MLST and microsatellite typing led to similar results (Garcia-Hermoso et *al.*, 2007).

We exploited the recently completed genome sequence of the *G. candidum* strain CLIB 918, or ATCC 204307, (Morel et al, in preparation) to develop protocols for typing this species. We developed a MLST scheme based on five housekeeping genes in 55 strains. A total of 30 sequence types (STs) were differentiated and a number of them formed a distinct clade entirely constituted of isolates from cheese. Contrasting with previous suggestions, our work reveals that the genetic diversity of *G. candidum* is low. An additional method, a inter LTR PCR fingerprinting was developed to differentiate strains sharing the same MLST sequence type and to provide a fast and reliable tool for studying the diversity of industrial isolates of *G. candidum*. Finally, by analyzing the distribution of mating types we show that clonality is much reduced in this yeast species.

Materiels and methods

Yeast strains

Isolates used in this study are listed in Table 1. FM strains were provided during the "Food Microbiomes" project by producers of starter cultures and cheeses, and were given FM numbers to keep their origin confidential; UCMA strains were provided by the University of Caen (France) and CLIB strains by CIRM-Levures (Thiverval-Grignon, France; http://www.inra.fr/cirmlevures). Most of the *G. candidum* isolates were from dairy products and are involved in the ripening of various cheeses, some are from other habitats such as soil; these strains were isolated in various places in France and the rest of the world. Other *G. candidum* strains were from Centralbureau voor Schimmelcultures (Utrecht, the Netherlands), Museum National d'Histoire Naturelle, Laboratoire de Cryptogamie (Paris, France), VTT Technical Research Center (Finland), and Kasetsart University (Bangkok, Thailand)

Growth conditions

Routinely, yeast strains were grown in YPD (Yeast Peptone Dextrose: yeast extract 10 g / L, bacto peptone 10 g/L glucose 10 g/L) at 28 ° C with shaking. For solid medium, 14g / L of agar was added to YPD. Colony morphology was observed on YPD agar and PDA (Potato Dextrose Agar) medium (bioMérieux). Aliquots of 50 µL of exponentially growing cultures were spotted onto the center of Petri dishes and incubated at 28°C for 7 days.

DNA extraction

Cultures grown in 3ml YPD medium overnight at 28°C were centrifuged at 2500*g* for 3min, and the cell pellets were washed in 750µl 50 mM EDTA. Each cell pellet was resuspended in 200 µl lysis buffer (1 % SDS, 2 % triton, 100 mM NaCl, 50 mM EDTA, 50mM Tris, pH=8), 200 µl chloroform/phenol (pH=8) and 300 mg glass beads and mechanically shaken for 4 mins. Then, 200 µl TE buffer was added and the samples centrifuged for 5 mins at 12,000*g*. The aqueous phase was transferred to a new tube, and two chloroform extractions were carried out. DNA was precipitated with an equal volume of 100% ethanol and centrifuged for 4 mins at 12,000*g*. The pellet was rinsed with 400 µl of 70 % ethanol, dried at room temperature for 15 mins, resuspended in 50µl TE buffer, and incubated with 2 µl RNase (10mg/ml) for 30 mins at 37°C. The DNA concentration was quantified on agarose gel.

PCR amplification

The oligonucleotide primers used in this study are listed in Supplementary Table S1. They were designed with Primer3 (http://fokker.wi.mit.edu/primer3) from the complete genome sequence of *G. candidum* CLIB 918 (Morel et al, in preparation). For inter-LTR PCR fingerprinting, LTR sequences were aligned using the ClustalX2 program (17846036) and primers were designed to correspond to parts of the conserved regions, i.e. oligonucleotides GC1_LTR_for and GC2_LTR_rev from the *G. candidum* LTR retrotransposon Tgc5. For MLST, 2 µl of DNA (containing between 25 and 50 ng) was added to 48 µl PCR reaction mix containing 0.8 mM dNTP mixture, 0.1 µM forward and reverse primers in the manufacturer's recommended buffer and 1 U TaKaRa Ex *Taq*. Reactions were run on a 2720 thermal cycler (Applied Biosystems) as follows: 5 min at 94°C followed by 30 cycles of 30 s at 94 °C, 40 s at temperatures between 35 and 50 °C and between 30 and 60 s at 72 °C, with a final extension step of 7 min at 72 °C. For sexual sign determinations, PCR were run with the primer pairs GECA_MATAfw/GECA_MATArv and GECA_MATB fw and rv on a 2720 thermal cycler (Applied Biosystems) as follows: 5 min at 94°C followed by 30 cycles of 30 s at 94 °C, 40 s at 50 °C and 60 s at 72 °C, with a final extension step of 7 min at 72 °C.

For inter-LTR PCR fingerprinting, PCR conditions with the primer pair GC1_LTR_for /GC2_LTR_rev were as follows: 94 °C for 4 min, four cycles of 94 °C for 30 s, 37 °C for 30 s and 72 °C for 2 min, followed by 30 cycles of 30 s at 94 °C, 40 s at temperature varying between 50 °C and 2 min at 72°Cwith a final extension at 72 °C for 4 min.

PCR products were visualized on a 2% agarose gel (wt/vol) (Q-Biogen, France) with 1 X TBE electrophoresis buffer (Q-Biogen, Illkirch, France) containing 0.2 mg/mL ethidium bromide, run at 120 V in a SUB-CELL GT electrophoresis system (BIORAD, Les Ulis, France) for 1 hr.

DNA sequence determination and phylogenetic analysis

PCR fragments were sequenced on both strands by Eurofins MWG Operon (Ebersberg, Germany). Sequences were assembled with the phred/phrap/consed package (Ewing and Green, 1998; Ewing et *al.*, 1998; Gordon et *al.*, 1998) and analyzed with various programs including BLAST implemented at NCBI (http://www.ncbi.nlm.nih.gov).Sequence alignments were generated by using ClustalX2 (Larkin et *al.*, 2007) and MUSCLE (Edgar, 2004) and manually adjusted with Genedoc (http://www.nrbsc.org/gfx/genedoc/). Phylogenetic trees were constructed with the Neighbor-Joining program implemented in ClustalX2 or with Phyml (Guindon and Gascuel, 2003). Phylogenetic trees were visualized with Seaview or NJplot (Gouy et *al.*, 2010; Perriere and Gouy, 1996). Bionumerics software 5.1 (Applied Maths, Belgium) was used for the acquisition of inter-LTR profiles and the integration MLST, inter-LTR and phenotypic data.

Results

Design of a MLST scheme for *G. candidum*

The literature on MLST was surveyed to identify genes which have been used to design MLST schemes in previous studies on yeasts (Bougnoux et *al.*, 2002; Dodgson et *al.*, 2003; Jacobsen et *al.*, 2007; McManus et *al.*, 2008; Munoz et *al.*, 2009; Tavanti et *al.*, 2005). A total of 19 sequences corresponding to the *G. candidum* orthologs of such genes were extracted from the genome of CLIB 918. Primers were designed and were tested in PCR amplification experiments on 55 strains assumed to be haploid on the basis of their mating type characteristics (see below). The study population included 35 *G. candidum* strains isolated from cheese, the *G. bryndzae* type strain CBS 11176^T, and various *G. sivicola* strains including the type strain CBS 9194^T. From the 19 genes tested, five markers were selected as showing the greatest sequence divergence between strains: *URA1*, *URA3*, *SAPT2*, *SAPT4 and NUP116.* Data for the five MLST markers are presented in Table 2. The PCR fragments analyzed were between 423 bp for *NUP116* and 514 bp for *URA1* yielding a concatenated sequence of 2373 bp containing 72 polymorphic sites.

Table 2 summarizes relevant information about the diversity of each locus. The number of polymorphic sites varied between seven for *URA1* and 22 for *SAPT2*. The number of allelic profiles between loci was five for *URA1*, five for *SAPT4*, five for *URA* 3, six for *NUP116* and 11 for *SAPT2*. The average percentage of divergence for all strains ranges from 0.77 % for *URA1* to 4.28% for *SAPT2*. The most variable marker was *SAPT2*, and *URA1* was the least variable. *URA1* allowed differentiation of four groups of strains, whereas *SAPT2* allowed differentiation of 11 groups (Table S2), where a group is defined as a single sequence type for the marker.

Geographical and industrial influence on strain grouping

Strains with the same concatenated sequence were classified as belonging to the same sequence types (STs). Thus, the combination of the alleles at the five loci led to the identification of 30 STs among the 55 strains. Strains from cheese were clearly separated from other strains. Overall, 23 strains were the sole member of their ST, and the other 32 strains were clustered in six STs. ST 18 contained two strains; ST 22, three strains; ST 21, four strains; ST 26 , ST 28 and ST 29, five strains each. ST 24 included eight strains (Table 3). The number of polymorphic sites between STs was between 0.04 % and 2.06 % (Table 4).

Trees were constructed by neighbor joining (Figure 1 and Figure 2). Twenty-one of the 35 nodes were supported by bootstrap values of over 60 %. Two main clades, called I and II, could be distinguished, with a small number of intermediate strains (from five STs) which did not fit within either of these two clades. Clade I contained 11 STs (11 strains) and clade II contained 14 STs (38 strains). Except for a strain isolated from Thailand, which showed substantial sequence divergence (0.97%), there was little divergence within either of these clades (0.5% for clade I and 0.45% for clade II). All STs of clade I contained only a single strain, whereas the STs containing numerous strains were in clade II: seven STs in clade II contained at least two strains and and six STs (ST19, ST 20, ST 23, ST 25, ST 27 and ST 30) contained a single strain. Interestingly, clade II contained only strains involved in cheese making (either from cheese made with raw milk or from starters used in the cheese industry), except for one isolate, FM 268 in ST 26, which was isolated from stools. Only five strains isolated from cheese were in clade I (Figure 2). Thus strains involved in cheese making preferentially cluster in clade II. Clade I was much more heterogeneous than clade II, and most strains of clade I were isolated either from the environment (NT 12, LCP 51.590, FM 270) or industrial plant (VTTC 4559 and FM 212). Other strains in clade I that were isolated from dairy sources are contaminants, for example FM 03, initially described as *G. silvicola* by our industrial supplier. However, four strains in clade I are used in cheese production (CLIB 1237, CLIB1267, CLIB 1283 and CLIB 1258).

In addition to most “cheese” strains clustering in one of the two clades, they also appeared to group according to their geographical origins. For example, strains originating from the Auvergne constituted ST 22 and strains from Haute-Savoie were in ST 21 and ST 28. The origin of the Slovakian strain CBS11176 is puzzling, since it was classified among French cheese isolates. This strain was first described as *Geotrichum bryndzea* (Sulo et *al.*, 2009), but it was recently shown to belong to *Gal. candidus*, the teleomorph of *G. candidum,* by (Groenewald et *al.*, 2012); the position of this strain in the phylogenetic tree confirmed the recent taxonomic change.

Although MLST was very informative about the divergence of environmental strains, it was not clear whether the large ST groups were due to clonality or lack of discrimination of the MLST markers. We therefore developed another typing method, and analyzed the distribution of the mating type among the strains.

Inter-LTR PCR distinguishes most of the *G. candidum* cheese strains

Sequence analysis of the complete genome of *G. candidum* revealed the presence of a Long Terminal Repeat (LTR)-retrotransposon related to the yeast Ty5 family (Morel et al, in preparation). LTRs are short sequences of around 300 bp widely repeated in some genomes, like those of the *Saccharomyces* and other Saccharomycotina yeasts (Bleykasten-Grosshans et *al.*, 2011; Dujon, 2010; Neuveglise et *al.*, 2002) either as flanking parts of the cognate retrotransposon or as solo elements resulting from removal of the retrotransposon through recombination between the two flanking LTRs. These repeated solo elements are numerous in some genomes, and this feature has been used to differentiate isolates within a single species by amplifying regions which separate LTRs, using primers in the conserved regions of LTRs; this approach has been applied to various species including *S. cerevisiae* (Legras and Karst, 2003; Ness et *al.*, 1993), *Debaryomyces hansenii* and *Kluyveromyces marxianus* (Sohier et *al.*, 2009).

The LTRs of *G. candidum* CLIB 918 were aligned with ClustalX and conserved regions used to design oligonucleotide primers (data not shown); these were used to amplify genomic DNA from the 32 clade II strains which could not be distinguished by MLST. Various primer pairs were tested; those leading to the most discriminating results, GECA_GC1for and GECA_GC2rev, were selected and used for subsequent analyses (Table S1). To ensure repeatability of the PCR inter-LTR fingerprinting method, various PCR conditions and genomic DNA extraction techniques were tested with three strains in independent experiments as described by (Sohier et *al.*, 2009) (data not shown).

Fingerprinting profiles were then generated for all strains in non-singleton STs (listed in Table 3). The amplified bands ranged from 400 to 1300 bp and the patterns of the various strains differed in

fragment number, size and intensity. To assess the diversity of strains which shared the same ST, we associated the inter-LTR profiles of these strains to the MLST tree containing only STs with at least two strains (Figure 3).

Most of the strains gave different inter LTR profiles; there were 28 different profiles among the 36 clade II isolates tested (Figure 3). Thus, strains in ST18, ST 22, ST24, ST26 and ST29 have different inter LTR profiles. However, CLIB 1239 and CLIB 1241 in ST 29 shared the same inter LTR profiles as did FM 76 and FM 77 in ST 26. A number of strains gave very similar profiles, for example those in ST 28 (CLIB 1244, CLIB 1245, CLIB 1246, CLIB1247 and CLIB 1253); these strains were all isolated from the Haute-Savoie and also shared the same MLST ST; their classification into the same cluster is therefore coherent. By contrast, FM 29, FM 30 and FM 31 showed different inter-LTR patterns, although they belonged to the same ST22. The combination of MLST and inter-LTR fingerprinting allowed powerful discrimination between cheese strains.

Some isolates which shared the same MLST profile, also shared the same carbon source assimilation characteristics. However, the phenotypic profiles concerning galactose, lactate and ribose assimilation did not differ substantially between the strains studied (see also Marcellino etal, 2001). All the isolates in ST 24 and ST 26 could assimilate galactose, whereas isolates in ST 28 and in ST 21 could not. In ST 29, CLIB 1256, unlike the other isolates of the ST, was not able to assimilate galactose. The findings were similar for the assimilation of lactate and ribose. Overall, the limited phenotypic diversity was not consistent with the observed variability of inter-LTR profiles.

Analysis of the distribution of mating types

G. candidum has a sexual state (Butler and Petersen, 1970), mating between compatible strains produces spores (de Hoog et al, 1986) and some strains have been reported to be homothallic (de Hoog and Smith, 2011). Nevertheless, very little is known about the life cycle of this species. Analysis of the complete genome sequence of *G. candidum* CLIB 918 led to the identification of the mating type locus: a single coding gene, named *MATA* on the basis of its sequence similarity and conserved organization with known *MATA* genes in other yeasts and fungal species (Morel et al, in preparation). We used the MLST typing results to learn more about the sexuality of *G. candidum*. Incidentally, mating types could also be used as an additional marker for strain differentiation.

To test for the allele in our isolates, primers were chosen to amplify a region of 501 bp within the *MATA* gene from the genome of CLIB 918. These primers were used for PCR with strains representative of all STs (Table 1). PCR products of approximately 280 bp were obtained for 31 strains. No PCR product was observed for 26 strains, suggesting that these strains do not carry the

MATA gene, and therefore that they may carry the opposite mating type. The genome of *G. candidum* CLIB 918 does not contain silent cassettes or genes like the one encoding the HO-like endonuclease indicating that *G. candidum* is heterothallic (Morel et al, in preparation). To confirm these possibilities, we attempted to identify the other sexual allele. Thus, primers located in the regions flanking the *MATA* gene were designed and used to amplify the corresponding region in three strains which did not give a positive signal with the *MATA*-specific primers. A region of 1.5 kb was successfully amplified from the three strains and was entirely sequenced: it contained an ORF, different from the *MATA* gene, in an otherwise identical environment. This gene was called *MATB*. To identify the two genes in our strains, we designed primers specific for sequences within the *MATB* gene to amplify a fragment of 236 bp (Table S1). PCR amplification with the *MATB*-specific primers yielded a DNA fragment of approximately the expected size in 26 isolates, and, as expected, failed to amplify any fragment from the strains carrying *MATA*.

Seven strains carried both *MATA* and *MATB* alleles (Table 1). These strains also carried two divergent alleles of some the genes used in the MLST scheme. As stated above, these strains were excluded from our MLST study.

Thus, in the population studied, there are 30 MATA strains and 25 MATB strains; the two signs are distributed evenly suggesting that sexuality is widespread in *G. candidum* (Table 1). Three STs (ST 21, 24 and 26) contained both strains with *MATA* and strains with *MATB* alleles (Figure 2 and 3). This presumably represents the smallest possible genetic scale of sex, and is clear evidence for recombination. The long branch which leads to clade II isolates indicates that clade II isolates are all derived from a common ancestor. The presence of strains with different *MAT* alleles in clade II suggests that genetic exchanges via recombination may have occurred within this clade, and that clade II is not clonal. The existence of the five intermediate strains corresponding to the five STs (ST 12, 13, 14, 15, 16) may be the consequence crossing between clades I and II.

Discussion

We report the development of tools for investigating the evolution of *G. candidum* and for studying industrial strains used in cheese production. Previous typing studies based on RAM-PCR, RAPD and CLP techniques suggest that there is substantial diversity within the *G. candidum* species (Gente et *al.*, 2002a; Gente et *al.*, 2002b; Marcellino et *al.*, 2001). This variability was correlated to observed phenotypic analysis and geographic origin.

A preliminary analysis in our laboratory (Morel and Casaregola, our unpublished data) showed that the actin coding gene in a large number of strains, including 28 strains described in (Marcelino et al, 2001), was almost identical. This gene is highly conserved and is not considered to be a good marker to study intraspecific diversity. However, our observations indicated that there was little sequence diversity in this species. The MLST analysis we report here involving 19 single copy genes confirms that genetic diversity in *G. candidum* is low, with the most polymorphic genes showing only 4.28 % divergence. Despite this low diversity, our MLST analysis provided two major results. The first is that *G. candidum* strains can be straightforwardly differentiated on the basis of the partial sequences of five genes: 30 different STs were obtained for 55 isolates. This classification grouped a number of strains according to their geographical origin: strains isolated from Haute-Savoie cheese were never associated with strains isolated in the Auvergne or Normandy. However, these Haute-Savoie strains belonged to different groups and are therefore different. Nevertheless, there were 32 isolates in clade II that could not be typed by this approach.

We were able to improve strain differentiation by developing inter-LTR PCR for *G. candidum*. This method is based on variability of insertions of transposons and their cognate LTRs. Our method successfully differentiated some of the strains that were not differentiated by MLST. However, it is not possible to assert that this variability is associated to functional changes because, in some cases, as with CLP, no correlation could be established between LTR location in the genome and morphology. However, this method is simple, reliable and fast, and may therefore be useful to differentiate between strains, for example, to follow yeast population dynamics in cheese during ripening.

Twenty-eight of the strains studied here have previously been studied by RAPD (Marcellino et *al.*, 2001). The previous RAPD results and our findings are generally coherent, although there are some discrepancies. For instance, the three strains, CLIB 1267 (=GC129), CLIB 1258 (=GC101= and CLIB 1237 (=GC37) isolated from milk used for manufacturing Camembert cheese (Normandy), ripened St Nectaire cheese (Auvergne) and Chaource cheese (Champagne-Ardenne), respectively, were

previously found to be closely related to other cheese strains (Marcellino et *al.*, 2001). However, we found that they are well separated from other cheese strains because they clearly belong to clade I, consistent with their geographical origin and the type of cheese from which they were isolated.

It has also been suggested that CLIB 1253 (=GC90), a member of ST 28 composed of isolates from ripening cheese from Annecy, was different from the other strains that we classified as also being members of the same ST. We clearly showed here that these five strains are indistinguishable on the basis of MLST, inter-LTR profile, phenotypic characteristics and mating type. These strains were isolated from two different cheeses, reblochon and tomme de Savoie, made in different dairies. Our results therefore indicate that these are isolates of the same strain present at more than one particular site and in more than one particular cheese.

Finally, two strains CLIB 1257 (=GC100) and CLIB1260 (=GC105) were reported to be almost identical in the RAPD study cited above. In our analysis, they belong to two different STs, and unlike CLIB 1257, CLIB 1260 cannot assimilate lactate. MLST was therefore able to differentiate between strains in the same RAPD group.

The 55 isolates studied here were classified in 30 STs. A total of 22 isolates were typed as they all display a unique ST. These 22 typed isolates include 11 isolates of clade I, five of intermediate clades and six of clade II. Nevertheless in the MLST scheme was not able to type 32 isolates in clade II. The inter LTR profiles allowed us to type all but five isolates of this population. Our MLST analysis does not appear to be completely coherent with RAPD typing results, both those we obtained using PCR inter-LTR and those using RAPD by Marcellino et al (2001). It has also been reported that *G. candidum* is highly variable as assessed by chromosome separation using PFGE (Gente et al, 2002). However, CLP very often results from ectopic recombination between similar sequences, like transposons (Casaregola et *al.*, 1998; Rachidi et *al.*, 1999; Zolan, 1995)). Accordingly, we observed a very large numbers of various kinds of transposons in the genome of *G. candidum* (Morel et *al.*, unpublished).

This discrepancy between MLST and RAPD may be inherent to the techniques used because methods based on the amplification of repeats do not measure genetic diversity directly, but rather, like PFGE, indicate chromosomal differences. Clade II strains could not be separated on the basis of sequence divergence, and displayed little phenotypic diversity; they are therefore presumably closely related. Similarly, the species *Schizosaccharomyces pombe* displays substantial CLP which is not associated with either sequence or phenotypic variability (Brown et *al.*, 2011).

The second major result of the MLST analysis is that almost all the cheese strains were classified in one clade and indeed in a limited number of STs, whereas the other clade contained a both

environmental and cheese strains, all distinct from each other. The remaining strains are on long branches between clade I and clade II. The intermediate position of strains in ST14, ST15 and ST 16 may be the result of mating between clade I and clade II or even clade II and ST 12 (previously *G. silvicola*). Such intermediate positions of strains between major industrial/specialized clades is reminiscent of population genomics findings for some *S. cerevisiae* strains, like laboratory strains W303 and S288c, clinical strains including YJM 789 and fermentation strains including DBVPG 1853 and NCYC 361 (Liti et *al.*, 2009). Likewise, clade II strains are reminiscent of the clades composed of the European wine yeasts or the Sake yeasts in the same study. Interestingly, clade I with its higher number of STs per strain resembles results obtained by Bai with environmental isolates of *S. cerevisiae* from primeval forest in China (Wang et al, 2012), and the various branches of the geographically distinct *S. cerevisiae* described by Liti et *al.* (2009).

There are at least two possible explanations of the origin of clade II strains: one is extensive adaptation to the cheese environment and human activities of cheese processing ; the other is natural selection of strains with the abilities to cope with these environments. We cannot currently determine which of these two possibilities is more likely.

The grouping of the *G. candidum* strains into two populations, dairy and environment isolates is reminiscent of the way *S. cerevisiae* strains group according to their origin or their involvement in particular processes (Legras and Karst, 2003; Liti et *al.*, 2009; Schacherer et *al.*, 2009). Studies on the population structure of both *S. cerevisiae* and *S. paradoxus* have shown that the divergence of the sequences within *S. paradoxus* is greater than that within *S. cerevisiae*. Strains of *S. cerevisiae* domesticated by man, and especially wine yeasts, showed little genetic diversity in contrast to wild strains of *S. paradoxus* (Liti et *al.*, 2006; Liti et *al.*, 2009). The overall sequence divergence in *G. candidum* clade II (0.30%) (Table 4) is even lower than that of *S. cerevisiae* wine yeasts (1.2%), consistent with a similar mode of evolution for these two industrial species. It is likely that *S. cerevisiae* wine yeasts have a common origin (Legras et *al.*, 2007), and frequent genetic exchanges have limited genetic diversity. We show here that clade II strains also have a common origin. *G. candidum* may have evolved by similar mechanisms, some of which are associated with human activities.

Indeed, unlike a number of fungi, *G. candidum* has been shown to have sexuality (Butler and Petersen, 1970), although only mating compatibilities and spore production have been assessed (de Hoog et *al.*, 1986; Groenewald et *al.*, 2012). Using molecular analysis to test for the presence of both mating types in *G. candidum* haploid strains, we report the first analysis of the extent of genetic exchange in *G.candidum*. The even distribution of the mating types in the study population is consistent with the absence of clonality that we observed, even in the sub-population forming the

MLST “cheese” clade. This implies that mating occurs in cheese. In agreement with this, we have previously observed that strains of the *Debaryomyces hansenii* complex isolated from cheese were mainly diploids or hybrids (Jacques et *al.*, 2009; Jacques et *al.*, 2010), and therefore suggested that the environment may influence the formation of diploids and hybrids. Few *G. candidum* diploids were observed: only seven of 62 strains tested carried both *MAT* alleles. The lack of clonality that we observed in *G. candidum* indicates that mating events are frequent and that meiosis in *G. candidum* is efficient.

All the members of clade II appear to have originated from a single ancestor (Figure 2). Our demonstration of the presence of both mating types in the strains of clade II indicates that there have been secondary contacts between strains from clade II and strains not belonging to this clade. The close relatedness of clade II strains suggests that (1) the cheese environment may select specific characteristics and (2) strains currently composing clade II may be the result of a purifying selection for “cheese alleles”.

The study of genetic diversity of *G. candidum* strains is important because of the trend in the cheese industry to use starters that can lead to a loss of biodiversity. We have also studied 30 other isolates used as starters for cheese-making provided by French industrial dairies. All strains grouped in clade II; they are scattered across the STs we report here, and none correspond to additional STs (unpublished observations). Our work provides information and methods to facilitate genetic improvement and genetic diversity within this species. It will make it easier to correlate genetic divergence with industrially important characteristics of *G. candidum*, including aroma production and morphology.

Acknowledgements

We thank Christelle Louis-Mondésir for expert technical assistance. We are grateful to Prof. Savitree Limtong for providing us with the isolate NT12. We also thank the *Centre National Interprofessionnel de l'Economie Laitière* (CNIEL) and the *Syndicat Professionnel des Producteurs d'Auxiliaires pour l'Industrie Laitière* (SPPAIL) for providing us with strains isolated from the cheese environment. This work received funding from the *Agence National pour la Recherche* grant “Food Microbiomes” (ANR-08-ALIA-007-02). GM was supported by a CIFRE fellowship with CNIEL.

References

Bleykasten-Grosshans, C., Jung, P.P., Fritsch, E.S., Potier, S., de Montigny, J., and Souciet, J.L. (2011). The Ty1 LTR-retrotransposon population in *Saccharomyces cerevisiae* genome: dynamics and sequence variations during mobility. FEMS yeast research *11*, 334-344.

Bougnoux, M.E., Aanensen, D.M., Morand, S., Theraud, M., Spratt, B.G., and d'Enfert, C. (2004). Multilocus sequence typing of *Candida albicans*: strategies, data exchange and applications. Infect Genet Evol *4*, 243-252.

Bougnoux, M.E., Morand, S., and d'Enfert, C. (2002). Usefulness of multilocus sequence typing for characterization of clinical isolates of *Candida albicans*. Journal of clinical microbiology *40*, 1290-1297.

Brown, W.R., Liti, G., Rosa, C., James, S., Roberts, I., Robert, V., Jolly, N., Tang, W., Baumann, P., Green, C.*, et al.* (2011). A Geographically Diverse Collection of *Schizosaccharomyces pombe* Isolates Shows Limited Phenotypic Variation but Extensive Karyotypic Diversity. G3 (Bethesda) *1*, 615-626.

Butler, E.E., and Petersen, L.J. (1970). Sexual reproduction on *Geotrichum candidum*. Science *169*, 481-482.

Casaregola, S., Nguyen, H.V., Lepingle, A., Brignon, P., Gendre, F., and Gaillardin, C. (1998). A family of laboratory strains of *Saccharomyces cerevisiae* carry rearrangements involving chromosomes I and III. Yeast *14*, 551-564.

de Hoog, G.S., and Smith, M.T. (2004). Ribosomal gene phylogeny and species delimitation in *Geotrichum* and its teleomorphs. Studies in Mycology *50*, 489–516.

de Hoog, G.S., Smith, T., and Guého, E. (1986). A Revision of the Genus *Geotrichum* and Its Telemorphs (Centraalbureau voar Schimmelcultures).

Desmasures, N., Bazin, F., and Guéguen, M. (1997). Microbiological composition of raw milk from selected farms in the Camembert region of Normandy. Journal of applied microbiology *83*, 53-58.

Dodgson, A.R., Pujol, C., Denning, D.W., Soll, D.R., and Fox, A.J. (2003). Multilocus sequence typing of *Candida glabrata* reveals geographically enriched clades. Journal of clinical microbiology *41*, 5709-5717.

Dujon, B. (2010). Yeast evolutionary genomics. Nature reviews Genetics *11*, 512-524.

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

Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome research *8*, 186-194.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome research *8*, 175-185.

Feng, X., Yao, Z., Ren, D., Liao, W., and Wu, J. (2008). Genotype and mating type analysis of *Cryptococcus neoformans* and *Cryptococcus gattii* isolates from China that mainly originated from non-HIV-infected patients. FEMS yeast research *8*, 930-938.

Garcia-Hermoso, D., Cabaret, O., Lecellier, G., Desnos-Ollivier, M., Hoinard, D., Raoux, D., Costa, J.M., Dromer, F., and Bretagne, S. (2007). Comparison of microsatellite length polymorphism

and multilocus sequence typing for DNA-Based typing of *Candida albicans*. Journal of clinical microbiology *45*, 3958-3963.

Gente, S., Desmasures, N., Jacopin, C., Plessis, G., Beliard, M., Panoff, J.M., and Gueguen, M. (2002a). Intra-species chromosome-length polymorphism in *Geotrichum candidum* revealed by pulsed field gel electrophoresis. International journal of food microbiology *76*, 127-134.

Gente, S., Desmasures, N., Panoff, J.M., and Gueguen, M. (2002b). Genetic diversity among *Geotrichum candidum* strains from various substrates studied using RAM and RAPD-PCR. Journal of applied microbiology *92*, 491-501.

Gordon, D., Abajian, C., and Green, P. (1998). Consed: a graphical tool for sequence finishing. Genome research *8*, 195-202.

Gouy, M., Guindon, S., and Gascuel, O. (2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Molecular biology and evolution *27*, 221-224.

Groenewald, M., Coutinho, T., Smith, M.T., and van der Walt, J.P. (2012). Species reassignment of *Geotrichum bryndzae*, *Geotrichum phurueaensis*, *Geotrichum silvicola* and *Geotrichum vulgare* based on phylogenetic analyses and mating compatibility. International journal of systematic and evolutionary microbiology.

Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic biology *52*, 696-704.

Haasl, R.J., and Payseur, B.A. (2011). Multi-locus inference of population structure: a comparison between single nucleotide polymorphisms and microsatellites. Heredity *106*, 158-171.

Jacobsen, M.D., Gow, N.A., Maiden, M.C., Shaw, D.J., and Odds, F.C. (2007). Strain typing and determination of population structure of *Candida krusei* by multilocus sequence typing. Journal of clinical microbiology *45*, 317-323.

Jacques, N., Mallet, S., and Casaregola, S. (2009). Delimitation of the species of the *Debaryomyces hansenii* complex by intron sequence analysis. International journal of systematic and evolutionary microbiology *59*, 1242-1251.

Jacques, N., Sacerdot, C., Derkaoui, M., Dujon, B., Ozier-Kalogeropoulos, O., and Casaregola, S. (2010). Population polymorphism of nuclear mitochondrial DNA insertions reveals widespread diploidy associated with loss of heterozygosity in Debaryomyces hansenii. Eukaryotic cell *9*, 449-459.

Klaassen, C.H. (2009). MLST versus microsatellites for typing *Aspergillus fumigatus* isolates. Medical mycology : official publication of the International Society for Human and Animal Mycology *47 Suppl 1*, S27-33.

Kurtzman, C.P., and Fell, J.W. (1998). The Yeasts - A Taxonomic Study (Elsevier Science).

Kurtzman, C.P., Fell, J.W., and Boekhout, T. (2011). The Yeasts: A Taxonomic Study (Elsevier).

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., *et al.* (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947-2948.

Legras, J.L., and Karst, F. (2003). Optimisation of interdelta analysis for *Saccharomyces cerevisiae* strain characterisation. FEMS microbiology letters *221*, 249-255.

Legras, J.L., Merdinoglu, D., Cornuet, J.M., and Karst, F. (2007). Bread, beer and wine: *Saccharomyces cerevisiae* diversity reflects human history. Molecular ecology *16*, 2091-2102.

Legras, J.L., Ruh, O., Merdinoglu, D., and Karst, F. (2005). Selection of hypervariable microsatellite loci for the characterization of *Saccharomyces cerevisiae* strains. International journal of food microbiology *102*, 73-83.

Liti, G., Barton, D.B., and Louis, E.J. (2006). Sequence diversity, reproductive isolation and species concepts in *Saccharomyces*. Genetics *174*, 839-850.

Liti, G., Carter, D.M., Moses, A.M., Warringer, J., Parts, L., James, S.A., Davey, R.P., Roberts, I.N., Burt, A., Koufopanou, V.*, et al.* (2009). Population genomics of domestic and wild yeasts. Nature *458*, 337-341.

Marcellino, N., Beuvier, E., Grappin, R., Gueguen, M., and Benson, D.R. (2001). Diversity of *Geotrichum candidum* strains isolated from traditional cheesemaking fabrications in France. Applied and environmental microbiology *67*, 4752-4759.

McManus, B.A., Coleman, D.C., Moran, G., Pinjon, E., Diogo, D., Bougnoux, M.E., Borecka-Melkusova, S., Bujdakova, H., Murphy, P., d'Enfert, C.*, et al.* (2008). Multilocus sequence typing reveals that the population structure of *Candida dubliniensis* is significantly less divergent than that of *Candida albicans*. Journal of clinical microbiology *46*, 652-664.

Munoz, R., Gomez, A., Robles, V., Rodriguez, P., Cebollero, E., Tabera, L., Carrascosa, A.V., and Gonzalez, R. (2009). Multilocus sequence typing of oenological *Saccharomyces cerevisiae* strains. Food microbiology *26*, 841-846.

Ness, F., Lavallée, F., Dubourdieu, D., Aigle, M., and Dulau, L. (1993). Identification of yeast strains using the polymerase chain reaction. Journal of the science of food and agriculture *62*, 89-94.

Neuveglise, C., Feldmann, H., Bon, E., Gaillardin, C., and Casaregola, S. (2002). Genomic evolution of the long terminal repeat retrotransposons in hemiascomycetous yeasts. Genome research *12*, 930-943.

Pan, W., Khayhan, K., Hagen, F., Wahyuningsih, R., Chakrabarti, A., Chowdhary, A., Ikeda, R., Taj-Aldeen, S.J., Khan, Z., Imran, D.*, et al.* (2012). Resistance of Asian *Cryptococcus neoformans* serotype A is confined to few microsatellite genotypes. PloS one *7*, e32868.

Perriere, G., and Gouy, M. (1996). WWW-query: an on-line retrieval system for biological sequence banks. Biochimie *78*, 364-369.

Rachidi, N., Barre, P., and Blondin, B. (1999). Multiple Ty-mediated chromosomal translocations lead to karyotype changes in a wine strain of *Saccharomyces cerevisiae*. Molecular & general genetics : MGG *261*, 841-850.

Schacherer, J., Shapiro, J.A., Ruderfer, D.M., and Kruglyak, L. (2009). Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. Nature *458*, 342-345.

Sohier, D., Dizes, A.S., Thuault, D., Neuveglise, C., Coton, E., and Casaregola, S. (2009). Important genetic diversity revealed by inter-LTR PCR fingerprinting of *Kluyveromyces marxianus* and *Debaryomyces hansenii* strains from French traditional cheeses. Dairy Sci Technol *89*, 569-581.

Sulo, P., Laurencik, M., Polakova, S., Minarik, G., and Slavikova, E. (2009). *Geotrichum bryndzae sp. nov.*, a novel asexual arthroconidial yeast species related to the genus *Galactomyces*. International journal of systematic and evolutionary microbiology *59*, 2370-2374.

Tavanti, A., Davidson, A.D., Johnson, E.M., Maiden, M.C., Shaw, D.J., Gow, N.A., and Odds, F.C. (2005). Multilocus sequence typing for differentiation of strains of *Candida tropicalis*. Journal of clinical microbiology *43*, 5593-5600.

Taylor, J.W., and Fisher, M.C. (2003). Fungal multilocus sequence typing--it's not just for bacteria. Current opinion in microbiology *6*, 351-356.

Vanhee, L.M., Symoens, F., Jacobsen, M.D., Nelis, H.J., and Coenye, T. (2009). Comparison of multiple typing methods for *Aspergillus fumigatus*. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases *15*, 643-650.

Zolan, M.E. (1995). Chromosome-length polymorphism in fungi. Microbiological reviews *59*, 686-698.

Figure legends

Figure 1: Unrooted Neighbor-Joining tree of the unique STs generated from the analysis of the five concatenated sequences in 55 G*. candidum* isolates. Clade I is indicated in red and Clade II in blue. Bar, 0.002 substitutions per site.

Figure 2: Mid-point rooting Neighbor-Joining tree of 55 *G. candidum* based on five concatenated sequences. Bootstrap values >60% (1000 repetitions) are indicated at the nodes. MATA and MATB isolates are indicated in blue and in red, respectively. Colored squares and colored circles indicate the technological use and the geographical origin of the isolates, respectively. Bar, 0.001 substitutions per site.

Figure 3: Multi-analysis of the 32 clade II isolates, which could not be typed by MLST. The MLST dendrogram was generated as described in the Materials and Methods. For each strain, the corresponding inter-LTR profile is shown. The ability to assimilate galactose (GAL), lactose (LAT) and ribose (RIB) is shown with a black box. *MATA* and *MATB* isolates are indicated with green and red squares, respectively. The geographical origin of the isolates and the ST to which they belong are shown.

Figure 1

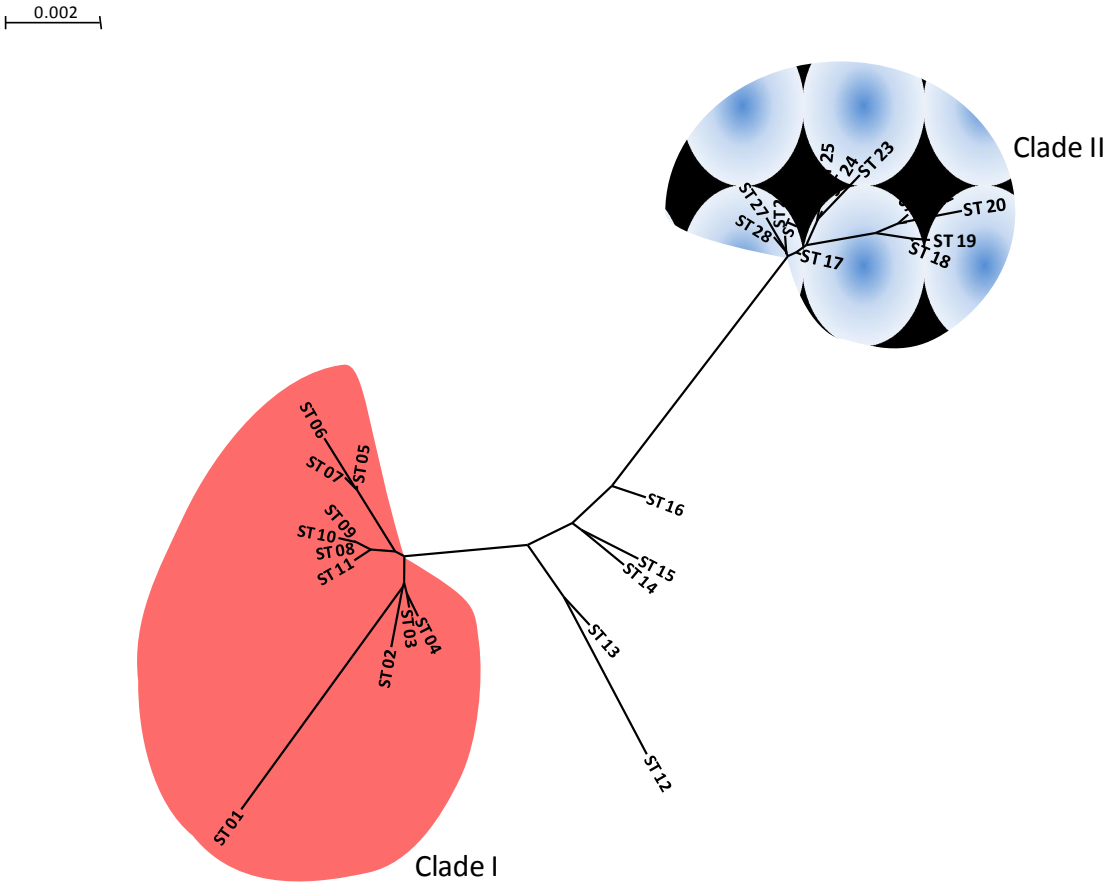

0.002
Clade II
ST 25
ST 24
ST 23
ST 27
ST 28
ST 20
ST 19
ST 18
ST 17
ST 16
ST 15
ST 14
ST 13
ST 12
ST 06
ST 07
ST 05
ST 09
ST 10
ST 08
ST 11
ST 04
ST 03
ST 02
ST 01
Clade I

Figure 2

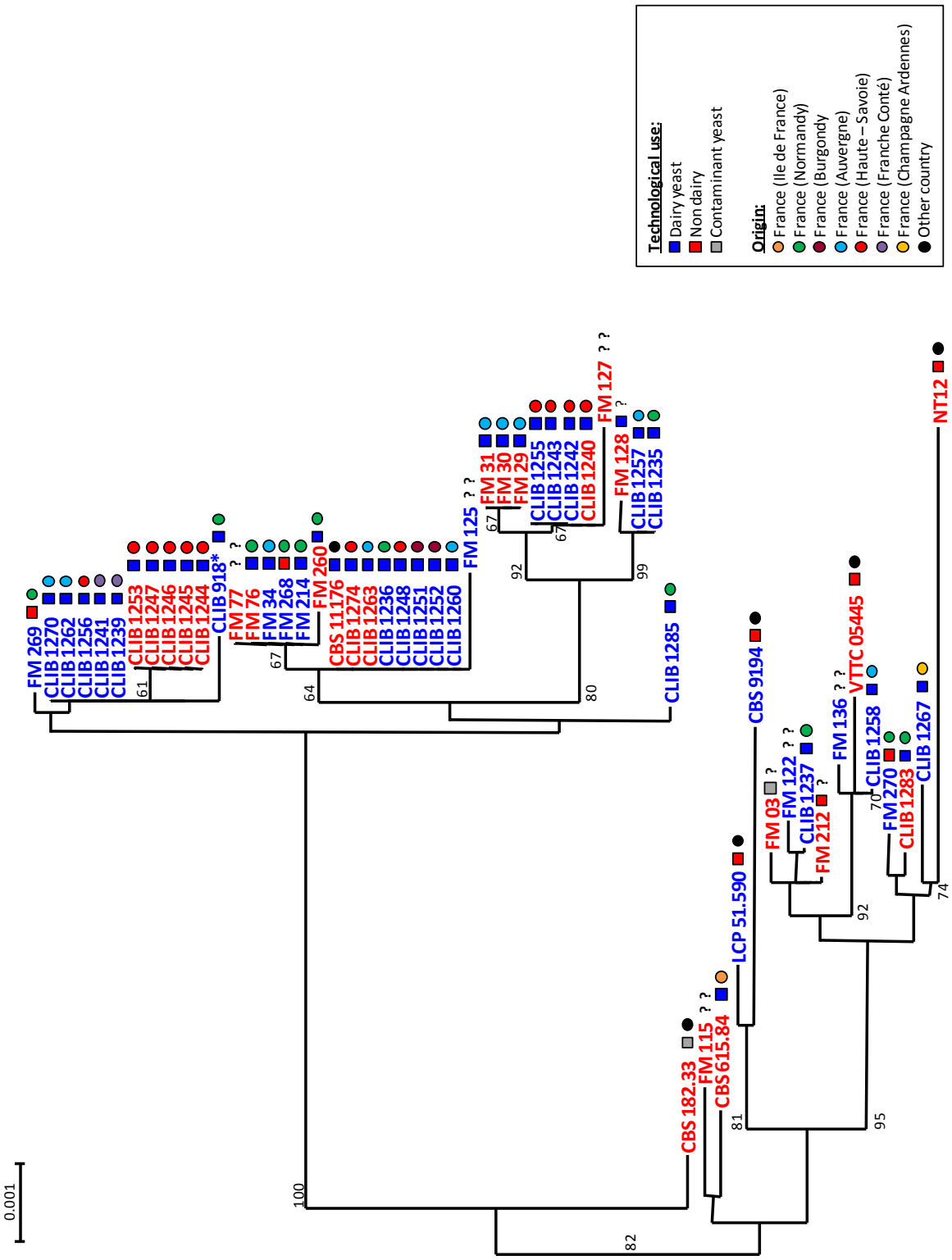

Technological use:
Dairy yeast
Non dairy
Contaminant yeast
Origin:
France (Ile de France)
France (Normandy)
France (Burgondy
France (Auvergne)
France (Haute – Savoie)
France (Franche Conté)
France (Champagne Ardennes)
Other country
FM 269
CLIB 1270
CLIB 1262
CLIB 1256
CLIB 1241
CLIB 1239
CLIB 1253
CLIB 1247
CLIB 1246
CLIB 1245
CLIB 1244
CLIB 918*
FM 77
FM 76
FM 34
FM 268
FM 214
FM 260
CBS 11176
CLIB 1274
CLIB 1263
CLIB 1236
CLIB 1248
CLIB 1251
CLIB 1252
CLIB 1260
FM 125
FM 31
FM 30
FM 29
CLIB 1255
CLIB 1243
CLIB 1242
CLIB 1240
FM 127
FM 128
CLIB 1257
CLIB 1235
CLIB 1285
CBS 9194
LCP 51.590
CBS 182.33
FM 115
CBS 615.84
FM 03
FM 122
CLIB 1237
FM 212
FM 136
VTTC 05445
CLIB 1258
FM 270
CLIB 1283
CLIB 1267
NT12
0.001

Figure 3

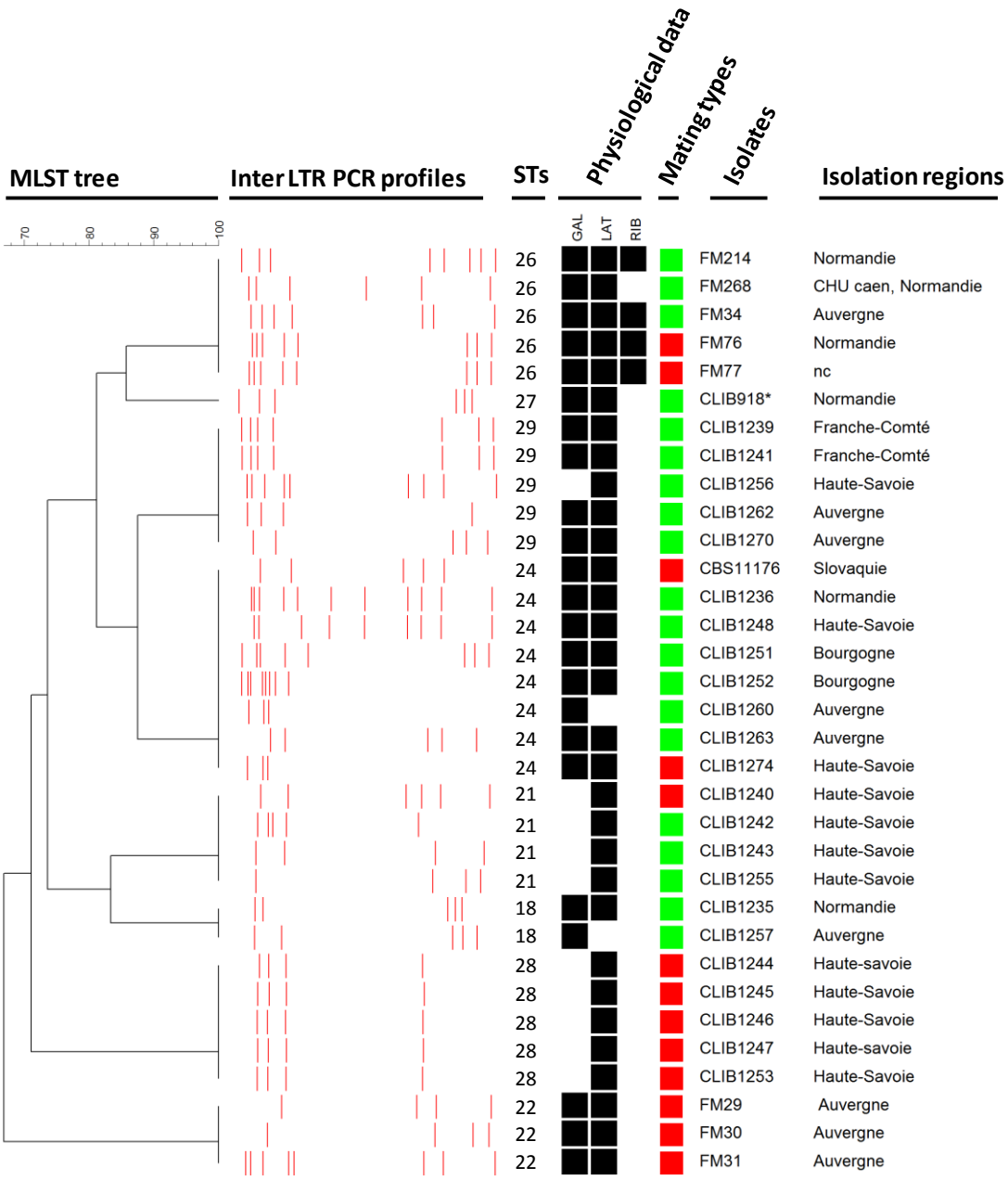

MLST tree
Inter LTR PCR profiles
STs
Physiological data
GAL
LAT
RIB
Mating types
Isolates
Isolation regions
70
80
90
100
26 FM214 Normandie
26 FM268 CHU caen, Normandie
26 FM34 Auvergne
26 FM76 Normandie
26 FM77 nc
27 CLIB918* Normandie
29 CLIB1239 Franche-Comté
29 CLIB1241 Franche-Comté
29 CLIB1256 Haute-Savoie
29 CLIB1262 Auvergne
29 CLIB1270 Auvergne
24 CBS11176 Slovaquie
24 CLIB1236 Normandie
24 CLIB1248 Haute-Savoie
24 CLIB1251 Bourgogne
24 CLIB1252 Bourgogne
24 CLIB1260 Auvergne
24 CLIB1263 Auvergne
24 CLIB1274 Haute-Savoie
21 CLIB1240 Haute-Savoie
21 CLIB1242 Haute-Savoie
21 CLIB1243 Haute-Savoie
21 CLIB1255 Haute-Savoie
18 CLIB1235 Normandie
18 CLIB1257 Auvergne
28 CLIB1244 Haute-savoie
28 CLIB1245 Haute-Savoie
28 CLIB1246 Haute-Savoie
28 CLIB1247 Haute-savoie
28 CLIB1253 Haute-Savoie
22 FM29 Auvergne
22 FM30 Auvergne
22 FM31 Auvergne

Table 1: List of strains studied

Isolate	Other names	Species	Origin	Source	Mating type
CLIB 1366	CBS 11176	*Geotrichum candidum*	Slovaquie	Cheese	MATB
CLIB 1367	CBS 178.71t	*Geotrichum candidum*	Germany	Soil polluted with oil	MATA / MATB
CLIB 1361	CBS 182.33	*Geotrichum candidum*	Italie	Yoghurt	MATB
CLIB 1368	CBS 615.84	*Geotrichum candidum*	Ile de France	Cheese	MATB
CLIB 1235	GC12	*Geotrichum candidum*	Normandy	Cheese	MATA
CLIB 1236	GC 21	*Geotrichum candidum*	Normandy	Cheese	MATA
CLIB 1237	GC37	*Geotrichum candidum*	Normandy	Cheese	MATA
CLIB 1239	GC43	*Geotrichum candidum*	Franche-Comté	Cheese	MATA
CLIB 1240	GC44	*Geotrichum candidum*	Haute-Savoie	Cheese	MATB
CLIB 1241	GC45	*Geotrichum candidum*	Franche-Comté	Cheese	MATA
CLIB 1242	GC42	*Geotrichum candidum*	Haute-Savoie	Cheese	MATA
CLIB 1243	GC51	*Geotrichum candidum*	Haute-Savoie	Cheese	MATA
CLIB 1244	GC59	*Geotrichum candidum*	Haute-savoie	Cheese	MATB
CLIB 1245	GC60	*Geotrichum candidum*	Haute-Savoie	Cheese	MATB
CLIB 1246	GC63	*Geotrichum candidum*	Haute-Savoie	Cheese	MATB
CLIB 1247	GC64	*Geotrichum candidum*	Haute-savoie	Cheese	MATB
CLIB 1248	GC74	*Geotrichum candidum*	Haute-Savoie	Cheese	MATA
CLIB 1249	GC76	*Geotrichum candidum*	Franche-Comté	Cheese	MATA / MATB
CLIB 1251	GC79	*Geotrichum candidum*	Burgondy	Cheese	MATA
CLIB 1252	GC84	*Geotrichum candidum*	Burgondy	Cheese	MATA
CLIB 1253	GC90	*Geotrichum candidum*	Haute-Savoie	Cheese	MATB
CLIB 1254		*Geotrichum candidum*	nc	nc	MATA / MATB
CLIB 1255	GC96	*Geotrichum candidum*	Haute-Savoie	Cheese	MATA
CLIB 1256	GC97	*Geotrichum candidum*	Haute-Savoie	Cheese	MATA
CLIB 1257	GC100	*Geotrichum candidum*	Auvergne	Cheese	MATA
CLIB 1258	GC101	*Geotrichum candidum*	Auvergne	Cheese	MATB
CLIB 1260	GC105	*Geotrichum candidum*	Auvergne	Cheese	MATA
CLIB 1262	GC110	*Geotrichum candidum*	Auvergne	Cheese	MATA
CLIB 1263	GC120	*Geotrichum candidum*	Auvergne	Cheese	MATB
CLIB 1267	GC129	*Geotrichum candidum*	Champagne-Ardenne	Cheese	MATA
CLIB 1270	GC146	*Geotrichum candidum*	Auvergne	Cheese	MATA
CLIB 1274	GC164	*Geotrichum candidum*	Haute-Savoie	Cheese	MATB
CLIB 1283		*Geotrichum candidum*	Normandy	Cheese	MATB
CLIB 1284		*Geotrichum candidum*	Normandy	Raw cream	MATA / MATB
CLIB 1285		*Geotrichum candidum*	Normandy	Dairy	MATA
CLIB 918*		*Geotrichum candidum*	Normandy	Cheese	MATA
FM 03		*Geotrichum candidum*	nc	Cheese contaminant	MATB
FM 115		*Geotrichum candidum*	nc	nc	MATB
FM 119		*Geotrichum candidum*	nc	nc	MATA / MATB

Table 1 (suite): List of strains studied

Isolate	Other names	Species	Origin	Source	Mating type
FM 122		*Geotrichum candidum*	nc	nc	MATA
FM 125		*Geotrichum candidum*	nc	nc	MATA
FM 127		*Geotrichum candidum*	nc	nc	MATB
FM 128		*Geotrichum candidum*	nc	nc	MATB
FM 136		*Geotrichum candidum*	nc	nc	MATA
FM 212		*Geotrichum candidum*	France	Corn silage	MATB
FM 213		*Geotrichum candidum*	nc	nc	MATA / MATB
FM 214		*Geotrichum candidum*	Normandy	Cow milk	MATA
FM 260		*Geotrichum candidum*	Isigny, Normandy	pis	MATB
FM 267		*Geotrichum candidum*	CHU caen, Normandy	Stools	MATA / MATB
FM 268		*Geotrichum candidum*	CHU caen, Normandy	Stools	MATA
FM 269		*Geotrichum candidum*	CHU caen, Normandy	Stools	MATA
FM 270		*Geotrichum candidum*	CHU caen, Normandy	Stools	MATA
FM 29		*Geotrichum candidum*	Auvergne	Cheese	MATB
FM 30		*Geotrichum candidum*	Auvergne	Cheese	MATB
FM 31		*Geotrichum candidum*	Auvergne	Cheese	MATB
FM 34		*Geotrichum candidum*	Auvergne	Cheese	MATA
FM 76		*Geotrichum candidum*	Normandy	Cow milk	MATB
FM 77		*Geotrichum candidum*	nc	nc	MATB
LCP 51.590		*Geotrichum candidum*	Spain	Sand	MATA
CLIB 1378	CBS 9194	*Geotrichum silvicola*	Brasil	Insect	MATA
NT 12		*Geotrichum silvicola*	Thailand	Rain forest	MATA
VTTC 4559		*Geotrichum silvicola*	Sweden	Malting system	MATB

[a]Abbreviations: CBS, Centralbureau voor Schimmelcultures, Utrecht, the Netherlands; CLIB : Centre international de ressources microbiologiques Thiverval-Grignon, France; LCP, Museum National d'Histoire Naturelle, Laboratoire de Cryptogamie Paris, France; FM, Food Microbiomes private collection; VTTC : Technical research center of Finland ; GC: Marcelino et *al.* (2001) iolates *Sequenced strain

Table 2: Characteristics of the five loci studied

Locus	Size of the amplicon examined (bp)*	Sequence divergence (%)	Number of polymorphic sites	Number of AP per locus
NUP116	423	2.84	12	6
URA1	465	1.51	7	5
URA3	470	2.13	10	5
SAPT2	514	4.28	22	11
SAPT4	501	4.19	21	5

*Bases of the sequences were discarded from the examination to ensure a clear reading of the chromatograms. AP, allelic profile

Table 3: .Genotypes and Sequence Types of the 55 *G. candidum* isolates tested

Strains	Markers					STs
	NUP116	***SAPT2***	***SAPT4***	***URA1***	***URA3***	
NT 12	6	11	2	5	4	ST 01
CLIB 1267	2	5	4	1	1	ST 02
CLIB 1283	2	3	2	1	1	ST 03
FM 270	2	3	2	4	3	ST 04
CLIB 1258	1	4	2	2	1	ST 05
VTTC 4559	1	4	2	2	5	ST 06
FM 136	1	3	2	2	1	ST 07
FM 212	2	4	2	2	1	ST 08
CLIB 1237	2	3	2	2	1	ST 09
FM 122	3	3	2	2	1	ST 10
FM 03	5	4	2	2	1	ST 11
CBS 9194	4	2	3	3	1	ST 12
LCP 51.590	2	10	5	2	3	ST 13
CBS 615.84	3	1	3	2	1	ST 14
FM 115	3	6	5	1	1	ST 15
CBS 182.33	2	1	2	1	1	ST 16
CLIB 1285	2	1	1	4	1	ST 17
CLIB 1235	1	1	1	2	1	ST 18
CLIB 1257	1	1	1	2	1	ST 18
FM 128	1	9	1	2	1	ST 19
FM 127	1	8	1	1	1	ST 20
CLIB 1240	1	1	1	1	1	ST 21
CLIB 1242	1	1	1	1	1	ST 21
CLIB 1243	1	1	1	1	1	ST 21
CLIB 1255	1	1	1	1	1	ST 21
FM 29	1	1	1	1	3	ST 22
FM 30	1	1	1	1	3	ST 22
FM 31	1	1	1	1	3	ST 22
FM 125	2	7	1	2	1	ST 23
CBS 11176	2	1	1	2	1	ST 24
CLIB 1236	2	1	1	2	1	ST 24
CLIB 1248	2	1	1	2	1	ST 24
CLIB 1251	2	1	1	2	1	ST 24
CLIB 1252	2	1	1	2	1	ST 24
CLIB 1260	2	1	1	2	1	ST 24
CLIB 1263	2	1	1	2	1	ST 24
CLIB 1274	2	1	1	2	1	ST 24
FM 260	3	1	1	2	3	ST 25

Table 3 (suite): .Genotypes and Sequence Types of the 55 *G. candidum* isolates tested

Strains	Markers					STs
	NUP116	***SAPT2***	***SAPT4***	***URA1***	***URA3***	
FM 214	2	1	1	2	3	ST 26
FM 268	2	1	1	2	3	ST 26
FM 34	2	1	1	2	3	ST 26
FM 76	2	1	1	2	3	ST 26
FM 77	2	1	1	2	3	ST 26
CLIB 918*	2	1	1	1	2	ST 27
CLIB 1244	5	1	1	1	1	ST 28
CLIB 1245	5	1	1	1	1	ST 28
CLIB 1246	5	1	1	1	1	ST 28
CLIB 1247	5	1	1	1	1	ST 28
CLIB 1253	5	1	1	1	1	ST 28
CLIB 1239	2	1	1	1	1	ST 29
CLIB 1241	2	1	1	1	1	ST 29
CLIB 1256	2	1	1	1	1	ST 29
CLIB 1262	2	1	1	1	1	ST 29
CLIB 1270	2	1	1	1	1	ST 29
FM 269	2	1	1	1	3	ST 30

Table 4: Sequence Type (ST) variability; Percent of SNPs between pairs of sequence types (2373 bp); the mean percentage of divergence for all strains is 0.90%

	Clade I											Intermediate				
	ST01	**ST02**	**ST03**	**ST04**	**ST05**	**ST06**	**ST07**	**ST08**	**ST09**	**ST10**	**ST11**	**ST12**	**ST13**	**ST14**	**ST15**	**ST16**
ST01	*															
ST02	0.72	*														
ST03	0.63	0.17	*													
ST04	0.72	0.25	0.08	*												
ST05	0.93	0.46	0.38	0.38	*											
ST06	0.97	0.59	0.51	0.51	0.13	*										
ST07	0.97	0.51	0.34	0.34	0.04	0.17	*									
ST08	0.72	0.25	0.17	0.17	0.21	0.34	0.25	*								
ST09	0.76	0.29	0.13	0.13	0.25	0.38	0.21	0.04	*							
ST10	0.8	0.34	0.17	0.17	0.29	0.42	0.25	0.08	0.04	*						
ST11	0.76	0.29	0.21	0.21	0.25	0.38	0.29	0.04	0.08	0.13	*					
ST12	1.18	0.88	0.8	0.88	1.01	1.14	1.05	0.8	0.84	0.88	0.84	*				
ST13	1.22	0.76	0.67	0.59	0.72	0.84	0.76	0.51	0.55	0.59	0.55	0.46	*			
ST14	1.43	0.97	0.88	0.88	0.93	1.05	0.97	0.72	0.76	0.72	0.76	0.67	0.38	*		
ST15	1.18	0.72	0.63	0.72	0.93	1.05	0.97	0.72	0.76	0.72	0.76	0.59	0.38	0.25	*	
ST16	1.05	0.59	0.51	0.59	0.8	0.93	0.84	0.59	0.63	0.67	0.63	0.8	0.59	0.38	0.29	*
ST17	1.77	1.31	1.22	1.22	1.43	1.56	1.47	1.22	1.26	1.31	1.26	1.43	1.14	0.93	0.93	0.72
ST18	2.06	1.6	1.52	1.52	1.14	1.26	1.18	1.35	1.39	1.43	1.39	1.64	1.26	1.05	1.22	1.01
ST19	2.06	1.6	1.52	1.52	1.14	1.26	1.18	1.35	1.39	1.43	1.39	1.64	1.31	1.1	1.26	1.05
ST20	2.06	1.6	1.52	1.6	1.39	1.52	1.43	1.6	1.64	1.69	1.64	1.73	1.52	1.31	1.22	1.01
ST21	1.94	1.47	1.39	1.47	1.26	1.39	1.31	1.47	1.52	1.56	1.52	1.6	1.39	1.18	1.1	0.88
ST22	1.98	1.52	1.43	1.43	1.31	1.43	1.35	1.52	1.56	1.6	1.56	1.64	1.35	1.22	1.14	0.93
ST23	1.98	1.52	1.43	1.43	1.47	1.6	1.52	1.26	1.31	1.35	1.31	1.56	1.18	0.97	1.14	0.93
ST24	1.85	1.39	1.31	1.31	1.35	1.47	1.39	1.14	1.18	1.22	1.18	1.43	1.05	0.84	1.01	0.8
ST25	1.94	1.47	1.39	1.31	1.43	1.56	1.47	1.22	1.26	1.22	1.26	1.52	1.05	0.84	1.01	0.88
ST26	1.9	1.43	1.35	1.26	1.39	1.52	1.43	1.18	1.22	1.26	1.22	1.47	1.01	0.88	1.05	0.84
ST27	1,64	1,35	1,26	1,35	1,56	1,68	1,6	1,35	1,39	1,43	1,39	1,47	1,26	1,05	0,97	0,76
ST28	1.77	1.31	1.22	1.31	1.52	1.64	1.56	1.31	1.35	1.39	1.26	1.43	1.22	1.01	0.93	0.72
ST29	1.73	1.26	1.18	1.26	1.47	1.6	1.52	1.26	1.31	1.35	1.31	1.39	1.18	0.97	0.88	0.67
ST30	1.77	1.31	1.22	1.22	1.52	1.64	1.56	1.31	1.35	1.39	1.35	1.43	1.14	1.01	0.93	0.72

	Clade II													
	ST17	**ST18**	**ST19**	**ST20**	**ST21**	**ST22**	**ST23**	**ST24**	**ST25**	**ST26**	**ST27**	**ST28**	**ST29**	**ST 30**
ST17	*													
ST18	0.29	*												
ST19	0.34	0.04	*											
ST20	0.38	0.25	0.29	*										
ST21	0.25	0.13	0.17	0.13	*									
ST22	0.29	0.17	0.21	0.17	0.04	*								
ST23	0.21	0.34	0.38	0.59	0.46	0.51	*							
ST24	0.08	0.21	0.25	0.46	0.34	0.38	0.13	*						
ST25	0.17	0.29	0.34	0.55	0.42	0.38	0.21	0.08	*					
ST26	0.13	0.25	0.29	0.51	0.38	0.34	0.17	0.04	0.04	*				
ST27	0,13	0,42	0,46	0,42	0,29	0,34	0,34	0,21	0,29	0,25	*			
ST28	0.08	0.38	0.42	0.38	0.25	0.29	0.29	0.17	0.25	0.21	0.13	*		
ST29	0.04	0.34	0.38	0.34	0.21	0.25	0.25	0.13	0.21	0.17	0.08	0.04	*	
ST30	0.08	0.38	0.42	0.38	0.25	0.21	0.29	0.17	0.17	0.13	0.13	0.08	0.04	*

Table S1: primers used in this study

		PRIMER SEQUENCE 5'-3'	Melting temperature (°C)
MLST Loci			
NUP116	Fw:	ACCGCTACAACTGGATTTGG	40
	Rv:	GAGACCTGTTTGAGGGCTTG	
SAPT2	Fw:	AGACCAACCGCTACTGTGCT	40
	Rv:	TGTCAGCACCTCTTCACTGG	
SAPT4	Fw:	ATCATTAACACCCCGGCATA	40
	Rv:	GTGTCACCAAGCAGAGCAAA	
URA1	Fw:	CAAGCCAATTGTGCTGAGAA	35
	Rv:	GGTGTCGTAGGGCAGTTGAT	
URA3	Fw:	GCCAAAAAGACCAACCTGTG	37
	Rv:	CCTCATCCATACGGTTCTGC	
MATING TYPE			
GECA_MATA	Fw:	GCATTCCAAAATAAAGCTGCTC	52
	Rv:	CACACTTCAGTCTATCCTAACTATC	
GECA_MATB	Fw:	ACGACGAAAACCCAACAC	50
	Rv:	TGCTCGAAGAAGCCAAAC	
GECA_MAT_ext	Fw:	CGCAGTAGAGAAGAATCGTAG	50
	Rv:	TCAAGACAACGGAGATGGAG	
INTER-LTR			
GC1_LTR_for		TCAACAATGGAATCCCAAC	37
GC2_LTR_rev		CATCTTAACACCGTATATGA	

Table S2: Intra Marker variability for each allele profile, in percent

URA1	**(465 bp)**										
	1	**2**	**3**	**4**	**5**						
1	*										
2	0,65	*									
3	1,08	1,29	*								
4	0,22	0,43	1,29	*							
5	0,43	1,08	0,65	0,65	*						
SAPT4	**(501 bp)**										
	1	**2**	**3**	**4**	**5**						
1	*										
2	3,19	*									
3	3,79	1,00	*								
4	3,39	0,20	1,20	*							
5	3,59	0,80	0,20	1,00	*						
URA3	**(470 bp)**										
	1	**2**	**3**	**4**	**5**						
1	*										
2	0,43	*									
3	0,21	0,64	*								
4	1,49	1,06	1,70	*							
5	0,64	1,06	0,85	1,70	*						
NUP116	**(423 bp)**										
	1	**2**	**3**	**4**	**5**	**6**					
1	*										
2	1,18	*									
3	1,41	0,24	*								
4	1,41	0,24	0,47	*							
5	1,41	0,24	0,47	0,47	*						
6	2,12	0,94	1,18	1,18	1,18	*					
SAPT2	**(514 bp)**										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1	*										
2	1,75	*									
3	2,53	1,56	*								
4	2,33	1,36	0,19	*							
5	2,72	1,75	0,58	0,39	*						
6	0,39	1,36	2,14	1,95	2,33	*					
7	0,58	2,33	3,11	2,92	3,31	0,97	*				
8	0,58	2,33	3,11	2,92	3,31	0,97	1,17	*			
9	0,19	1,75	2,53	2,33	2,72	0,58	0,78	0,78	*		
10	1,36	0,39	1,56	1,36	1,75	0,97	1,95	1,95	1,56	*	
11	2,53	1,56	0,39	0,19	0,58	2,14	3,11	3,11	2,53	1,56	*

Chapitre 3

Séquençage du génome de *Geotrichum candidum* CLIB 918

1 INTRODUCTION

Afin d'augmenter les données pour les analyses méta-génomique du projet ANR ALIA food-microbiomes et d'obtenir l'une des rares levures majeures des fromages à ne pas avoir été séquencées, nous avons décidé de séquencer le génome de la levure *G. candidum*. Afin d'utiliser au mieux les données de séquençage, il est nécessaire de fournir des données d'annotations fiables de tout le génome ainsi que d'isoler et de caractériser les éléments génique comme l'ADN mitochondrial et les transposons.

Un premier travail expérimental et de bibliographie a consisté à choisir une souche représentative de l'espèce. Notre choix s'est porté sur la souche CLIB 918 = ATCC 204307, car c'est une souche française, haploïde, isolée de fromage (Pont l'Evêque) et qui a fait l'objet de nombreuses études à la fois technologiques, physiologiques et moléculaires.

Les résultats du séquençage et de l'annotation du génome seront analysés dans une première partie. Dans une seconde partie sera présenté l'assemblage et l'annotation du génome mitochondri*al.* Enfin dans une dernière partie nous tacherons de mieux comprendre pourquoi *G. candidum* a longtemps été confondu avec un champignon filamenteux. Nous traiterons alors des spécificités de son génome, présence de transposons, contenu en gène et mating type.

2 MATERIELS ET METHODES

2.1 Séquençage et assemblage du génome de *G. candidum* CLIB 918

L'ADN génomique de *G. candidum* CLIB 918 = ATCC 204307 a été isolé selon méthode standard d'extraction au phénol chloroforme décrit par (Jacques et *al.*, 2009). La séquence a été générée par le Génoscope (Evry, France). La stratégie choisie combine (i) un « run » de 454 pyrosequencing, Titanium, 8 kbp mate-pair, sur le séquenceur Roche Genome Sequencer FLX, (ii) un « run » de 454 pyrosequencing standard (Roche) et (iii) un « run » de séquençage Illumina (Solexa).

L'assemblage de l'ensemble des séquences « 454 » obtenues a été effectué en utilisant le logiciel Newbler Assembler software version 2.3 (Génoscope, Evry, France). Les reads ont donc été assemblées en contigs puis les contigs en scaffolds. Ensuite, les « reads » illumina ont été utilisées pour corriger les possibles erreurs de séquençage 454 et permettent l'obtention un génome de très bonne qualité.

2.2 Détection des éléments transposables

Afin de détecter les éléments transposables du génome de *Geotrichum candidum*, dont les reads avaient été préalablement écartées de l'assemblage comme toutes les séquences répétées. L'assemblage a été soumis à un pipeline d'annotation et de détection des transposons. En utilisant la méthode d'identification des transposons (de novo TE identification) sur le pipeline "REPET" (http://urgi.versailles.inra.fr) en appliquant le protocole standard (Flutre et *al.*, 2011). Parallèlement, les reads séquences répétées ont été comparées a une banque de séquence de transposons par blast. Les reads ainsi récupéré ont pu être assemblées grâce au package phred/Phrap/Consed. La structure secondaire du transposon MITE été déterminée en utilisant le programme mfold v3.2 avec les paramètres par défaut (Zuker, 2003).

2.3 Annotation du génome nucléaire

La prédiction des gènes du génome de *G. candidum* CLIB 918 été réalisée en utilisant la plate-forme d'annotation génomique de l'URGI, comprenant des pipelines, des bases de données et des interfaces, développés pour les champignons (http://urgi.versailles.inra.fr/). La prédiction des gènes a été réalisée en utilisant le pipeline Eugene v.4 (Foissac et *al.*, 2008). Les modèles de gènes prédits par Eugène reposent sur une combinaison de plusieurs méthodes *in silico* (*ab initio* et de similarité).

Les logiciels de prédiction de gènes a*b initio* sont Eugene_IMM (Schiex et *al.*, 2001) les modèles sont trouvées selon les probabilités des séquences codantes ou non) et SpliceMachine (Degroeve et *al.*, 2005) qui prédit les CDS les sites d'épissages d'intron. Ces méthodes ont été combinée avec une analyse par similarité de séquence (BLASTX) sur les bases de données de protéines fongiques. Les différents résultats ont été utilisés par Eugène afin de prévoir des modèles de gènes. Les trois logiciels *ab initio* de prédiction de gènes ont été « entrainés » à l'aide d'un ensemble de gènes annotés manuellement et obtenu après assemblage de données de séquençage RNAseq en utilisant l'assembleur SOAP *de novo* (Li et *al.*, 2008). Au total un ensemble de 1300 gènes a été utilisé pour entrainer Eugène. Un tiers de l'ensemble a été utilisé pour les logiciels *ab initio*, un tiers pour optimiser les paramètres d'Eugène et le dernier tiers pour calculer la précision d'Eugène. Eugène a prédit un ensemble de 6948 gènes dans le génome de *G. candidum* CLIB 918. Une première annotation fonctionnelle automatique a été effectuée sur la base de recherches BLASTP avec une longueur équivalente de 90% minimum et une e-value <e-10. Toutes prédictions automatiques sont en cours de vérification par curation manuelle par le consortium d'annotation de *G. candidum* sur la plateforme BOGAS (http://bioinformatics.psb.ugent.be/webtools/bogas/). Les ARNt ont été prédites par tRNAscan-SE (Lowe et Eddy, 1997) avec les paramètres par défaut.

2.4 Confirmation des séquences introniques et de l'annotation structurale.

Trois extractions d'ARN total issus de milieux différents ont été effectuées selon le protocole d'extraction décrit par (Mansour et *al.*, 2008) et *pooler* ensemble pour le séquençage. Les trois milieux de culture utilisées sont YPD (milieu complet), YNB_{N5000} (milieu minimum), et SCM (milieu fromage synthétique) (Leclercq-Perlat et *al.*, 2000). La qualité des extractions d'ARN a été confirmée sur la plateforme bio analyseur Agilent 2100 (Agilent Technologies). Les banques d'ADNc ont été construites grâce au soutien de *Génolevures*. Le séquençage d'ARN total a été effectué par le Génoscope. Enfin l'ensemble des « reads » a été intégré au logiciel GenomeView (http://genomeview.org/) et lié la plateforme d'annotation BOGAS pour faciliter la curation manuelle de la structure de chaque gène.

2.5 Assemblage et annotation du génome mitochondrial

Les reads correspondant aux séquences des extrémités des contigs et scaffolds correspondant a l'ADN mitochondrial ont été identifiées par blastn, Les traces ont été extraites et assemblées en utilisant le package phred/phrap/consed (Ewing et Green, 1998; Ewing et *al.*, 1998; Gordon et *al.*, 1998). Un premier assemblage à alors pu être créé et le génome a pu être alors re-circularisé. Dans

un second temps, l'assemblage a été confirmé par amplification PCR. Les produits de PCR attendus ont été obtenus (

Figure 25, page 118).

Pour annoter le génome mitochondrial, le programme BLASTX (Altschul et *al.*, 1997) a été utilisé pour comparer le génome mitochondrial (ADNmt) de *G. candidum* avec les autres protéines d'ADNmt d'autre levures : *S. cerevisiae* (Foury et *al.*, 1998), *Candida albicans* (Anderson et *al.*, 2001), *Yarrowia lipolytica* (Kerscher et *al.*, 2001), *Kluyveromyces lactis* (Zivanovic et *al.*, 2005), *Lachancea thermotolerans* (Talla et *al.*, 2005), *Candida glabrata* (Koszul et *al.*, 2003). Les gènes ribosomiques ont été identifiés par blastn contre l'ADNmt de *Y. lipolytica* et *S. cerevisiae*, tandis que les gènes de les ARNt ont été isolés en utilisant le programme : tRNAscan-SE (Lowe et Eddy, 1997) avec les paramètres par défaut.

2.6 PCR, séquençage et assemblages additionnels

Les amorces ont été déterminées avec le logiciel Primer3 (http://fokker.wi.mit.edu/primer3) (**Tableau 7**). Les produits PCR ont été séquencés sur les deux brins par Eurofins MWG Operon (Ebersberg, Germany). Le package phred/phrap/consed (Ewing et Green, 1998; Ewing et *al.*, 1998; Gordon et *al.*, 1998) a été utilisé pour les assemblages additionnels.

Noms	Séquences des oligonucléotides
3025_for	3'-AGCGCTTTGCTAAGTTCTCC-5'
3025_rev	3'-CGTAATGTAAACCGACACAGG-5'
3164_rev	3'-TGAATTAGAGCTTCATTCCCAAG-5'
3164_for	3'-AATCCCATAACTCCTAATCCTGT-5'
3141_rev	3'-TGAATAACATCAACACCACCAG-5'
3141_for	3'-CAAGAATGAAAGGACTTGAACCA-5'
3026_rev	3'-CCAATTAAGAACGAAGCAAAGC-5'
3026_for	3'-GGACAAGGAATTTCGCTACA-5'

Tableau 7 : Liste des amorces utilisées pour confirmer l'assemblage in silico du génome mitochondrial de *G. candidum* CLIB 918

2.7 Recherche d'orthologues

Les protéines orthologues à celles de *G. candidum* ont été recherchées dans les génomes de levures et champignons filamenteux séquencés grâce à une étape de BLASTP avec une P-value établie de maximum 1E-5. Les génomes et les bases de données utilisées dans ce travail sont listés dans le **Tableau 8**.

Espèces	Souches	Sources
Ashbya gossypii	ATCC10895	Ashbya Gossypi Db
Aspergillus fumigatus	Af293	NCBI
Aspergillus nidulans	FGSC A4	BROAD
Candida glabrata	CBS138	NCBI
Candida lusitaniae	ATCC42720	BROAD
Coccidioides immitis	RS	BROAD
Cryptococcus neoformans	JEC21	NCBI
Debaryomyces hansenii	CBS767	NCBI
Fusarium graminearum	PH-1	BROAD
Geotrichum candium	CLIB 918	BOGAS
Kluyveromyces lactis	CLIB210	NCBI
Komagataella pastoris	CBS 7435	BOGAS
Lachancea kluyveri	CBS 3082	Génolevures
Lachancea thermotolerans	CBS 6340	Génolevures
Magnaporthe grisea	70-15	BROAD
Millerozyma sorbitophila	CBS 7064	Génolevures
Neurospora crassa	OR74A	BROAD
Ogataea parapolymorpha	CBS 4732	JGI
Penicillium chrysogenum	Wisconsin54-1255	JGI
Phanerochaete chrysosporium	RP-78	JGI
Saccharomyces bayanus	623-6C	MIT
Saccharomyces cerevisiae	S288C	NCBI
Saccharomyces paradoxus	Weihenstephan 34/70	MIT
Scheffersomyces stipitis	CBS 6054	JGI
Schizosaccharomyces pombe	972h	NCBI
Sclerotinia sclerotiorum	1980 UF-70	BROAD
Trichoderma reesei	QM6a	JGI
Ustilago maydis	521	BROAD
Yarrowia lipolytica	CLIB122	NCBI
Zygosaccharomyces rouxii	CBS 732	Génolevures

Tableau 8 : Liste des génomes et bases de donnée utilisés dans cette étude

Pour affirmer les orthologues fungiques de *Geotrichum candidum* et la voie métabolique dans la quelle les gènes sont impliqués, la base de donnée Fungipath a été utilisée (Grossetete et *al.*, 2010).

2.8 Analyses phylogénétiques

Les séquences des protéiques orthologues ont été alignées en utilisant MUSCLE (Edgar, 2004) avec les valeurs par défauts. Les régions protéiques alignées sans ambigüité ont été conservées, les gaps dans l'alignement de séquence ont été supprimés. Les analyses phylogénétiques ont été effectuées en utilisant la méthode de maximum de vraisemblance PHYML v2.4.4 (Guindon et Gascuel, 2003). Une analyse de bootstrap a été effectuée avec 100 réplications pour s'assurer de la robustesse de chaque nœud.

3 RESULTATS ET DISCUSSION

3.1 Assemblage du génome de *G. candidum* CLIB 918

Une séquence de grande qualité a été obtenue en utilisant une stratégie combinant plusieurs méthodes de séquençage: (i) 454 pyrosequencing Titanium sur le séquenceur FLX (Roche), (ii) 8 kbp mate-pair 454 pyrosequencing et (iii) séquençage illumina (Solexa). Le séquençage effectué au Génoscope a fournis 3.152.715 lectures de haute qualité, soit 850.062.477 bases (tableau 2). L'assemblage de novo des reads a été effectué en utilisant le logiciel Newbler. Celui-ci a généré 1.688 contigs de plus de 500 pb. La taille cumulée des contigs et de 23,2 Mb. 94,8% des reads ont été alignés et la couverture de l'assemblage est estimée à 36X. Grâce aux données Mate-pair, 1370 contigs ont été assemblés en 134 scaffolds. 41 ont une taille supérieure à 50 kb. La taille cumulée des scaffolds est de 24,8 Mb. Afin de palier les erreurs éventuelles inhérentes à la technologie 454 notamment dans les homopolymères, les reads illumina (taux d'erreur <0.4%) ont été cartographiées et alignées sur le génome. La taille du génome est donc au minimum de 24,8 Mb mais pourrait être supérieure si l'on ajoutait les transposons entiers. Ce qui fait du génome de *G. candidum* le plus grand génome de levure aujourd'hui séquencé, la taille des plus grands génomes connus sont 20Mb pour *Yarrowia lipolytica* et 21,27 Mb pour *Lipomyces starkeyi.*

Les principales données issues du séquençage du génome sont présentées dans le **Tableau 9**.

Caractéristiques du projet de séquençage	**Résultats**	
Nombre de reads	3 322 644	
Nombre de bases séquencées	850 062 477	
Nombre de reads alignées	3 152 715	**(94.8%)**
Taille de la séquence assemblée (pb)	23 255 818	
Nombre de contigs	1 688	
Plus grand contig (pb)	141 117	
N50 des contigs (pb)	26 778	
Nombre de scaffolds	134	
Nombre de contigs entiers dans les scaffolds	1 370	**(81.1%)**
Taille cumulée des scaffolds	24 865 483	
N50 des scaffolds (pb)	1 159 651	
Moyenne G+C (%)	44	

Tableau 9 : Résultats du séquençage et de l'assemblage du génome *G. candidum* CLIB 918

L'analyse de l'assemblage final n'a pas révélé la présence du locus rDNA dans l'assemblage. Cependant à partir de l'analyse des séquences répétées, nous avons pu extraire des reads présentant de l'homologie de séquence avec les rDNA connus et nous avons pu assembler manuellement une partie de celui-ci. Les premières analyses ont montré que à l'instar de *Sch. pombe* et *Y. lipolytica*, *G. candidum* possède au moins deux clusters de rDNA. Le 5S est lui aussi absent de l'assemblage. Il est vraisemblablement inclus dans les clusters de rDNA. La complexité des séquences présentant de l'homologie avec les rDNA suggère qu'il existe plus de deux clusters. De plus il a été montre que *G. candidum* possède un polymorphisme dans sa séquence D1D2 (Alper et *al.*, 2011).

Le génome mitochondrial est fragmenté en 4 contigs les contigs numéro 3025, 3164, 3141 et 3026. Ceux-ci ont été identifiés via le biais de GC% et grâce au nombre élevé de reads composant ces contigs (Voir partie 3.3, page 117).

3.2 Annotation du génome de *G. candidum* CLIB 918

Les 130 scaffolds de *G. candidum* CLIB 918 ont été automatiquement annotés en utilisant le pipeline d'annotation Eugene. Apres des recherches Blastn et BlastX contre la base de données NCBI GenBank 100 scaffolds ne possèdent pas de séquence codant pour des protéines et n'ont pas été analysés plus avant. Par conséquent, nous nous sommes intéressés au 30 scaffolds restants ; les annotations des 27 plus grand scaffolds ont été transférées sur la plateforme d'annotation BOGAS. Les 27 scaffolds ont une taille totale de 24,2MB et contiennent 6663 gènes codant pour une protéine. Les autres scaffolds constitués de contigs dont la taille est inférieure à 2 kb ont été concaténés en tant que nouveau scaffold chimérique (scaffold 32). Au total, 6802 séquences codantes pour des protéines (CDS) ont été prédites par Eugène. 38,2% de la séquence génomique est considéré comme codant pour une protéine. La taille des scaffolds et leur contenu en gène est présenté dans le **Tableau 10**.

Scaffold #	taille (nt)	nombre de gènes	Scaffold #	taille (nt)	nombre de gènes
scaffold 01	2.581.591	680	scaffold 15	711.411	232
scaffold 02	2.106.585	553	scaffold 16	665.493	187
scaffold 03	1.796.091	487	scaffold 17	608.101	165
scaffold 04	1.686.461	495	scaffold 18	562.025	163
scaffold 05	1.653.926	469	scaffold 19	507.124	126
scaffold 06	1.264.337	327	scaffold 20	368.852	112
scaffold 07	1.211.666	333	scaffold 21	313.304	93
scaffold 08	1.159.651	314	scaffold 22	313.014	92
scaffold 09	1.063.388	266	scaffold 23	307.402	83
scaffold 10	918.757	236	scaffold 24	271.541	100
scaffold 11	909.630	254	scaffold 25	237.622	79
scaffold 12	890.415	264	scaffold 26	220.840	64
scaffold 13	871.522	235	scaffold 27	219.481	51
scaffold 14	795.198	203	scaffold 32	620.629	139

Tableau 10 : Taille des scaffolds et leur contenu en gènes

68% des gènes sont des gènes composés d'un seul exon, 22% des gènes présentent deux exons. Les gènes restants contiennent 2 introns (7%), 3 introns (2%), 4 introns (0,6%) et 5 introns ou plus (0,3%), voir aussi **Tableau 11**. A nouveau, *G. candidum* se démarque des autres levures de part la quantité de gènes introniques, 32% des gènes de *G. candidum* possèdent au moins un intron, ils représentent 14,9% des gènes de *Y. lipolytica*, 11,4% chez *P. pastoris*, 5% de *S. cerevisiae.*

Contrairement aux autre levures, une mauvaise conservation des séquences consensus d'épissage 5' été observée, les bases, GT en 5' et AG en 3' sont quant à elles conservées **(Figure 24**). La taille moyenne des introns est de 136pb. Une première analyse indique que la région S2 séparant le point de branchement et le consensus d'épissage est de 8-12 pb. Sur les 2180 gènes introniques, 688 possèdent des introns en position 3'.

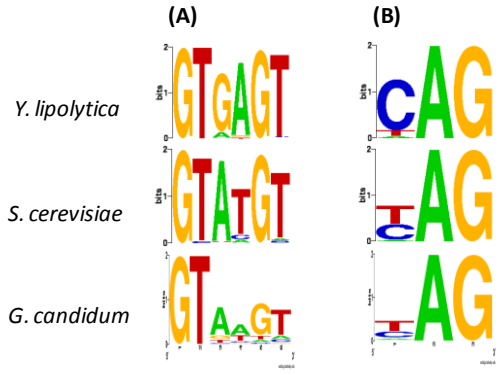


Figure 24 : Séquences consensus des sites d'epissages d'introns 5' (B) et 3' (A) obtenu grâce à WebLogo (http://weblogo.berkeley.edu/logo.cgi)

Une pré-annotation fonctionnelle automatique a été effectuée sur la base d'une conservation de longueur de 90% et d'une e-value de 1E-5 par rapport aux résultats BlastP sur une banque comprenant les génomes de levures et champignons filamenteux listés dans le **Tableau 8**. Avec cette méthode, 4997 descriptions de gènes ont été assignées aux CDS. 1646 CDS sont toujours annotées « unknown protein » (**Tableau 11**).

Caractéristiques de l'annotation	Résultats	
Nombre de gènes codant des protéines	6802	
Nombre de gènes possédant un seul exon	4622	**68%**
Nombre de gènes possédant plusieurs exons	2180	**32%**
Gènes avec deux exons	1499	
Gènes avec trois exons	467	
Gènes avec quatre exons	134	
Gènes avec cinq exons ou plus	61	
Gènes avec annotation référant à *S. cerevisiae*	3293	**48%**
Gènes avec annotation référant à une autre levure	1430	**21%**
Gènes avec annotation référant à un champignon filamenteux	274	**4%**
Nombre de gènes annotés entant que "hypothetical protein"	155	**2,20%**
Nombre de gène annotés entant que "unknown protein"	1646	**24%**

Tableau 11: Résultats de l'annotation structurale Eugene et de l'annotation fonctionnelle

3.3 Le génome de l'ADN mitochondrial, assemblage et annotation

Quatre scaffolds mitochondriaux ont été détectés dans l'assemblage du génome de *G. candidum* CLIB 918 sur la base de leur GC% et de leur haute concentration en reads. L'appartenance de ces 4 scaffolds au mtDNA a été confirmée par Blastn contre la base de données NCBI GenBank. Les 4 scaffolds correspondent aux contigs numéro 3025, 3026, 3141 et 3164. Ces quatre scaffolds ont été assemble et annoté à la main. La combinaison de deux méthodes m'a permis d'assembler les contigs et de re-circulariser le génome mitochondrial : (i) les reads et traces jouxtant chaque scaffolds ont été récupérés puis assemblées en utilisant le package phred/Phrap/Consed. L'assemblage des reads ainsi obtenus à permis d'identifier la séquence des autres scaffolds. (ii) L'assemblage des scaffolds a été validé par PCR (

Figure 25).

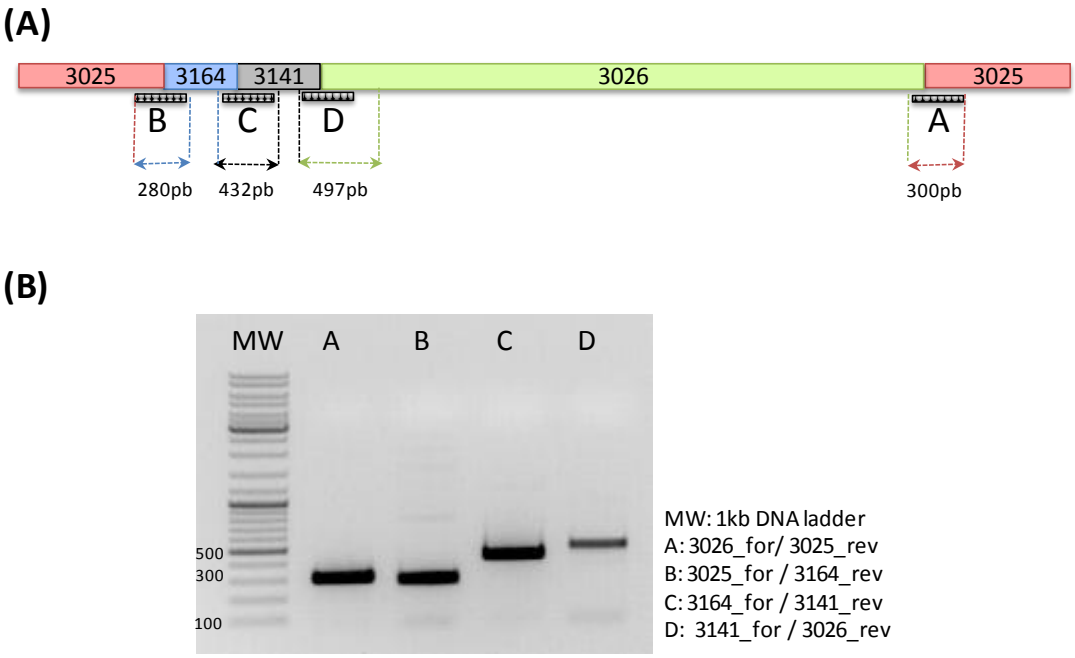


Figure 25 : Reconstruction in sillico (A) et validation par méthode PCR (B) de l'assemblage du génome mitochondrial de *G. candidum* CLIB 918

La taille du génome mitochondrial de *G. candidum* CLIB 918 est de 29 kb, ce qui est 18,9 kb plus petit que celui de *Y. lipolytica* (Kerscher et *al.*, 2001)(AJ307410.1). Mais qui est compris entre la taille du génome mitochondrial de *C. glabrata* 20 kb (Koszul et *al.*, 2003), plus petit génome mitochondrial chez les levures hémiascomycètes et celui de *Podospora anserina*, champignon filamenteux, 100kb. Le GC % du génome mitochondrial de *G. candidum* (28%) est supérieur à celui de *Y. lipolytica* (22,7%).

L'annotation du génome mitochondrial de *G. candidum* CLIB 918 a permis de détecter 7 gènes codant pour des protéines *COB COXI COXII COXIII*, *ATP8*, *ATP6*, *ATP9* et 6 gènes du complex I, ubiquinone oxidoreductase complex, *NADH1 NADH2 NDH3 NADH4 NADH5 NADH6* ainsi que 23 tRNA et 2 gènes du rRNA (**Figure 26**). Etonnamment, un seul intron du groupe 2 à été trouvé dans le gène codant pour la protéine *COB1*, celui-ci codant pour sa propre GIY-YIG endonuclease. Maintenant que de nombreux génomes mitochondriaux sont disponibles, le fait de posséder peu d'introns est moins surprenant, mais cela reste quand même rare puisque seul *Candida phangngensis* propose un seul gène intronique *COX1* comportant deux introns (Gaillardin et *al.*, 2012).Comme souvent observé chez les champignons ascomycètes, tous les gènes sont codés sur le même brin, mais des exceptions sont notables, chez S. cerevisiae, ARNt (thr1) se trouve sur le brin opposé (Foury et *al.*, 1998) ou chez *D. hansenii* , où l'on trouve une insertion (Lépingle et *al.*, 2000). Remarquablement, le génome mitochondrial ici étudié, possède un gène codant pour la protéine ribosomique VAR1, une protéine

homologue à RPM1 (codant pour une sous unité de la RNaseP mitochondrial) retrouvée habituellement chez les espèces du genre *Saccharomyces* (Langkjaer et *al.*, 2003).

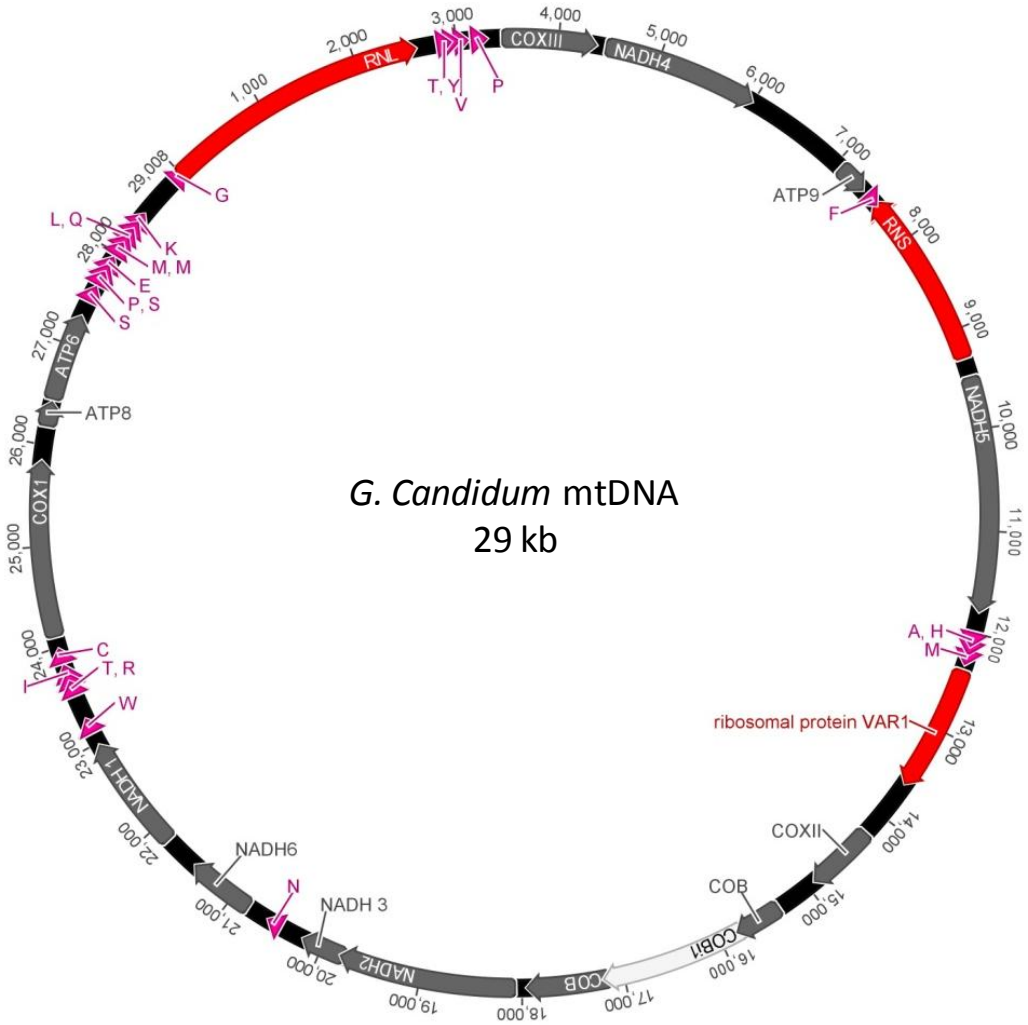


Figure 26 : Carte du génome mitochondrial *G.candidum* CLIB 918. Les ORFs (Open Reading Frame) exoniques sont présentés en gris, les ORFs fonctionnels introniques en blanc, l'ARN ribosomiques (RNL, RNS) et VAR1 en rouge. Les ARNt et le code une lettre des acides aminés sont indiqués en rose.

3.4 Position phylogénétique de *G. candidum* dans l'arbre des champignons ascomycètes

G. candidum a longtemps été considéré comme un champignon filamenteux. Alors qu'il est à présent certain que cette espèce soit une levure hémiascomycète, et bien que de nombreuse analyses phylogénétiques impliquant *G. candidum* ait été effectuées (de Hoog et Smith, 2004; Kurtzman et Robnett, 1995). Le placement de *G. candidum* est relativement incertain et ne dépend que de l'ADN ribosomique. Nous avons donc voulu placer au sein de l'arbre global des champignons, en complément de notre analyse de quatre gènes (voir chapitre I).

Les 246 protéines sélectionnées pour leur performance phylogénétique par (Aguileta et *al.*, 2008) ont été isolées chez 28 espèces fongiques (**Tableau 8**). Les protéines ont ensuite été alignées séparément avec le programme clustalX puis concaténées en une seule et unique séquence pour chaque espèce. Après une curation Gblocks, une phylogénie par maximum de vraisemblance phyML a été construite avec un bootstrap de 100 sur un alignement de 66.981 acides aminés (**Figure 27**). Tous les nœuds sont bien supportées par de haute valeur de bootstrap et et la topologie de l'arbre est en agrément avec les phylogénies précédemment publiées (Aguileta et *al.*, 2008; Fitzpatrick et *al.*, 2006). La topologie de l'arbre montre que l'espèce la plus proche de *G. candidum* est *Y. lipolytica*. Voisin tout relatif comme nous le montre la longueur des branches séparant ces deux espèces. Ces deux espèces forment entre elles un groupe distinct des autres saccharomycotina à la partie basale de l'arbre des hémiascomycètes.

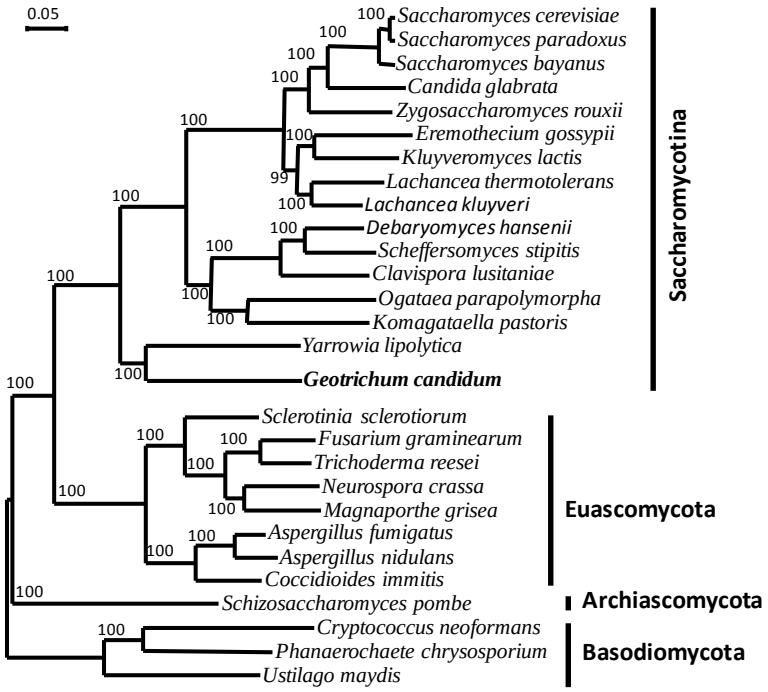


Figure 27 : Arbre phylogénétique des 28 champignons ascomycètes obtenu grace aux 246 protéines simple copie définies par (Aguileta et *al.*, 2008).

3.5 Analyse du contenu en élément transposable

Plusieurs types de transposons ont pu être trouvés dans le génome de *G. candidum* CLIB 918. Des éléments de classe I (transposons à ARN) ont pu être identifiés dans le génome, ce sont des éléments de type « Ty like », rétro transposons à LTR (***L**ong **T**erminal **R**epeats*) ou « LINE » (***L**ong **I**nterspersed **N**uclear **E**lement*). Nous avons identifié peu de transposon de type TY3/gypsy et ceux-ci étaient dégénérés. Par contre, sur la base du nombre de reads, nous avons pu identifier un grand nombre de Ty5-like que nous avons appelé Tgc5. Comme pour les clusters d'ADN ribosomique, l'assembleur a écarte toutes les séquences répétées, dont la plupart des transposons. Ceux-ci sont responsables de la plupart des gaps du génome. On trouve à proximité de ceux-ci des fragments de transposases. Néanmoins, des éléments associés à Tgc5, ses LTR, sont présents dans le génome. Nous nous sommes d'ailleurs servis de ces séquences pour mettre au point la méthode de typage PCR inter-LTR (voir Chapitre 2). De façon remarquable, tous ces éléments LTR sont localisés dans des clusters que l'on retrouve dans 11 scaffolds (un cluster par scaffold, sauf pour le scaffold 4 qui a deux clusters). Excepté les clusters du scaffold 4, tous les clusters de LTR sont localisés en début ou en fin de scaffolds, associés à des gaps. Ceci rappelle la situation chez *D. hansenii* dans laquelle les Ty5 sont aussi localisés en cluster dans les chromosomes, à raison de un cluster par scaffold (Neuveglise et Casaregola, non publlié; Lynch et *al.*, 2010).

Contrairement aux autre levures (Bleykasten-Grosshans et Neuveglise, 2011), plusieurs familles de transposons de Classe II (transposons à ADN) ont pu être identifiés grâce au pipeline REPET. Nous avons trouvé des éléments « Mutator-like » et « Tc-like ».

Etonnamment, le pipeline REPET a permis d'isoler du génome un élément inédit chez la levure, un MITE (*Miniature inverted-repeat transposable element*). Les éléments MITEs sont abondants dans les génomes des plantes et des animaux, cependant peu sont rencontrés chez les champignons ascomycètes. Ils ont été décrits chez *Fusarium oxysporum* (Bergemann et *al.*, 2008) or *Epichloë festucae* (Fleetwood et *al.*, 2011), par exemple. La taille de l'élément trouvé chez *G. candidum* est de 457 pb riche avec un GC% de 36% et est présent en 5 copies dans le génome. Plusieurs reliques sont identifiables par Blast. Les MITEs décrits par ailleurs ont pour taille moyenne 246 pb et un GC% de 24%, caractéristiques proche du MITE trouvé chez *G. candidum*. Les MITEs sont des éléments non autonomes, ils ne possèdent pas de gènes essentiels à la transposition, c'est-à-dire de gènes codant pour une transposase. Chez *G. candidum* et les autres éléments décrits montrent la capacité de former une structure secondaire stable en tête d'épingle (*hairpin-like secondary structures*), voir **Figure 28**. La composition en élément transposable de *G. candidum* est remarquable par sa grande diversité, et rappelle le contenu en transposons des génomes fongiques: *copia*, *gypsy*, *LINE*, *Tc-like*, *Mutator-like* et *MITE*. L'analyse initiale indique que ces nombreux éléments peuvent être

responsables des gaps dans la séquence du génome. Certains de ces éléments, Tc-like ou Mutator peuvent être impliqués dans les événements HGT.

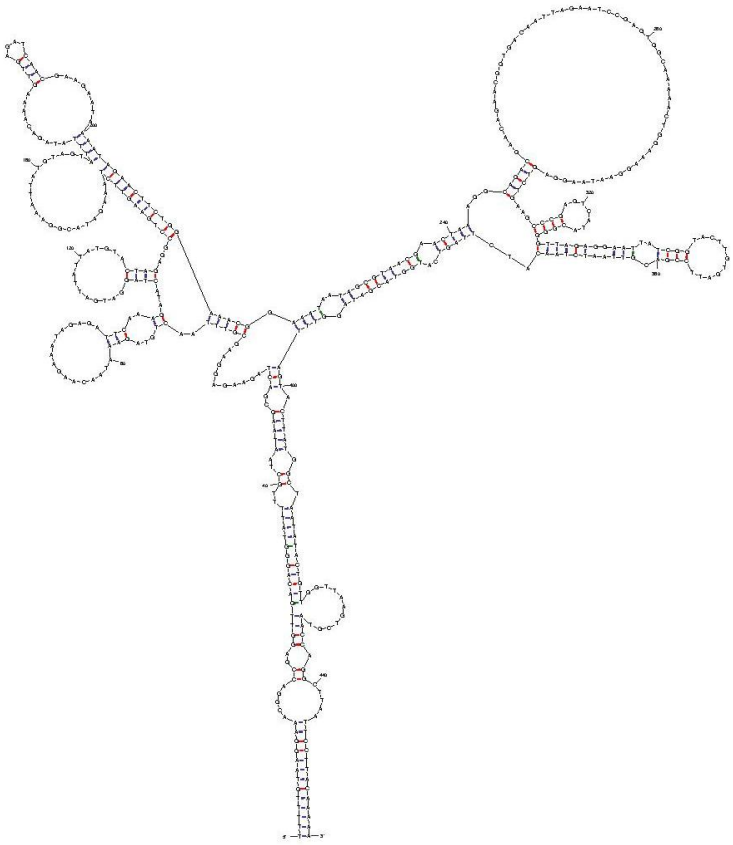


Figure 28 : Structure secondaire du MITE trouvé chez *G. candidum* ΔG=-30.85

3.6 Analyse du contenu en gène du génome

3.6.1 Gènes dupliqués chez G. candidum CLIB 918

En étudiant l'annotation fonctionnelle automatique du génome effectué sur BOGAS, j'ai pu déterminer 1209 gènes à l'annotation identique et dont au moins 529 gènes possèdent au moins deux copies au sein du génome. Parmi ces 529 gènes, 384 possèdent une annotation *Gene Ontology* (GO). La distribution des GO les plus fréquemment trouvés sont présentées dans la **Figure 29**. La plupart des gènes dupliqués sont impliqués des le trafique membranaire, un processus métabolique. 24 % de ces protéines n'ont pas de Gène Onthology associé.

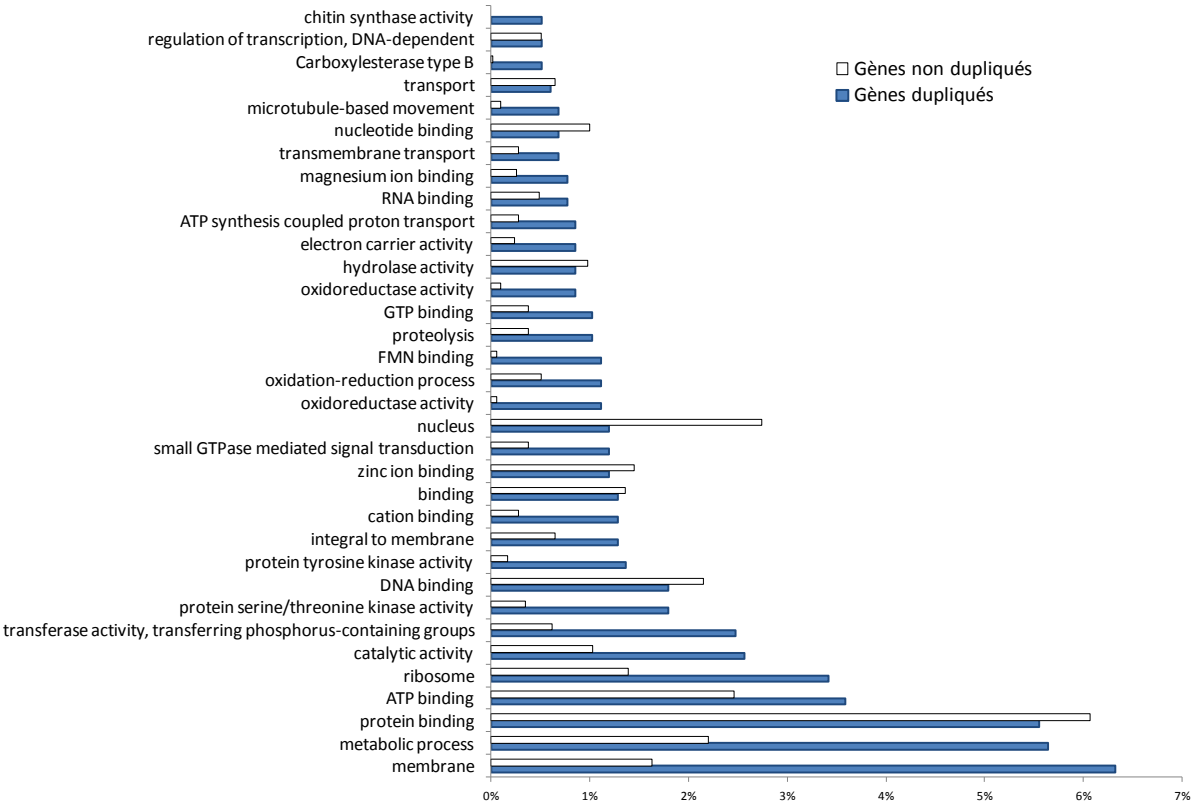


Figure 29 : Distribution des GO les plus retrouvées chez les gènes dupliqués et leur distribution parmi les gènes non dupliquées

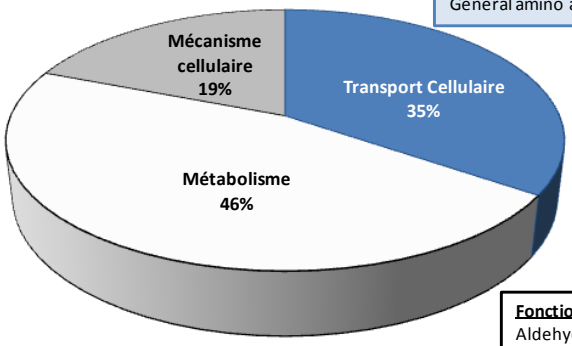
On retrouve 1 gène présent en 10 copies, il s'agit d'un gène homologue au gène ARI1 codant pour une aldéhyde réductase NADPH-dépendante. Ce gène a été identifié comme un gène qui participe à la tolérance de la levure et la détoxification de furfural mais aussi peut être impliqué dans la réduction de certains inhibiteurs générés pendant le processus de conversion de la biomasse lignocellulosique (Liu et Moon, 2009). On retrouve aussi 10 gènes orthologues au gène de *S. cerevisiae* YGL157W/ARI1 chez *Y. lipolytica* (Dujon et *al.*, 2004).

Parmi les 6 gènes présents en 6 copies, on retrouve des gènes codant pour des perméases, des réductases, une endo polygalacturonase ou une estérase. On retrouve ensuite 12 gènes présents en 5 copies, 17 en 4 copies, 46 en 3 copies et 448 en double copies. Les gènes présents en au moins 4 copies sont présentés dans la **Figure 30**. La grande majorité des gènes dupliqués sont impliqués dans le transport cellulaire ou le métabolisme. Même si l'on sait que le nombre de copie d'un gène n'est pas forcément corrélé avec son expression, il existe cependant dans certains cas, un effet dose. Certains de ces gènes peuvent fournir des pistes intéressantes pour la sélection de souches fromagères. Par exemple, lorsque qu'il est surexprimé chez *S. cerevisiae*, le gène *PST2*, présent en 4 copies chez *G. candidum*, induit une accumulation de S-adenosylméthionine, précurseur de la voie de

bio synthèse des polyamines. Nous retrouvons deux copies des gènes *TPO1* et *TPO2* décrit comme codant pour des tranporteurs de polyamines chez *S. cerevisiae.*

Fonction	nb de gènes
Karyogamy	4
Histidine kinase osmosensor	4
Serine-threonine phosphatase	4
Flavodoxin-like proteins	4
Multi-drug and toxin extrusion	4

Fonction	nb de gènes
Allantoate permease	6
Oligopeptide transporter	6
Nicotinic acid plasma membrane permease	5
Monocarboxylate permeases	5
Membrane multidrug transporter	5
FAD transporter	5
ABC transporter	4
Multidrug resistance transporters	4
Ferric reductase	4
General amino acid permease	4



Fonction	nb de gènes
Aldehyde reductase	10
Esterase	6
Endo-polygalacturonase	6
Peroxisomal protein	5
Amidase	5
Dihydrokaempferol 4-reductase	5
Xylose and arabinose reductase	5
Glycerol dehydrogenase	4
Aldehyde dehydrogenase	4
Chitin synthase II	5
FAD/FMN binding	4
Putative mannosidase	4

Figure 30 : Fonctions putatives des gènes retrouvés en 4 copies ou plus

3.6.2 Gènes codant pour une estérase, un exemple d'une expansion de famille de gènes GL3C3695

Un ensemble de 6 gènes appartenant à la famille Génolevures GL3C3695 (www.genolevures.org) a été identifié dans le génome de *G. candidum* CLIB 918. La famille GL3C3695 correspond aux gènes codant pour une estérase ou carboxyl esterase/lipase type B, lipase coupant les fonctions ester des lipides. L'étude phylogénétique des gènes de *G. candidum* (**Figure 31**) révèlent que l'un d'entre eux semble phylogénétiquement proche de *Y. lipolytica* tandis que les 5 autres semblent quant à eux proches des gènes d'autres levures. La différentes places de ces gènes dans l'arbre phylogénétique pourrait expliquer les différentes activités décrites chez les lipases de type B de *Geotrichum candidum*, fréquemment utilisées en industrie (Brabcova et *al.*, 2010).

La distance très proche entre les trois gènes GECA03s04685g, GECA03s04718g et GECA03s0652g peut être expliquée par une expansion récente de la famille de gène. De plus les gènes GECA03s04718g et GECA03s04685g forment une duplication en tandem.

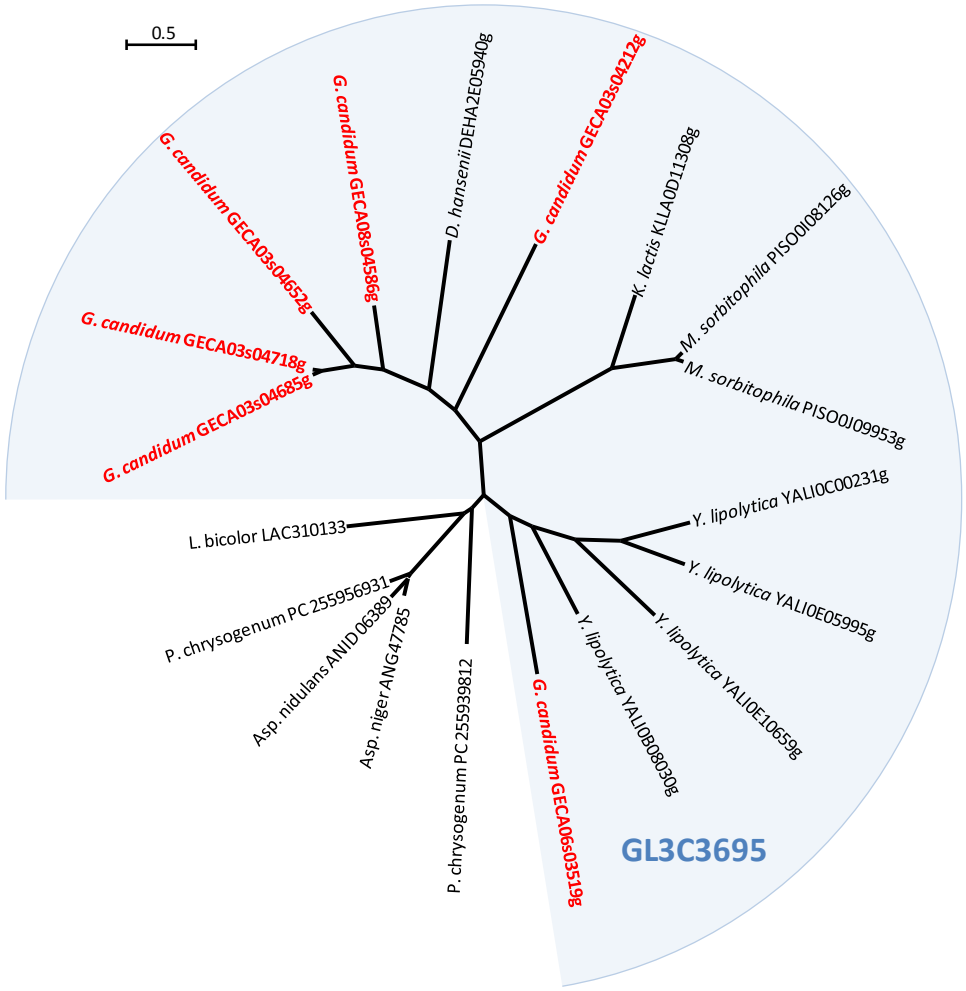


Figure 31 : Arbre phylogénétique non raciné PhyML des Carboxylesterase/lipase type B des hémiascomycètes (GL3C3695) et champignons filamenteux. La zone bleue couvre les gènes de la famille Génolevures GL3C695.

3.7 Le génome de *G. candidum* CLIB 918 révèle un haut degré de conservation de gènes de champignons filamenteux

Parmi les 6802 gènes décris par Eugène, 583 gènes ont été automatiquement annotés comme plus proche d'un gène de champignons filamenteux que d'une levure. Ils possèdent un orthologue chez les champignons filamenteux sur la base d'un résultat de BlastP avec une conservation de longueur de protéines supérieur ou égale à 90% et une e-value inferieure à $1E^{-10}$. Une pré-analyse phylogénétique a montré que non seulement 134 de ces 583 gènes sont phylogénétiquement proches des champignons filamenteux mais que ces gènes ne sont pas présents chez les levures hémiascomycètes. L'annotation structurale de ces gènes a été vérifiée gènes par gènes.

Pour déterminer si ces gènes peuvent nous donner un aperçu de l'histoire évolutive de *G. candidum*, il sera question ici, d'évaluer la distinction entre les gènes de champignons filamenteux acquis par transfert horizontal de gènes et ceux qui sont issus d'une conservation de ces gènes conservation des gènes au cours de l'évolution.

L'ensemble des protéines de *G. candidum* ont été soumis à un BlastP sur deux banques protéiques que j'ai moi-même crée. La première contient les protéines des génomes de levures auxquels ont été enlevé celle de l'espèce étudiée, la seconde contient les protéines issues des génomes de champignons filamenteux (**Tableau 8**). Dans les deux cas, seul le meilleur résultat blast avec une e-value inferieure à 1E-20 est conservé. Les espèces ont été choisies afin de couvrir au mieux l'ensemble de l'arbre phylogénétique des Ascomycètes.

Chaque séquence protéique de *G. candidum* possède alors un % de similarité par rapport à une séquence protéique de levure et/ou de champignon filamenteux. Plusieurs cas de figure se présentent dès lors. (i) La séquence protéique de *G. candidum* possède un orthologue uniquement chez les levures. (ii) La séquence protéique possède un orthologue uniquement chez les champignons filamenteux. (iii) la séquence protéique possède un orthologue chez les levures et les champignons filamenteux. Et, (iv) La séquence protéique de *G. candidum* ne possède aucune similarité avec un gène connu. Cette dernière est alors ôtée de la présente étude.

Un rapport entre le pourcentage de similarité du gène de levure et du gène de champignon filamenteux a été calculé. Ainsi plus la valeur de ce rapport tend vers zéro plus le gène est similaire à un gène de champignons filamenteux. Un rapport fixé à zéro signifie alors que le gène n'a pas d'équivalent chez les levures. Afin de comparer le résultat obtenu avec *G. candidum*, la même analyse a été effectué sur trois génomes de levures hémiascomycètes : *Y. lipolytica*, *S. cerevisiae* et *D. hansenii* (**Figure 32**).

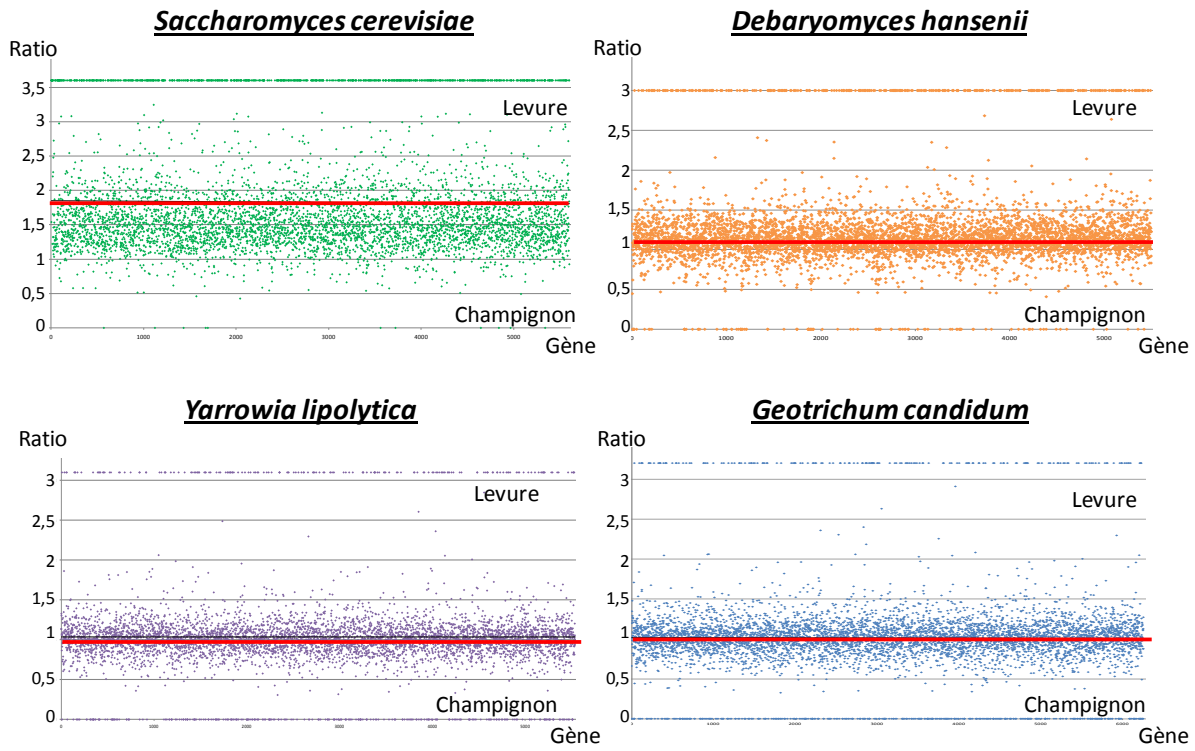


Figure 32 : Ratio entre les % de similarité Blastp levure et % de similarité Blastp champignons filamenteux. Pour les 4 espèces, tous les gènes du génome ont été soumis à l'analyse BLAST et un taux ont été calculés. A: *S. cerevisiae*, B: *D. hansenii*; C: *Y. lipolityca* et D: *G. candidum* En rouge: les moyennes des ratios de chaque espèce sont présentées**.**

Remarquablement, chez *G. candidum,* 315 gènes codant pour des protéines possèdent un orthologue chez les champignons filamenteux et non chez les levures. Ce nombre est supérieur à celui qui a été trouvé chez les autres levures : *24* chez *S. cerevisiae*, 115 chez *D. hansenii* et 287 chez *Y. lipolytica*.

L'analyse des GO des 315 gènes révèle que certains d'entre eux sont impliqués dans une fonction moléculaire (activité cellulase ou une activité catabolique, par exemple), dans un constituant cellulaire, centre de l'organisation des microtubules et d'autres semble être impliquée dans le processus biologique global (organisation du cytosquelette, régulation de la transcription ...), voir **Figure 33**. Comme nous pouvons le constater, le ratio de gènes impliqués dans le métabolisme est bien plus grand que lors de l'analyse globale du génome (46%).

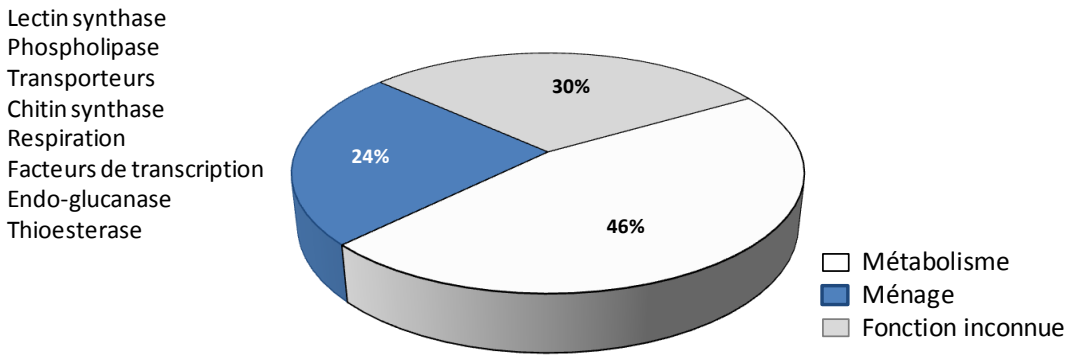


Figure 33 : Fonctions putatives des gènes de champignons filamenteux

Devant le nombre important de gènes provenant des champignons, nous avons voulu comprendre quelles en étaient les origines. Plusieurs hypothèses sont alors possibles : (i) *G. candidum* est en ancien hybride entre un champignon filamenteux et une levure ; (ii) *G. candidum* a acquis par transfert horizontal de gène des gènes de champignons ; ou (iii) *G. candidum* a conservé des gènes de champignons qui auraient été perdus chez les autres levures hémiascomycètes.

Une analyse spécifique des introns de ces gènes a été effectuée. 24% d'entre eux possèdent au moins un intron ce qui est plus faible que les 32% de gènes introniques annotés. De plus la position des introns de ces gènes ne révèle pas de présence significative en 3'. Cela signifie que le processus de perte d'intron (Dujon, 2010; Stajich et *al.*, 2007) aurait eu lieu très tôt après la différenciation avec les champignons.

Dans le cas d'un gain de gène suite à un croisement récent entre un champignon filamenteux et une levure, nous pouvons nous attendre à ce que les gènes de champignons se trouvent localisés dans le génome comme précédemment observé pour les transferts horizontaux eucaryotes vers eucaryotes (Novo et *al.*, 2009). Nous avons alors mappé les gènes de champignons sur les 5 plus grands scaffolds (**Figure 34**).

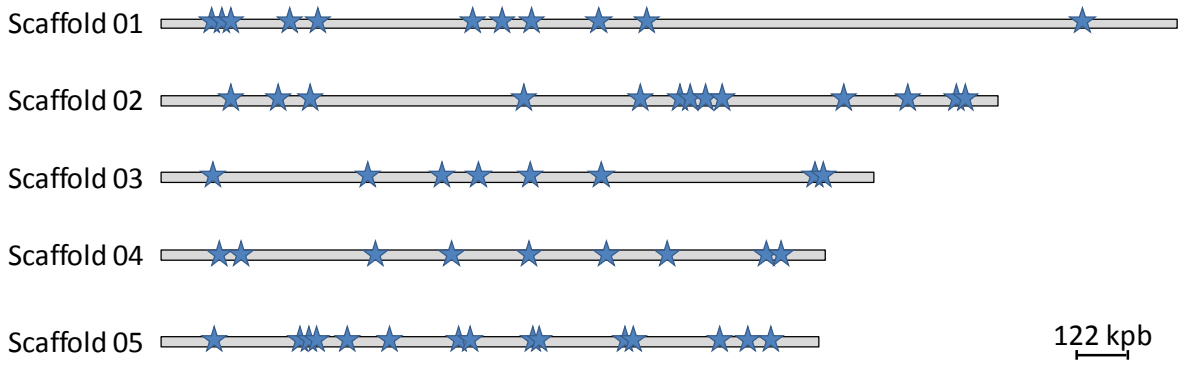


Figure 34 : Répartition des gènes d'origine fungique au sein des 5 plus grands scaffolds. Les étoiles représentent un gène d'origine fongique

Les gènes sont répandus dans le génome de *G. candidum* et il ne semble pas y avoir de régions contenant plus de gènes issus de champignon que d'autres clades. Le fait d'avoir une répartition de ces gènes dans le génome permet d'invalider l'hypothèse d'acquisition de ces gènes via un transfert horizontal tel une introgression récente. Cela peut être cependant le résultat d'une hybridation ancestrale suivie de remaniements chromosomiques massifs qui peut être amplifié par la plasticité importante du génome de *G. candidum* liée à son nombre d'éléments répétés.

L'analyse des résultats Blast ne nous montre pas de résultat satisfaisant quand au putatif donneur de gènes, cela peut être du à un manque de données génomiques de champignons. De plus pour écarter ou valider cette possibilité, les 315 séquences protéiques ont été comparées à un génome d'une espèce hémiascomycètes récemment séquencée par le JGI (http://www.jgi.doe.gov/), *Lipomyces starkeyi*. Ainsi parmi, les 315 gènes trouvés, 161 n'ont pas d'équivalent chez *L. starkeyi* et ont donc pu être conservés spécifiquement par *G. candidum* ou acquis par transfert horizontal de gène. On retrouve une partie de ces gènes dans d'autres levures éloignées.

Je préfère alors privilégier les deux autres hypothèses possibles : la conservation de gènes de champignons filamenteux perdue chez les autres levures et/ou l'existence de transferts horizontaux de gènes de champignons filamenteux vers *G. candidum*. Pour valider ces deux hypothèses une analyse phylogénétique de chacun de ces gènes est nécessaire.

Certains de ces gènes ont d'ores et déjà été analysés, et deux exemples vont être présentés dans ce manuscrit : (i) exemple des gènes codant pour une polygalacturonase et (ii) exemple d'un gène codant pour une spermine / spermidine synthases

3.8 Deux exemples de gène d'origine fungique présent chez *G. candidum* CLIB918

3.8.1 Gènes codant pour une polygalacturonase chez G. candidum CLIB 918

La pectine est un des composés assurant l'intégrité de la paroi cellulaire des végétaux supérieurs. Elle possède une structure composée de polysacharides. La dégradation de la pectine s'effectue grâce à l'action de deux enzymes, une pectine methylesterase (EC 3.1.1.11) et une polygacturonase. Les PGs peuvent avoir une activité au hasard, endo-polygacturonase (EC 3.2.1.15), ou en fin de chaine, exo-polygacturonase (EC 3.2.1.67), voir **Figure 35.**

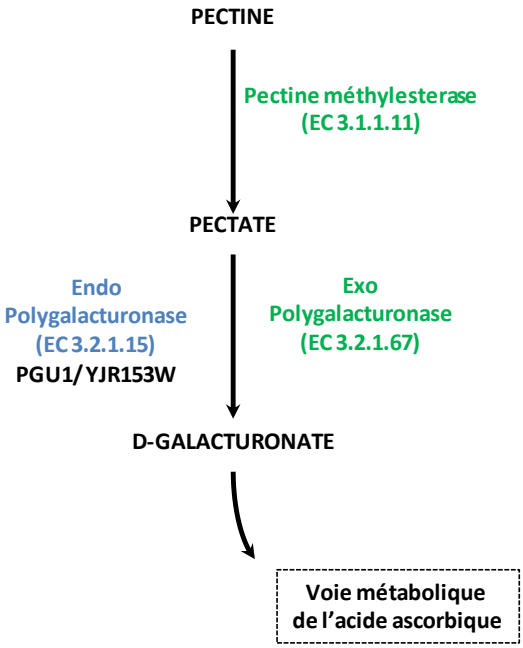


Figure 35 : Enzymes impliquées dans la dégradation de la pectine. En vert sont présentées les enzymes produites exclusivement par les champignons filamenteux, en bleu les enzymes produites par les champignons filamenteux et les levures hémiascomycètes. Le gène retrouvé chez S. cerevisiae est indiqué en gras.

Les enzymes pectinolytiques sont très fréquemment utilisées pour la fabrication de sucres issus de biomasse, la fabrication et la clarification de jus de fruits y compris les mouts de raisin, la macération de fruits ou de légumes ou la clarification d'huiles végétales (Bailey et Pessa, 1990; Serrat et *al.*, 2002).

En ce sens, les avantages possibles de l'utilisation de PG de levures a été mis en avant (REF). Celles-ci ne produisent que des PG à activité endo. Des polygacturonases ont notamment été décrites dans diverses espèces du sous embranchement des Saccharomycotina telles que *S. cerevisiae* ou une souche de *K. marxianus*. (Gainvors et *al.*, 2006; Serrat et *al.*, 2002).

Chez *G. candidum*, 6 gènes annotés comme endo-polygalacturonase ont été trouvés. Ces gènes semblent phylogénétiquement proches des champignons filamenteux. Afin de mieux comprendre l'histoire évolutive de ces gènes, une analyse phylogénétique a été effectuée. Les séquences protéiques des différentes espèces ont été alignées et les régions comportant des gaps ont été supprimées. L'arbre phylogénétique présenté en **Figure 36** a été construit sur la base d'un alignement de séquence protéique de 260 acides aminés.

L'analyse phylogénétique des séquences fongiques de polygalacturonase montre l'existence de deux familles. Les six gènes retrouvés chez *G. candidum* appartiennent tous à la même famille que celle

retrouvée chez les champignons filamenteux, à l'exception distincte de *Schizophyllum commune,* champignon basidiomycète, considéré ici comme *outgroup* (**Figure 36**).

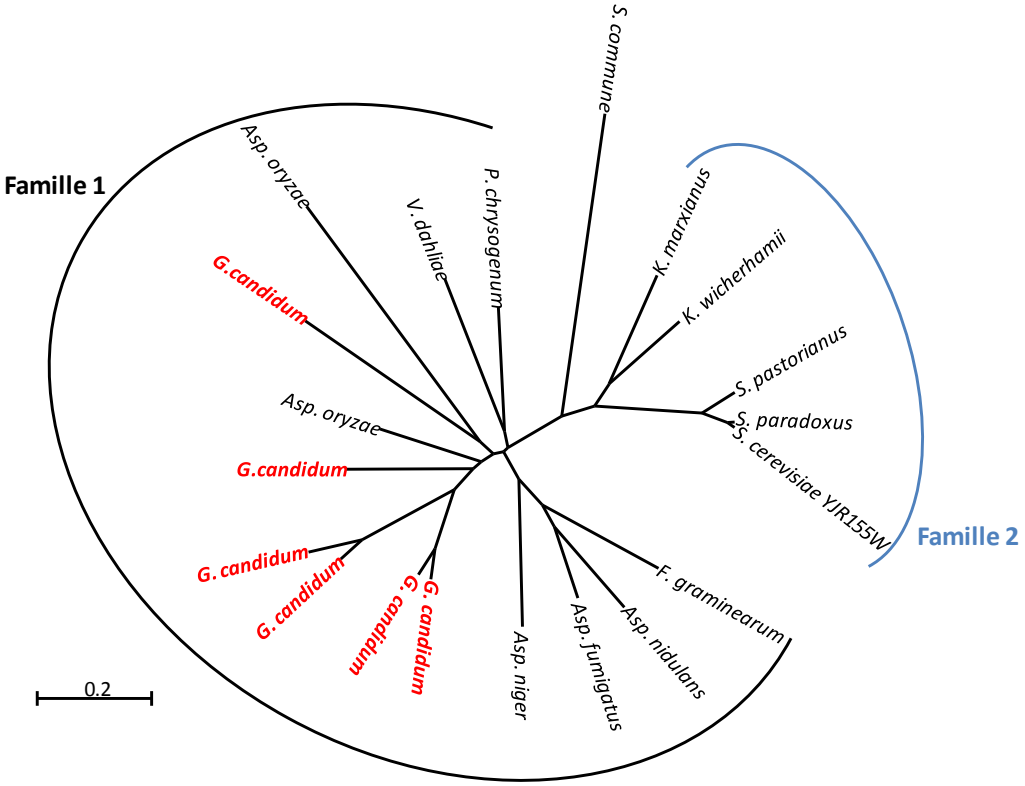


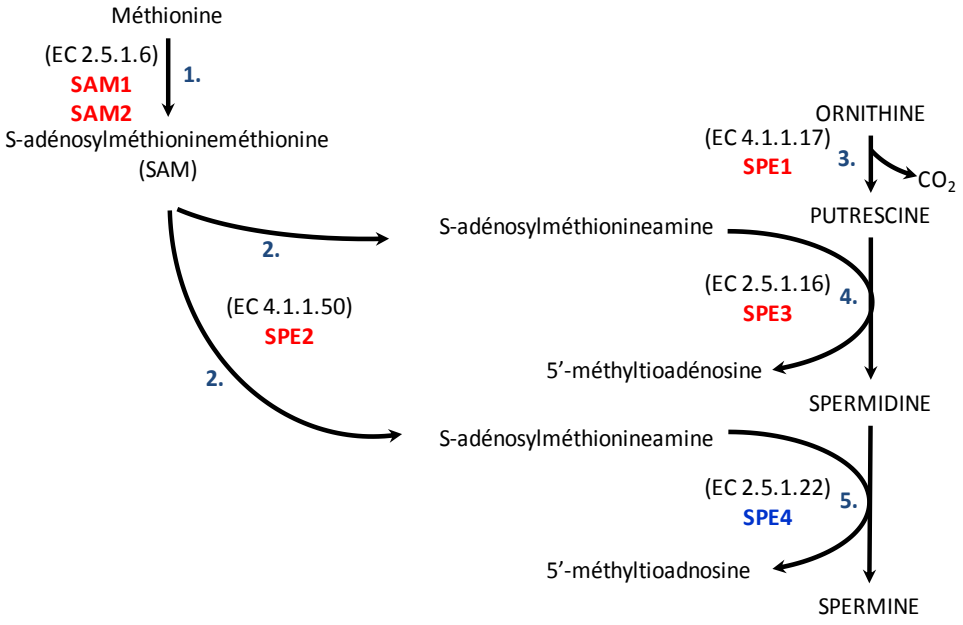
Figure 36 : Arbre phylogénétique non raciné des gènes de polygalacturonase d'origine fongique. En rouge sont présentés les gènes de *G. candidum*.

Les activités des polygalacturonases de *G. candidum* sont en train d'être testées. Cela pourra conférer à *G. candidum* un nouvel intérêt biotechnologique.

Cependant, deux cas de figures peuvent être proposés quand à l'origine évolutive de ce gène : (i) l'absence de gène de la famille 1 chez les Saccharomycotina est due à une perte de ces gènes, sauf chez *G. candidum* ou (ii) que la seule présence de ces gènes dans *G. candidum* est due à un événement HGT dont la source serait un champignon filamenteux, suivi d'un élargissement de la famille.

3.8.2 Gène codant pour la spermine/spermidine synthase chez G. candidum, un exemple de transfert horizontal chez G. candidum

La spermine et la spermidine sont des polyamines. Ces molécules non soufrées sont étroitement liées au métabolisme du soufre. Chez *S. cerevisiae*, la synthèse des polyamines nécessite l'intervention de cinq enzymes (**Figure 37**). Une S-adenosylméthionine synthase, codée par *SAM1* et *SAM2* produit de la S-adenosylméthionine (SAM) à partir de la Méthionine. Ensuite, une S-adenosylméthionine décarboxylase, codée par *SPE2*, génère de la *S*adénosylméthioninamine à partir de la S-adenosylméthionine . Cette molécule est le substrat d'une spermidine synthase et d'une spermine synthase, codées par *SPE3* et *SPE4*, conduisant respectivement à la production de spermidine et de spermine. La formation de putrescine est nécessaire à la synthèse de ces deux polyamines. Cette molécule est produite à partir d'ornithine par une ornithine décarboxylase, codée par *SPE1*. L'ornithine, dont le précurseur est l'arginine, est synthétisée au niveau du cycle de l'urée. Les enzymes codées par *SPE3* et *SPE4* produisent non seulement de la spermidine et de la spermine, mais aussi de la méthylthioadénosine (Figure 37). Ce composé, qui est impliqué dans le métabolisme de l'adénine, peut aussi servir de précurseur pour la synthèse de méthionine. Cette voie est appelée cycle de la méthylthioadénosine (ou recyclage de la méthionine).



1. S-adénosylméthionine synthétase, 2. S-adénosylméthionine décarboxylase 3. Ornithine decarboxylase; 4. Spermidine Synthase; 5. Spermine Synthase

Figure 37 : Synthèse des polyamines. Les nombre entre parenthèses correspondent aux numéros Enzyme Commission, les noms des gènes impliqués chez S. cerevisiae sont en gras, la couleur rouge signifie que les gènes possèdent des orthologues chez toutes les espèces ascomycètes, la couleur bleue signifie que les orthologues sont trouvés seulement chez les hémiascomycètes

Le gène *SPE3* codant pour la spermidine synthase est retrouvé chez toutes les espèces ascomycètes, le gène *SPE4* codant pour la spermine synthase n'est retrouvé que chez les hémiascomycètes. Les deux enzymes sont absentes de la plupart des génomes de basidiomycète à l'exception de le *Melanspora larici populina*, qui possède dans son génome une enzyme codée par un orthologue de *SPE3*.

Cependant il est retrouvé dans l'ensemble des champignons filamenteux et les basidiomycètes une troisième enzyme, la spermine/spermidine synthase. Ce gène est aussi retrouvé chez *G. candidum* CLIB 918.

De plus la recherche des gènes SPE3 et SPE4 dans le gènome de *G. candidum* a révélé que seule une de ces deux enzymes est présente dans ces deux génomes. Afin, d'établir avec certitude la fonction du gène retrouvé, une phylogénie a été effectuée avec les séquences protéiques des enzymes spermine et de spermidine synthase de différentes levures. . Cette analyse révèle que le gène codant pour la spermidine synthase est absent chez *G. candidum*. La situation de présence et absence de gènes est résumé dans le **Tableau 12**.

	Spermidine Synthase EC 2.5.1.16	**Spermine synthase EC 2.5.1.22**	**Spermine/Spermidine Synthase**
Basidiomycète	oui	non	oui
Euascomycète	oui	non	oui
Hémiascomycète	oui	oui	non
Geotrichum candidum	**non**	**oui**	**oui**

Tableau 12 : Présence des spermine et spermidine synthases dans le règne des champignons

Hamasaki *et al.* ont montré qu'un mutant déficient en spermidine synthase est auxotrophe pour les polyamines tandis qu'un mutant déficient en spermine synthase n'est pas affecté dans sa croissance et possède une morphologie normale (Hamasaki-Katagiri et *al.*, 1998). Cela est du au fait qu'un mutant ΔSPE3 est déficient en spermine et en spermidine.

Nous proposons que l'absence de spermidine synthase chez *G. candidum* doit être compensée. De ce fait nous retrouvons chez *G. candidum* un gène qui pourrait coder pour une spermine/spermidine synthase. Afin de mieux comprendre l'histoire évolutive de ce gène, une analyse phylogénétique a été effectuée. Les séquences protéiques des différentes espèces de basidiomycètes et d'euascomycètes ont été alignées et les régions comportant des gaps ont été supprimées. L'arbre phylogénétique présenté et a été construit sur la base d'un alignement de séquence protéique de 439 acides aminés voir **Figure 38**.

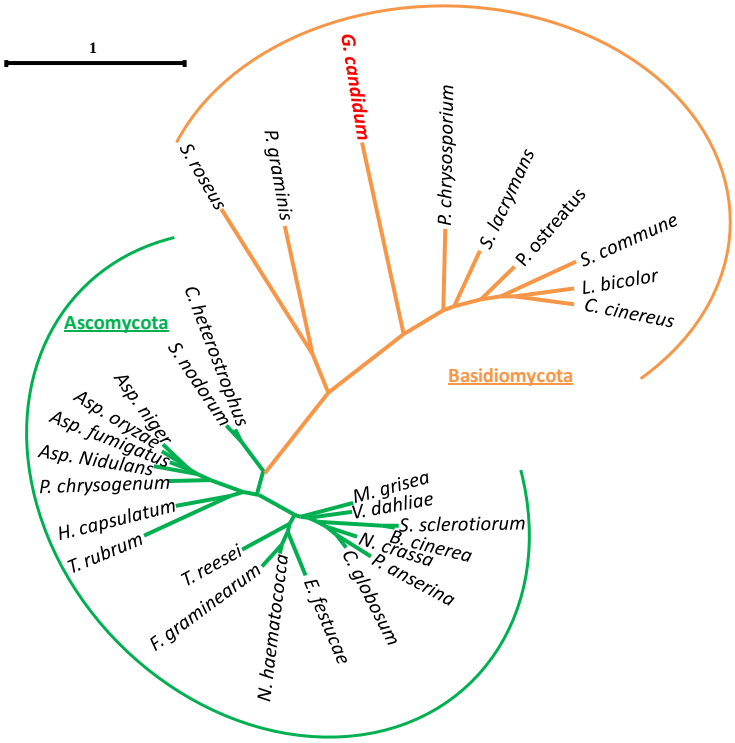


Figure 38 : Arbre non raciné PHYML de gènes de spermine/spermidine synthase retrouvé chez les ascomycètes. En orange sont présenté les branches des Basidiomycètes et vert les branches des Euascomycètes, le gène de *G. candidum* est en rouge.

La position phylogénétique du gène de *G. candidum* est ici très intéressante. En effet, celui-ci se trouve en plein milieu des Basidiomycètes. L'explication la plus vraisemblable est que ce gène a été acquis par transfert horizontal de gène dont la source serait un Basidiomycète non encore séquencé.

Les deux exemples de gènes présentés ici ainsi que le grand nombre de gènes d'origine fongique peuvent jouer un rôle dans l'adaptation de *G. candidum* aux différentes niches écologiques dans lesquelles il est fréquemment isolé. Les polygalcturonases sont des enzymes connues qui jouent un rôle dans la dégradation de la pectine et peuvent peut être conférer une meilleur adaptation aux milieux végétal. Nous retrouvons aussi chez *G. candidum* des endoglucanases, enzymes dégradant la cellulose tendant vers cette hypothèse. La Spermine/spermidine synthase joue un rôle dans le métabolisme de composés soufrés, la production de composés soufrés volatils de *Geotrichum candidum* dans le fromage est source d'arôme. Les autres gènes sont en cours d'analyse en relation avec leurs rôles possibles dans l'adaptation de *G. candidum* lors de la fabrication du fromage ou dans d'autres niches écologiques.

3.9 Signe sexuel et locus MAT chez *G. candidum*.

Nous avons vu que la levure *G. candidum* est dans le clade des *Dipodascaceae*. Comme nous l'avons montré dans le chapitre 2, l'absence de cassette et de HO fait qu'il existe dans la population de *G. candidum* deux signes sexuels distincts, appelés arbitrairement dans cette étude MATA et MATB. Comme *Y. lipolytica*, *G. candidum* est donc hétérothallique.

La souche séquencée ici présente le signe sexuel MATA. Signe déterminé par la recherche d'orthologue BLASTP avec la protéine codée par le gène GECA02s09283g. Fait intéressant ce gène présente une homologie de séquence avec un domaine appelé « *HMG (high mobility group) box binding protein* ». Ce domaine est retrouvé dans l'ensemble des gènes *MAT* de champignon filamenteux et chez les hémiascomycètes : *Y. lipolytica, K. lactis, C. albicans*. L'alignement de ce domaine avec les autres gènes *MAT* d'origine fongique est présenté dans la

Figure 39**.** Le gène opposé, *MATB* a aussi pu être séquencé et possède quant à lui un domaine dit « *alpha box* » que l'on retrouve chez tous les ascomycètes (voir chapitre II).

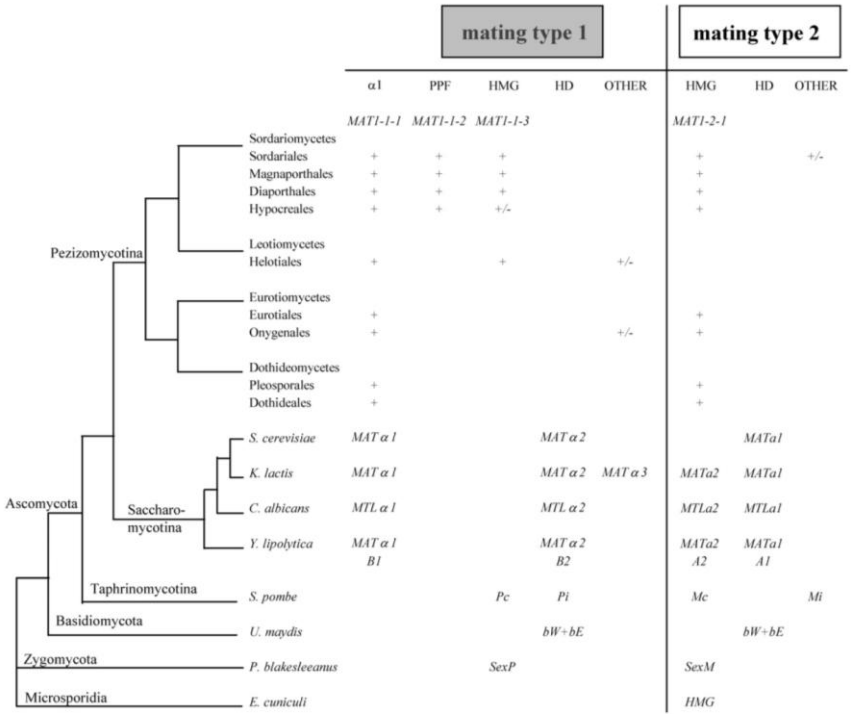


Figure 39 : Structure des Mating types dans l'ensemble du règne fungique (Martin et *al.*, 2010)

Les 2 domaines conservés, ont pu être alignés avec les séquences protéiques retrouvées dans la base de donnés NCBI (**Figure 40**).Malheureusement, ces séquences ne nous permettent pas d'obtenir une phylogénie robuste. Contrairement aux autres levures *G. candidum* possède un seul gène dans le

locus *MATA* et *B*. C'est une configuration que l'on peut retrouver chez d'autres champignons filamenteux (comme par exemple les espèces du genre *Aspergillus*). Une nouvelle fois, une caractéristique de champignon filamenteux est retrouvée chez *G. candidum*.

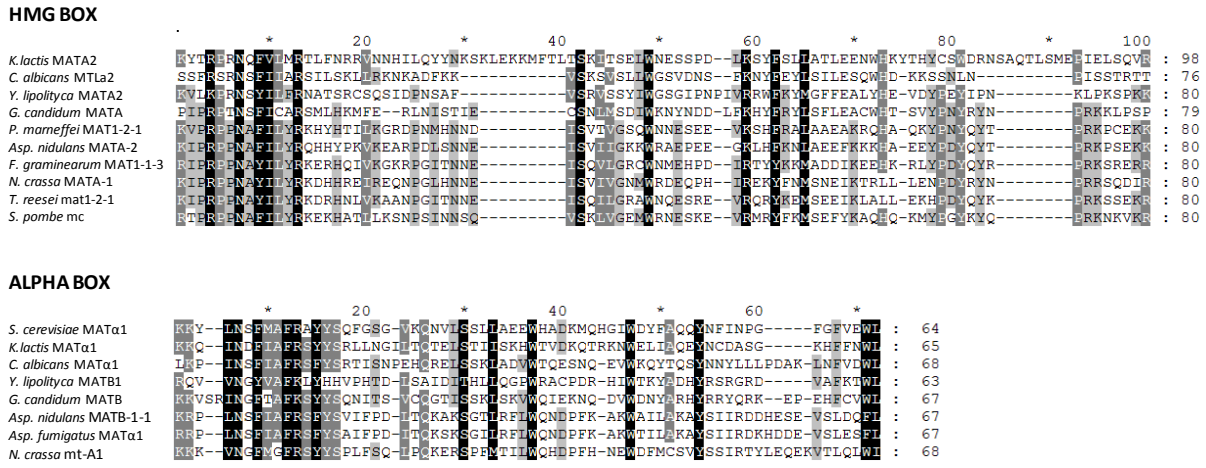


Figure 40 : Alignement des domaines HMG box et alpha box chez les champignons

L'accès au gène du mating type de *Geotrichum candidum* nous fournit un nouvel outil pour étudier l'évolution du locus MAT chez les levures avec un nouveau candidat de la famille de Dipodascacae. La **Figure 41** présente l'organisation du locus MAT dans *Geotrichum candidum* et la compare avec l'organisation remarquée dans Butler et *al.* (2004). Pour chaque espèce, sont représentés l'organisation des 2 idiomorphes du gène du type sexuel. Les lignes bleues connectent les gènes orthologues. Les gènes conservés sont colorés: en rouge, l'idiomorphe a ; en vert, α ; en orange, les homologues de *S. cerevisiae APC5* (YOR249C) ; en brun, les homologues de *S. cerevisiae COX13* (YGL191W); en violet, les homologues de *S.* cerevisiae *SLA2* (YNL243W) ; en violet clair, *SUI1* (YNL244C) ; en rose, DIC1 (YLR348C) ; en jaune, *APN2* (YBL019W) ; en gris, *eiF1* () et en noir les gènes espèces spécifiques. Ainsi si en position 3' du locus la situation parait assez claire, on retrouve chez *G. candidum* les gènes, homologues à *S. cerevisiae SLA2*, *SUI1* et *eIF1.* La synthénie des gènes SLA2 et SUI1 semble largement conservée chez toutes les levures indiquées ici, mis à part chez Oogatea polymorpha présentant alors le gène *eIF1* en place de *SUI1*. Un gène homologue à *eIF1* est de plus retrouvé chez *G. candidum* en position 3' de *SLA2* et *SUI1*. En position 5' du locus le cas de *G. candidum* parait cependant moins clair. Contrairement à ce que l'on trouve chez Y. lipolytica et chez les champignons filamenteux tel que *T. reesei* ou *N. crassa*, le gène homologue à APN2 de S. cerevisiae n'est par retrouvé chez *G. candidum*. On retrouve cependant le gène *APC5* gène qui se

trouve plus en amont du locus chez les champignons. Cette disposition de gènes nous montre certainement le résultat d'un remaniement chromosomique aux abords du locus.

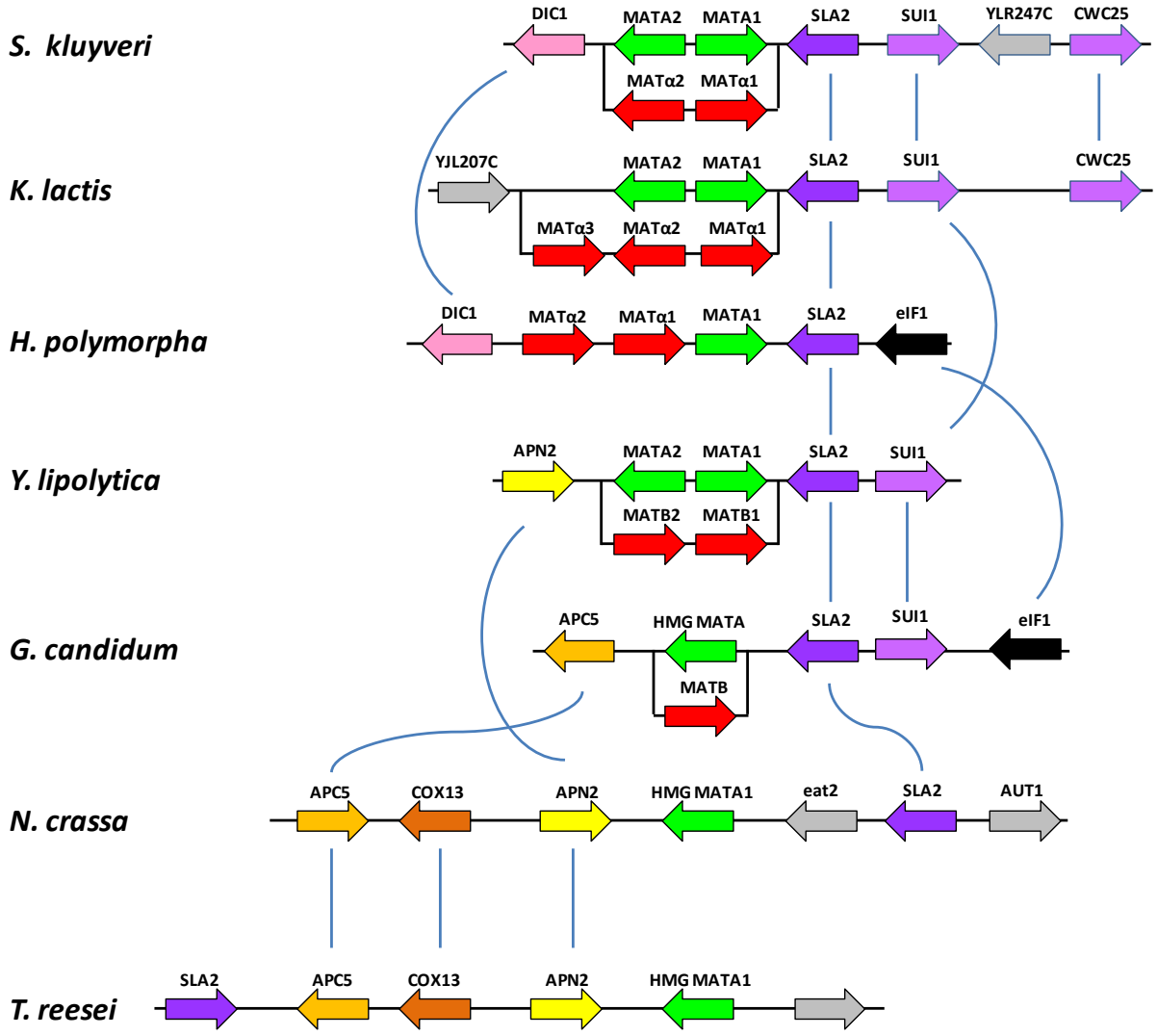


Figure 41 : Organisation comparative du locus MAT chez cinq espèces de levures et de deux champignons filamenteux, *N. crassa* et *T. reesei*.

4 CONCLUSIONS

L'analyse globale du génome indique que de nombreuses caractéristiques de champignons filamenteux sont présentes chez *G. candidum*. Le génome de *G. candidum* est d'une une grande taille, 24,8 Mb. Le profil d'intron est peu différencié et ceux-ci sont de petite taille avec une taille moyenne de 136 pb. *G. candidum* possède une grande variété de transposons proche de ceux

trouvés chez les champignons. Enfin la structure du locus du Mating type possède un seul gène comme certains champignons filamenteux.

Enfin, au moins 315 gènes sont retrouvés chez *G. candidum* et non dans les autres levures séquencés. Cela ouvre une nouvelle perspective dans la compréhension de l'évolution des génomes de levures et nous pouvons avancer une nouvelle théorie selon laquelle *G. candidum* serait une levure « intermédiaire » entre champignons filamenteux et levures hémiascomycètes.

CONCLUSIONS ET PERSPECTIVES

Lors de ce travail, nous avons essayé d'améliorer la vision globale que l'on avait de *G. candidum* et de positionner cette espèce dans l'arbre des champignons. Trois études complémentaires ont été effectuées : (i) une étude visant à délimiter *G. candidum* et les genres qui lui sont associés; (ii) une étude visant à mieux comprendre l'évolution de *G. candidum* tout en fournissant également des outils pour étudier les souches industrielles utilisées dans la fabrication du fromage ; et (iii) une étude du génome de *G. candidum* permettant de mieux comprendre son histoire évolutive et ses propriétés technologiques.

Au cours de cette partie conclusions et perspectives, les trois études menées au cours de cette thèse seront analysées indépendamment l'une de l'autre en suivant un schéma commun. Ainsi pour chaque étude, je ferai tout d'abord un rappel des objectifs à atteindre en détaillant les différentes étapes pour y parvenir. Enfin, je discuterai des apports fondamentaux des résultats de ces trois études et je développerai les projets envisageables pour la poursuite de ce travail.

Délimitation de l'espèce *Geotrichum candidum/Galactomyces candidus* et phylogénie des espèces du genre Geotrichum et des genres associés

Afin, d'étudier la diversité de l'espèce *G. candidum* dans les fromages et de choisir un candidat pour le séquençage, la première tache que j'ai accompli a été d'identifier les différentes souches du CIRM levures et les souches d'origine industrielle. Cela a alors posé la question de qu'est ce que *G. candidum*, question à laquelle je tâche de répondre dans cette première étude.

En effet, l'histoire taxonomique de la levure *Geotrichum candidum* est longue et complexe (de Hoog et Smith, 2004). Découverte en 1809, cette levure a très souvent changé de nom, on compte 38 synonymes dans le livre de référence The yeasts, Kurtzman et *al.* (2011). Considérée longtemps comme un champignon filamenteux avec des caractéristiques de levures, ce n'est que très récemment qu'elle a été intégrée au Saccharomycotina (levures hémiascomycètes). Afin de s'assurer de la nature taxonomique du matériel de départ, les 51 souches appelées *G. candidum* et fournies par les participants au projet *Food Microbiomes*, ainsi que 52 souches du CIRM-Levures ont d'abord été testées pour leur appartenance à l'espèce *G. candidum* par des méthodes moléculaires et aussi pour évaluer la variabilité intra-spécifique intrinsèque de la population. Le marqueur moléculaire utilisé est celui de la séquence codante du gène *ACT1* (Daniel et Meyer, 2003). Un total de 650 pb de 77 séquences a été comparé aux deux souches types de l'espèce. Très peu de diversité intra-spécifique a été observée. C'est un résultat surprenant qui est en désaccord avec ce que l'on pensait sur la variabilité intra-spécifique chez cette espèce décrite par *Chromosome Length Polymorphisms* (CLP) sur la taille des chromosomes (Gente et *al.*, 2002a), les expériences de *Random Amplified*

Polymorphic DNA (RAPD) (Gente et *al.*, 2002b) et la variabilité des séquences ribosomiques (Alper et *al.*, 2011). Nous avons montré que les méthodes citées ne constituent pas une bonne façon d'évaluer la diversité spécifique chez *G. candidum*. Notre approche, basée sur l'analyse des séquences des gènes codants, a été confirmée par les analyses MLST de 55 souches (voir plus bas).

Nous avons néanmoins identifiées 2 souches fournies par les industrielles qui divergeaient de *G. candidum*. Elles ont été identifiées entant que *G. fragrans*. Finalement, nous n'avons pas rencontré de problèmes pour identifier *G. candidum* et elle est la seule espèce du genre *Geotrichum* utilisée dans les fromages.

Les phylogénies existantes de *Geotrichum candidum* et des espèces proches *Galactomyces sp.*, *Dipodascus sp.*, *Magnusiomyces sp.* et *Saprochaete sp.* ont, jusqu'à présent, été basées sur la comparaison de séquences de l'ADN ribosomique (SSU et ITS). La validité de ces marqueurs pour les études de phylogénie est aujourd'hui contestée. Il vient d'ailleurs d'être montré que les séquences ribosomique présentaient une grande variabilité intra-spécifique (Alper et *al.*, 2011). Une analyse multigénique a été entreprise en utilisant la concaténation de gènes de ménage présents en simple copie dans le génome et codants pour des protéines : l'exon2 du gène codant pour l'actine *ACT1*, les gènes codant pour des sous-unités de l'ARN polymérase *RPB1* et *RPB2* et TEF1α, gène codant pour le facteur d'*élongation alpha.*

Nous avons donc dans un premier temps confirmé que *G. candidum* appartenait bien au Saccharomycotina et qu'elle a une position basale dans l'arbre des Saccharomycotina avec pour plus proche parent *Yarrowia lipolytica* (Dujon et *al.*, 2004).

De plus, l'étude a aussi pu mettre en évidence les dangers intrinsèques du à une taxonomie très changeante et non logique. En effet les anamorphes des espèces du genre *Geotrichum* ne sont pas des *Candidum* mais des *Geotrichum*. Ainsi, *G. candidum* ou *Gal. candidum* est très fréquemment confondu avec *Gal. geotrichum*. Cela a eu pour effet la description des deux nouvelles espèces invalidées depuis par les travaux de Groenewald et *al.* (2012) : *G. bryndzae* et *G. silvicola*.

Nous proposons ici, afin d'éviter toutes nouvelles erreurs de réinstaurer le nom d'espèce *G. candidum* En effet, cette espèce est très utilisé dans l'industrie biotechnologique et agro-alimentaire sous la dénomination *Geotrichum candidum*. De plus, le genre *Galactomyces* a été décrit comme noms de genre d'espèces téléomorphes des espèces *Dipodascus* et *Geotrichum* (de Hoog et Smith, 2004).

Le 18^{e} congrès international de botanique, qui s'est tenu en Juillet 2011, a adopté un amendement à l'article 59 du Code international de nomenclature botanique qui abandonne la double nomenclature traditionnellement utilisé en mycologie pour désigner les formes sexuées et asexuées d'une espèce

fongique (Norvell, 2011). La règle de « Un nom, un champignon » s'impose et un nom doit être choisi pour le genre *Geotrichum* / *Galactomyces*. Le choix du genre décrit plus récent est recommandé par le code, mais dans le cas de *Geotrichum* / *Galactomyces* les diverses raisons présentées ci-dessus plaident fortement en faveur de *Geotrichum.*

L'accès récent au génome de *G. candidum* CLIB 918 (voir plus bas) pourra à l'avenir, conduire au développement de nouveaux marqueurs nécessaires pour obtenir une phylogénie comprenant l'ensemble des espèces associés aux clades *Geotrichum* et *Magnusiomyces* étant donné que celles-ci ne vont pas être séquencées, par exemple *MCM7*, *NIP1*, *TSR1*, *CDC60* ou *UBA1* (Aguileta et *al.*, 2008).

Mis en place d'un schéma MLST pour le typage de *Geotrichum candidum*

L'objectif était ici de répondre à la nécessité d'une meilleure compréhension de l'évolution de *G. candidum* tout en fournissant également des outils pour étudier les souches industrielles utilisées dans la fabrication du fromage. La disponibilité du génome de *G. candidum* nous a fait choisir la méthode de typage la plus reproductible qui est la MLST. Cette approche devait aussi nous permettre de mieux connaître la diversité intra-spécifique de l'espèce *Geotrichum candidum.* En s'inspirant de schémas MLST mis en place pour d'autres levures, un schéma MLST a été établi avec l'analyse d'environ 500pb de 5 gènes de ménage, *NUP116, SAPT2, SAPT4, URA1, URA3* sur 55 souches de *G. candidum* d'origine variée. La première difficulté rencontrée ici, a été de trouver des loci suffisamment divergeant pour mettre en place le schéma MLST. En effet l'étude et le séquençage de 14 gènes supplémentaire s'est montré infructueux. Ce qui nous montre encore une fois que la diversité génétique de G. candidum a été quelque peu surestimée. Cette étude MLST a permis de mettre en évidence 30 séquences types distinctes et de détecter trois groupes. Le premier est composé de 11 souches clairement distinctes (11 ST) d'origine diverse (sable, yaourt, silo à grain, crème, fromage…) ; le second est composé d'un plus grand nombre de souches, 40 au total, toutes isolées de produits laitiers, auxquelles s'additionnent deux souches isolées de fèces, mais dont l'origine est vraisemblablement un produit laitier, puisqu'elles ne sont pas distinguables des souches isolées de fromage.

Cette analyse indique aussi que la divergence génétique qui existe entre ces souches est faible. Cette faible diversité observée pourrait être due à une spécialisation récente de *G. candidum* dans l'écosystème fromager, son mode de propagation ou l'utilisation systématique d'un petit groupe de souches technologiquement performantes. Chez certains champignons, il existe très peu d'échanges génétiques. C'est pourquoi nous avons cherché à connaitre le mode de propagation de cette espèce du point de vue moléculaire. A partir de la séquence génomique de *G. candidum*, il a été possible

d'identifier le locus sexuel et un gène qui présentait une similarité de séquence avec des gènes *MAT* des champignons. Nous avons ainsi identifié l'un des gènes du signe sexuel que nous avons appelé *MATA* chez la souche séquencée. L'autre gène, *MATB*, a été mis en évidence lors de l'analyse de la même région dans plusieurs souches. L'analyse du génome nous permet d'affirmer que *G. candidum* est hétérothallique. Le signe sexuel des 55 souches de *G. candidum* a ainsi pu être déterminé et le gène *MATB* a été séquencé. En détectant la présence des deux types sexuels dans les souches haploïdes *G. candidum*, nous avons pu analyser pour la première fois l'ampleur des échanges génétiques dans *G. candidum*. La répartition assez uniforme des types sexuels dans la population étudiée est cohérente avec l'absence de clonalité, même dans la sous-population issue du fromage. Cela impliquerait que les échanges de matériel génétique ont même lieu dans le fromage. En accord avec cela, il a déjà été observé que les souches du complexe *Debaryomyces hansenii* isolé du fromage étaient principalement des diploïdes ou hybrides (Jacques et *al.*, 2009; Jacques et *al.*, 2010), ce qui laisse penser que l'environnement peut influer sur la formation de diploïdes et hybrides. La présence de diploïdes chez G. candidum a été rarement observée, nous n'en avons trouvé que sept parmi les 62 souches. Ainsi l'absence de clonalité que nous avons observé dans G. candidum indique plutôt que les événements de recombinaison sont fréquents et que la méiose est un processus efficace de *G. candidum*.

Le deuxième résultat majeur de l'analyse MLST est le regroupement quasi de toutes les souches de fromage dans un clade dans un nombre limité de ST, tandis que l'autre clade contient un mélange de souches environnementales et du fromage tous distincts les uns des autres. Les souches restantes sont sur de longues branches entre clade I et II clade. La position intermédiaire des souches pouvant être le résultat de l'accouplement entre le clade I et clade II. Un tel placement intermédiaire de souches entre les principaux clades industriels / spécialisée n'est pas sans rappeler certaines souches de *S. cerevisiae* comme le décrit Liti et *al.* (2009) dans leur étude de génomique des populations. De même, les souches du clade II ne sont pas sans rappeler les configurations retrouvées chez les levures de vin européennes ou les levures de saké.

Deux hypothèses peuvent être avancées pour l'origine des souches du clade II: l'une serait une adaptation forte à l'environnement fromage et la fabrication du fromage liée aux activités de l'homme, l'autre serait la sélection naturelle de souches ayant la capacité de faire face à un tel environnement. A ce stade, nous ne pouvons pas choisir entre ces deux possibilités.

Étonnamment, le regroupement des souches de *G. candidum* en deux populations distinctes n'est pas sans rappeler la façon dont les souches de *S. cerevisiae* se groupe en fonction de leur origine ou de leur implication dans un processus spécifique (Legras et *al.,* 2007; Liti et *al.*, 2009; Schacherer et *al.*, 2009). De plus, les souches de *S. cerevisiae* domestiqué par l'homme, et en particulier les levures

de vin, ont montré peu de diversité génétique contrairement aux souches sauvages de *S. paradoxus* (Liti et *al.*, 2006; Liti et *al.*, 2009). La divergence de séquence globale pour les souches de *G. candidum* issues du clade « fromage » (0,3%) est encore inférieure à celles déterminé pour la levure *S. cerevisiae* vin (1,2%). Cela reste en accord l'hypothèse d'un processus évolutif similaire chez deux espèces industrielles. Il est probable que les souches de *S. cerevisiae* issues du vin ont la même origine (Legras et *al.*, 2007). L'évolution de *G. candidum* peut avoir lieu en utilisant les mêmes mécanismes, dont certains sont liés aux activités humaines.

La recherche de variabilité intra-spécifique étant longue et couteuse, nous avons mis au point une nouvelle méthode de typage, non pas pour remplacer la MLST, mais plutôt en soutien à cette dernière. Cette méthode est basée sur la variabilité des insertions de séquences répétées, les Long Terminal Repeats (LTR) qui bordent les transposons *Ty*. Auparavant développée chez K. marxianus et D. hansenii (Sohier et *al.*, 2009) ou S. cerevisiae (Legras et Karst, 2003), cette méthode permet l'amplification PCR des séquences séparant les LTR, conduisant à un profil unique par souche. L'avantage de cette méthode est qu'elle permet de suivre les variations récentes dans le génome au contraire de la MLST qui dépend de la fixation effective de mutations. En conséquence, cette méthode a pu différencier des souches qui ne l'étaient pas par MLST. Rapide et peu couteuse, on peut envisager utiliser cette méthode en routine pour le suivi de la dynamique des populations de levures dans le fromage en cours d'affinage. La variabilité obtenue par analyse des profils de PCR inter LTR est bien plus importantes que celle obtenue par MLST. Elle rejoint donc les précédentes analyses effectuée par RAPD ou CLP (Gente et *al.*, 2002a; Gente et *al.*, 2002b). L'explication la plus plausible est que les deux divergences observées ne sont pas nécessairement liées. Ainsi, si l'on étudie via la MLST la divergence génétique via la fixation de mutation, les trois autres méthodes mettent l'accent sur une divergence de structure au sein même du génome de *G. candidum*. Cette divergence peut être du à des événements de recombinaisons liés à la sexualité fréquents chez *G. candidum*, mais comme nous le verrons plus tard, cette divergence structurale peut être corrélée avec le nombre importants de transposons détectés dans le génome de cette levure. En effet comme le montre différentes études (Casaregola et *al.*, 1998; Rachidi et *al.*, 1999; Zolan, 1995), les transposons peuvent avoir un rôle dans la plasticité des génomes.

Ces travaux vont être poursuivis avec l'ajout de deux nouveaux marqueurs. On peut déjà observer que certaines souches semblent se regrouper selon leur région d'origine, ce qui serait in argument positif pour une spécificité de terroir et/ou de production. C'est en particulier le cas pour les souches isolées de Haute-Savoie qui ne sont jamais associées avec les souches d'Auvergne ou de Normandie.

Il pourrait être à l'avenir intéressant à l'avenir de chercher des gènes correspondant à une sélection purifiante ou une sélective positive plus rapide. Les traces de sélection purifiante ($dN/dS << 1$)

pourront être recherchées pour pointer les gènes qui, s'ils déterminent l'association de la levure avec le fromage, sont soumis à une forte sélection. Et si certains gènes peuvent être des possibles gains de fonction par mutation, signalés par une sélection diversifiante (dN/dS > 1).

Enfin, il apparait qu'un typage par l'étude du polymorphisme des microsatellites pourra aussi être envisagé. En effet bien que plus laboratoire « dépendant » cette méthode peut être préconisé pour l'étude de la variabilité d'espèces récentes (Haasl et Payseur, 2011).

Obtention et analyse du génome de *Geotrichum candidum* CLIB 918

Afin d'augmenter les données pour les analyses méta-génomique et d'obtenir le génome de l'une des rares levures majeures des fromages à ne pas avoir été séquencée, nous avons décidé de séquencer le génome de la levure *G. candidum*. Un premier travail expérimental et de bibliographie a consisté à choisir une souche représentative de l'espèce. Notre choix s'est porté sur la souche FM74 (déposée au CIRM-Levures sous le nom CLIB 918 = ATCC 204307), car c'est une souche française, haploïde, isolée de fromage (Pont l'Evêque) et qui a fait l'objet de nombreuses études à la fois technologiques, physiologiques et moléculaires.

L'étude globale du génome de *G. candidum* a révélé que cette espèce pourrait être considérée comme une levure intermédiaire. En effet de nombreux éléments rapprochent *G. candidum* des champignons.

- **La taille du génome** de *G. candidum* est de 24,8 Mb. Ce qui fait du génome de *G. candidum*, le plus grand génome de levure haploïde séquencée à ce jour. L'annotation automatique Eugène s'est montrée très performante pour prédire les gènes de *Geotrichum candidum*. L'obtention et le bon *mapping* des séquences d'ARN totaux est une véritable chance pour affiner cette annotation. Un total de 6802 gènes codant pour des protéines ont été annoté par Eugène.32% de ces gènes sont composé d'au moins un intron, ils sont 14,9% chez *Y. lipolytica* et 5% chez *S. cerevisiae*. Enfin, 38,2% de la séquence génomique est constituée de séquence codante.
- ***G. candidum* possède un grand nombre de gènes introniques** (32%) avec une séquence consensus des sites d'épissages d'introns en 5' peu conservé et une taille moyenne de 138 pb
- **Une grande variété de transposons** a été détectée par une analyse des séquences par le logiciel REPET. Nous ne retrouvons de transposons avec en particulier des transposons uniquement trouvés jusqu'à maintenant chez les champignons filamenteux et pas chez les levures, les MITEs.
- **315 gènes sans orthologues chez les levures hémiascomycètes** ont été trouvé par une analyse par comparaison de similarité de séquence aux diverses organismes séquencés. Ces gènes font

partie d'un ensemble de 583 gènes annoté automatiquement par rapport aux champignons filamenteux. Parmi eux certains sont à la fois dans les deux taxons. Dans ce cas, il existe une famille de gènes de type levures et une famille de gènes de type champignons filamenteux ; les gènes de *G. candidum* d'intérêt sont soit dans l'une des familles, celle des champignons filamenteux, soit dans les deux familles. L'analyse des introns présent dans les gènes de champignon conservé chez *G. candidum* a permis de montré que la perte d'introns s'est effectué assez tôt après la différenciation avec les champignons

- **Une structure du locus du mating-type** proche de celle des champignons avec seulement 1 seul gène présentant une région conservée HMG-box ou alpha-box dans chaque mating type.

L'ensemble des gènes de champignons filamenteux identifiés sont en train d'être testés par analyse phylogénétique. Une collaboration a été mise en place avec l'équipe de Toni Gabaldon (***CRG's Bioinformatics and Genomics,*** Barcelone), afin de systématiser l'analyse phylogénétique des gènes détectés. L'analyse devra fournir un arbre phylogénétique pour chaque gène et nous permettra de statuer pour chacun d'entre eux entre l'hypothèse d'un transfert horizontal de gène ou une conservation de gène d'origine fongique chez *G. candidum*. A l'avenir, les données de séquençage devenant de plus en plus nombreuses y compris chez les champignons filamenteux, il est certain que l'existence de nouveaux transferts de gène eucaryote à eucaryote seront à nouveaux décrits.

L'ensemble des gènes de champignons filamenteux chez *G. candidum* est principalement impliqué dans le métabolisme, mais on y retrouve également certains gènes de ménage, ce qui pourrait suggérer que ces gènes pourraient être issus d'un événement de transfert horizontal massif, lors d'une hybridation par exemple. En effet, les gènes de ménage ne sont pas sélectionnés lors de transferts horizontaux chez les eucaryotes. Une autre possibilité est que ceux-ci aient été conservés pour assurer l'expression des gènes du métabolisme conférant un réel avantage dans l'adaptation de *G. candidum* dans son milieu.

Les analyses phylogénétiques indiquent également qu'une autre catégorie de gènes, clairement issus de transferts horizontaux eux aussi, ont des origines différentes, comme une spermine/spermidine synthase dont la source la plus probable se trouve chez les basidiomycètes, alors que les autres transferts horizontaux semblent provenir de champignons Euascomycètes. La présence de ce gène chez d'autres espèces du genre *Geotrichum* est entrain d'être recherché. Une étude récente présente la phylogénie des gènes de spermine et spermidine synthase chez tous les organismes. L'arbre phylogénétique montre que les de spermines synthases trouvés chez les levures et les gènes de spermidine synthases trouvés chez tous les Ascomycètes sont paraphylétique. Il dérive alors d'un

ancêtre commun (Minguet et *al.*, 2008). L'absence de la spermine synthase chez les champignons filamenteux peut alors s'expliquer par la non-duplication ancestrale du gène chez ces derniers. Le gène ancestral serait alors une spermidine synthase présente chez tous les ascomycètes. En ce sens, chez *G. candidum* sont présent un gène de spermine synthase mais pas de spermidine synthase. La perte de spermidine synthase tend vers une seule histoire évolutive. L'hypothèse la plus vraisemblable est que ce dernier ait perdu le gène de spermidine synthase et que cette perte ait été compensée par l'apport d'une spermide/spermidine synthase d'origine fongique. A l'avenir, il pourrait être intéressant de caractériser la spermide/spermidine synthase de *G. candidum* en la clonant chez une souche de *S. cerevisiae ΔSPE3* et sur une souche *ΔSPE4.* La complémentation de *ΔSPE3* permettrait de conforter notre hypothèse quant au rôle de cette enzyme chez *G. candidum*. Si la souche *ΔSPE4* complémentée produit de la spermine, la double capacité de la spermine/spermidine synthase serait démontrée. Les différents résultats contradictoire quant à la production de spermine chez les champignons filamenteux (Marshall et *al.*, 1979; Nickerson et *al.*, 1977; Paulus et *al.*, 1982), pourrait être expliquée par le fait que cette dernière enzyme est soumise à des régulations du aux conditions de cultures.

La présence du gène de spermine/spermidine synthase chez les espèces proches de *G. candidum* est en train d'être testée au laboratoire. Cela permettra d'estimer la date du HGT. La présence de la spermidine synthase devra aussi être testée pour essayer de corréler la perte du gène et le gain par transfert horizontal. Existe-t-il des souches possédant trois enzymes impliqués dans la synthèse des polyamines, est ce spécifique à *G. candidum*, est ce que l'apport de ce gène est lié à des activités associé au métabolisme du soufre.

Nos travaux permettent de mieux comprendre la confusion qui existait sur la position taxonomique de cette levure qui est restée longtemps classée parmi les champignons, car elle présente du fait de l'origine d'un certain nombre de ses gènes, plusieurs caractéristiques de champignons filamenteux. Des efforts sont apportés à une meilleure compréhension de l'apport de ces gènes dans l'adaptation de *G. candidum* à l'environnement fromager. Au-delà de l'utilisation du génome dans les travaux de métagénomiques du projet, la mise à disposition de ce génome annoté pour la communauté scientifique constitue un apport substantiel de notre travail dans ce projet.

Le séquençage de *G. candidum* révèle une nouvelle perspective dans la compréhension de l'évolution des levures et leur adaptation à leur milieu. Le grand nombre de gènes de champignons conservés par *G. candidum* au cours de l'évolution, révèle peut être une adaptation en deux étapes jusqu'a aujourd'hui. *G. candidum* est retrouvé fréquemment dans un environnement végétal et la conservation de gènes comme les endoglucanases et le polygalacturonase révèle une adaptation à cet environnement. On peut alors très bien concevoir, que *G. candidum* ait colonisé les fromages

lorsque ceux-ci étaient encore posés sur des planches de bois. L'activité humaine n'aurait alors été que le révélateur d'une sélection naturelle de souches ayant la capacité de faire face à un tel environnement.

La séquence a été utilisée dans les projets ANR « Food Microbiomes » et « ExEco ». Il a été montré que *G. candidum* est la levure la plus transcriptionellement active dans un écosystème fromager artificiel contenant plusieurs bactéries et trois levures *Kluyveromyces lactis*, *Debaryomyces hansenii* et *Geotrichum candidum*.

Il reste bien entendu du travail a effectué sur le génome mais j'espère que les pistes que j'ai suivi dans ce manuscrit constitueront une bonne base pour mes successeurs. A l'avenir je pense qu'il serait intéressant de séquencer d'autres espèces de ce clade présentées dans le chapitre 1 des résultats. Le séquençage de novo d'une espèce du genre *Magnusiomyces* pourrait révéler de nouveaux transferts horizontaux ou de nouveaux gènes de champignons filamenteux conservés. Cela permettra de confirmer les transferts horizontaux décrits chez *G. candidum.*

L'accès à une telle séquence permettra de mieux comprendre le processus évolutif de *G. candidum*. De plus le séquençage d'une souche de *G. candidum* isolée de l'environnement pourrait nous renseigner sur les évolutions récentes de *G. candidum* dans l'écosystème fromager comme cela a été montré chez *S. cerevisiae* EC 11118.

BIBLIOGRAPHIE

Aguileta, G., Marthey, S., Chiapello, H., Lebrun, M.H., Rodolphe, F., Fournier, E., Gendrault-Jacquemard, A., et Giraud, T. (2008). Assessing the performance of single-copy genes for recovering robust phylogenies. Systematic biology *57*, 613-627.

Albertin, W., Marullo, P., Aigle, M., Bourgais, A., Bely, M., Dillmann, C., et Sicard, D. (2009). Evidence for autotetraploidy associated with reproductive isolation in *Saccharomyces cerevisiae*: towards a new domesticated species. Journal of evolutionary biology *22*, 2157-2170.

Alper, I., Frenette, M., et Labrie, S. (2011). Ribosomal DNA polymorphisms in the yeast *Geotrichum candidum*. Fungal biology *115*, 1259-1269.

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

Anderson, J.B., Wickens, C., Khan, M., Cowen, L.E., Federspiel, N., Jones, T., et Kohn, L.M. (2001). Infrequent genetic exchange and recombination in the mitochondrial genome of *Candida albicans*. Journal of bacteriology *183*, 865-872.

Andersson, J.O. (2005). Lateral gene transfer in eukaryotes. Cellular and molecular life sciences : CMLS *62*, 1182-1197.

Andrighetto, C., Psomas, E., Tzanetakis, N., Suzzi, G., et Lombardi, A. (2000). Randomly amplified polymorphic DNA (RAPD) PCR for the identification of yeasts isolated from dairy products. Letters in applied microbiology *30*, 5-9.

Arfi, K., Spinnler, H.E., Tache, R., et Bonnarme, P. (2002). Production of volatile compounds by cheese-ripening yeasts: requirement for a methanethiol donor for S-methyl thioacetate synthesis by *Kluyveromyces lactis*. Applied microbiology and biotechnology *58*, 503-510.

Ayoub, M.J., Legras, J.L., Saliba, R., et Gaillardin, C. (2006). Application of Multi Locus Sequence Typing to the analysis of the biodiversity of indigenous *Saccharomyces cerevisiae* wine yeasts from Lebanon. Journal of applied microbiology *100*, 699-711.

Bachmann, H.P., Bobst, C., Butikofer, U., Dalla Torre, M., Frolich-Wyder, M.T., et Furst, M. (2003). Sticky cheese smear and natural white mould. MILCHWISSENSCHAFT • MILK SCIENCE INTERNATIONAL *58*, 117-232.

Bailey, M.J., and Pessa, E. (1990). Strain and process for production of polygalacturonase. Enzyme and Microbial Technology *12*, 266-271.

Bain, J.M., Tavanti, A., Davidson, A.D., Jacobsen, M.D., Shaw, D., Gow, N.A., et Odds, F.C. (2007). Multilocus sequence typing of the pathogenic fungus *Aspergillus fumigatus*. Journal of clinical microbiology *45*, 1469-1477.

Baleiras Couto, M.M., Eijsma, B., Hofstra, H., Huis in't Veld, J.H., et van der Vossen, J.M. (1996). Evaluation of molecular typing techniques to assign genetic diversity among *Saccharomyces cerevisiae* strains. Applied and environmental microbiology *62*, 41-46.

Barth, M., Hankinson, T., Zhuang, H., et Breidt, F. (2009). Microbiological Spoilage of Fruits and Vegetables. Compendium of the Microbiological Spoilage of Foods and Beverages, 135-183.

Belloch, C., Perez-Torrado, R., Gonzalez, S.S., Perez-Ortin, J.E., Garcia-Martinez, J., Querol, A., et Barrio, E. (2009). Chimeric genomes of natural hybrids of *Saccharomyces cerevisiae* and Saccharomyces kudriavzevii. Applied and environmental microbiology *75*, 2534-2544.

Bergemann, M., Lespinet, O., M'Barek, S.B., Daboussi, M.J., et Dufresne, M. (2008). Genome-wide analysis of the *Fusarium oxysporum* mimp family of MITEs and mobilization of both native and de novo created mimps. Journal of molecular evolution *67*, 631-642.

Besancon, X., Smet, C., Chabalier, C., Rivemale, M., Reverbel, J.P., Ratomahenina, R., et Galzy, P. (1992). Study of surface yeast flora of Roquefort cheese. International journal of food microbiology *17*, 9-18.

Bleykasten-Grosshans, C., Jung, P.P., Fritsch, E.S., Potier, S., de Montigny, J., et Souciet, J.L. (2011). The *Ty*1 LTR-retrotransposon population in *Saccharomyces cerevisiae* genome: dynamics and sequence variations during mobility. FEMS yeast research *11*, 334-344.

Bleykasten-Grosshans, C., et Neuveglise, C. (2011). Transposable elements in yeasts. Comptes rendus biologies *334*, 679-686.

Bonnarme, P., Arfi, K., Dury, C., Helinck, S., Yvon, M., et Spinnler, H.E. (2001a). Sulfur compound production by *Geotrichum candidum* from L-methionine: importance of the transamination step. FEMS microbiology letters *205*, 247-252.

Bonnarme, P., Lapadatescu, C., Yvon, M., et Spinnler, H.E. (2001b). L-methionine degradation potentialities of cheese-ripening microorganisms. The Journal of dairy research *68*, 663-674.

Borneman, A.R., Desany, B.A., Riches, D., Affourtit, J.P., Forgan, A.H., Pretorius, I.S., Egholm, M., et Chambers, P.J. (2012). The genome sequence of the wine yeast VIN7 reveals an allotriploid hybrid genome with *Saccharomyces cerevisiae* and *Saccharomyces kudriavzevii* origins. FEMS yeast research *12*, 88-96.

Bougnoux, M.E., Aanensen, D.M., Morand, S., Theraud, M., Spratt, B.G., et d'Enfert, C. (2004). Multilocus sequence typing of *Candida albicans*: strategies, data exchange and applications. Infect Genet Evol *4*, 243-252.

Bougnoux, M.E., Morand, S., et d'Enfert, C. (2002). Usefulness of multilocus sequence typing for characterization of clinical isolates of *Candida albicans*. Journal of clinical microbiology *40*, 1290-1297.

Boutrou, R., et Gueguen, M. (2005). Interests in *Geotrichum candidum* for cheese technology. International journal of food microbiology *102*, 1-20.

Brabcova, J., Zarevucka, M., et Mackova, M. (2010). Differences in hydrolytic abilities of two crude lipases from *Geotrichum candidum* 4013. Yeast *27*, 1029-1038.

Brown, C.A., Murray, A.W., et Verstrepen, K.J. (2010). Rapid expansion and functional divergence of subtelomeric gene families in yeasts. Curr Biol *20*, 895-903.

Brown, W.R., Liti, G., Rosa, C., James, S., Roberts, I., Robert, V., Jolly, N., Tang, W., Baumann, P., Green, C., *et al.* (2011). A Geographically Diverse Collection of *Schizosaccharomyces pombe* Isolates Shows Limited Phenotypic Variation but Extensive Karyotypic Diversity. G3 (Bethesda) *1*, 615-626.

Butinar, L., Spencer-Martins, I., et Gunde-Cimerman, N. (2007). Yeasts in high Arctic glaciers: the discovery of a new habitat for eukaryotic microorganisms. Antonie van Leeuwenhoek *91*, 277-289.

Butler, E.E., et Petersen, L.J. (1970). Sexual reproduction on *Geotrichum candidum*. Science *169*, 481-482.

Butler, E.E., et Petersen, L.J. (1972). *Endomyces geotrichum* a perfect state of *Geotrichum candidum*. Mycologia *64*, 365-374.

Butler, G., Kenny, C., Fagan, A., Kurischko, C., Gaillardin, C., et Wolfe, K.H. (2004). Evolution of the MAT locus and its Ho endonuclease in yeast species. Proceedings of the National Academy of Sciences of the United States of America *101*, 1632-1637.

Carr, P.D., et Ollis, D.L. (2009). Alpha/beta hydrolase fold: an update. Protein and peptide letters *16*, 1137-1148.

Casaregola, S., Nguyen, H.V., Lapathitis, G., Kotyk, A., et Gaillardin, C. (2001). Analysis of the constitution of the beer yeast genome by PCR, sequencing and subtelomeric sequence hybridization. International journal of systematic and evolutionary microbiology *51*, 1607-1618.

Casaregola, S., Nguyen, H.V., Lepingle, A., Brignon, P., Gendre, F., et Gaillardin, C. (1998). A family of laboratory strains of *Saccharomyces cerevisiae* carry rearrangements involving chromosomes I and III. Yeast *14*, 551-564.

Cholet, O., Henaut, A., Casaregola, S., et Bonnarme, P. (2007). Gene expression and biochemical analysis of cheese-ripening yeasts: focus on catabolism of L-methionine, lactate, and lactose. Applied and environmental microbiology *73*, 2561-2570.

Cohen, N.E., Shen, R., et Carmel, L. (2012). The role of reverse transcriptase in intron gain and loss mechanisms. Molecular biology and evolution *29*, 179-186.

Corredor, M., Davila, A.M., Casaregola, S., et Gaillardin, C. (2003). Chromosomal polymorphism in the yeast species *Debaryomyces hansenii*. Antonie van Leeuwenhoek *84*, 81-88.

Corsetti, A., Rossi, J., et Gobbetti, M. (2001). Interactions between yeasts and bacteria in the smear surface-ripened cheeses. International journal of food microbiology *69*, 1-10.

Cosentino, S., Fadda, M.E., Deplano, M., Mulargia, A.F., et Palmas, F. (2001). Yeasts associated with Sardinian ewe's dairy products. International journal of food microbiology *69*, 53-58.

Daniel, H.M., et Meyer, W. (2003). Evaluation of ribosomal RNA and actin gene sequences for the identification of ascomycetous yeasts. International journal of food microbiology *86*, 61-78.

de Hoog, G.S., et Smith, M.T. (2004). Ribosomal gene phylogeny and species delimitation in *Geotrichum* and its teleomorphs. Studies in Mycology *50*, 489–516.

de Hoog, G.S., et Smith, M.T. (2011). Chapter 31 - *Galactomyces* Redhead & Malloch (1977). In The Yeasts (Fifth Edition) (London: Elsevier), pp. 413-420.

de Hoog, G.S., Smith, T., et Guého, E. (1986). A Revision of the Genus *Geotrichum* and Its Telemorphs (Centraalbureau voar Schimmelcultures).

Degroeve, S., Saeys, Y., De Baets, B., Rouze, P., et Van de Peer, Y. (2005). SpliceMachine: predicting splice sites from high-dimensional local context representations. Bioinformatics *21*, 1332-1338.

Demarigny, Y., Berger, C., Desmasures, N., Gueguen, M., et Spinnler, H.E. (2000). Flavour sulphides are produced from methionine by two different pathways by *Geotrichum candidum*. The Journal of dairy research *67*, 371-380.

Dequin, S., et Casaregola, S. (2011). The genomes of fermentative *Saccharomyces*. Comptes rendus biologies *334*, 687-693.

Desmasures, N., Bazin, F., et Guéguen, M. (1997). Microbiological composition of raw milk from selected farms in the Camembert region of Normandy. Journal of applied microbiology *83*, 53-58.

Desnos-Ollivier, M., Ragon, M., Robert, V., Raoux, D., Gantier, J.C., et Dromer, F. (2008). *Debaryomyces hansenii* (*Candida famata*), a rare human fungal pathogen often misidentified as *Pichia guilliermondii* (*Candida guilliermondii*). Journal of clinical microbiology *46*, 3237-3242.

Dieuleveux, V., Rarah Ratih Adjie Maheswari Chataud, J., et Gueguen, M. (1997). Inhibition of *Listeria monocytogenes* by *Geotrichum candidum*, Vol 15 (Chatenay-Malabry, FRANCE: Société; informations études et édition en nutrition et alimentation).

Dieuleveux, V., Van Der Pyl, D., Chataud, J., et Gueguen, M. (1998). Purification and characterization of anti-Listeria compounds produced by *Geotrichum candidum*. Applied and environmental microbiology *64*, 800-803.

Dodgson, A.R., Pujol, C., Denning, D.W., Soll, D.R., et Fox, A.J. (2003). Multilocus sequence typing of *Candida glabrata* reveals geographically enriched clades. Journal of clinical microbiology *41*, 5709-5717.

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., et *al.* (2004). Genome evolution in yeasts. Nature *430*, 35-44.

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

Eliskases-Lechner, F., et Ginzinger, W. (1995). The bacterial flora of surface-ripened cheeses with special regard to coryneforms. Le Lait (Print) *75*, 571-583.

Ewing, B., et Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome research *8*, 186-194.

Ewing, B., Hillier, L., Wendl, M.C., et Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome research *8*, 175-185.

Fabre, E., Muller, H., Therizols, P., Lafontaine, I., Dujon, B., et Fairhead, C. (2005). Comparative genomics in hemiascomycete yeasts: evolution of sex, silencing, and subtelomeres. Molecular biology and evolution *22*, 856-873.

Fadda, M.E., Cosentino, S., Deplano, M., et Palmas, F. (2001). Yeast populations in Sardinian feta cheese. International journal of food microbiology *69*, 153-156.

Fadda, M.E., Mossa, V., Pisano, M.B., Deplano, M., et Cosentino, S. (2004). Occurrence and characterization of yeasts isolated from artisanal Fiore Sardo cheese. International journal of food microbiology *95*, 51-59.

Fairhead, C., et Dujon, B. (2006). Structure of *Kluyveromyces lactis* subtelomeres: duplications and gene content. FEMS yeast research *6*, 428-441.

Feng, X., Yao, Z., Ren, D., Liao, W., et Wu, J. (2008). Genotype and mating type analysis of *Cryptococcus neoformans* and *Cryptococcus gattii* isolates from China that mainly originated from non-HIV-infected patients. FEMS yeast research *8*, 930-938.

Fickers, P., Marty, A., and Nicaud, J.M. (2011). The lipases from *Yarrowia lipolytica*: genetics, production, regulation, biochemical characterization and biotechnological applications. Biotechnology advances *29*, 632-644.

Fink, G.R. (1986). Translational control of transcription in eukaryotes. Cell *45*, 155-156.

Fitzpatrick, D.A. (2012). Horizontal gene transfer in fungi. FEMS microbiology letters *329*, 1-8.

Fitzpatrick, D.A., Logue, M.E., Stajich, J.E., et Butler, G. (2006). A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. BMC evolutionary biology *6*, 99.

Fleet, G.H. (1990). Yeasts in dairy products. The Journal of applied bacteriology *68*, 199-211.

Fleet, G.H. (1999). Microorganisms in food ecosystems. International journal of food microbiology *50*, 101-117.

Fleet, G.H. (2003). Yeast interactions and wine flavour. International journal of food microbiology *86*, 11-22.

Fleet, G.H. (2007). Yeasts in foods and beverages: impact on product quality and safety. Current opinion in biotechnology *18*, 170-175.

Fleetwood, D.J., Khan, A.K., Johnson, R.D., Young, C.A., Mittal, S., Wrenn, R.E., Hesse, U., Foster, S.J., Schardl, C.L., et Scott, B. (2011). Abundant degenerate miniature inverted-repeat transposable elements in genomes of epichloid fungal endophytes of grasses. Genome biology and evolution *3*, 1253-1264.

Flutre, T., Duprat, E., Feuillet, C., et Quesneville, H. (2011). Considering transposable element diversification in de novo annotation approaches. PloS one *6*, e16526.

Foissac, S., Gouzy, J., Rombauts, S., Mathe, C., Amselem, J., Sterck, L., de Peer, Y.V., Rouze, P., et Schiex, T. (2008). Genome Annotation in Plants and Fungi: EuGene as a Model Platform. Current Bioinformatics *3*, 87-97.

Foury, F., Roganti, T., Lecrenier, N., et Purnelle, B. (1998). The complete sequence of the mitochondrial genome of *Saccharomyces cerevisiae*. FEBS letters *440*, 325-331.

Fox, P.F.M., P. L. H. (2004). Cheese, chemistry, physics and microbiology. General aspects. Elsevier Academic Press.

Friedman, R., et Hughes, A.L. (2001). Gene duplication and the structure of eukaryotic genomes. Genome research *11*, 373-381.

Gaillardin, C., Neuveglise, C., Kerscher, S., et Nicaud, J.M. (2012). Mitochondrial genomes of yeasts of the *Yarrowia* clade. FEMS yeast research *12*, 317-331.

Gainvors, A., Nedjaoum, N., Gognies, S., Muzart, M., Nedjma, M., et Belarbi, A. (2006). Purification and characterization of acidic endo-polygalacturonase encoded by the PGL1-1 gene from *Saccharomyces cerevisiae*. FEMS microbiology letters *183*, 131-135.

Galagan, J.E., Calvo, S.E., Borkovich, K.A., Selker, E.U., Read, N.D., Jaffe, D., FitzHugh, W., Ma, L.J., Smirnov, S., Purcell, S., *et al.* (2003). The genome sequence of the filamentous fungus *Neurospora crassa*. Nature *422*, 859-868.

Galeote, V., Bigey, F., Beyne, E., Novo, M., Legras, J.L., Casaregola, S., et Dequin, S. (2011). Amplification of a *Zygosaccharomyces bailii* DNA segment in wine yeast genomes by extrachromosomal circular DNA formation. PloS one *6*, e17872.

Garcia-Guinea, J., Cardenes, V., Martinez, A.T., et Martinez, M.J. (2001). Fungal bioturbation paths in a compact disk. Die Naturwissenschaften *88*, 351-354.

Garcia-Hermoso, D., Cabaret, O., Lecellier, G., Desnos-Ollivier, M., Hoinard, D., Raoux, D., Costa, J.M., Dromer, F., et Bretagne, S. (2007). Comparison of microsatellite length polymorphism and multilocus sequence typing for DNA-Based typing of *Candida albicans*. Journal of clinical microbiology *45*, 3958-3963.

Gardini, F., Suzzi, G., Lombardi, A., Galgano, F., Crudele, M.A., Andrighetto, C., Schirone, M., et Tofalo, R. (2001). A survey of yeasts in traditional sausages of southern Italy. FEMS yeast research *1*, 161-167.

Gente, S., Desmasures, N., Jacopin, C., Plessis, G., Beliard, M., Panoff, J.M., et Gueguen, M. (2002a). Intra-species chromosome-length polymorphism in *Geotrichum candidum* revealed by pulsed field gel electrophoresis. International journal of food microbiology *76*, 127-134.

Gente, S., Desmasures, N., Panoff, J.M., et Gueguen, M. (2002b). Genetic diversity among *Geotrichum candidum* strains from various substrates studied using RAM and RAPD-PCR. Journal of applied microbiology *92*, 491-501.

Gente, S., Larpin, S., Cholet, O., Gueguen, M., Vernoux, J.P., et Desmasures, N. (2007). Development of primers for detecting dominant yeasts in smear-ripened cheeses. The Journal of dairy research *74*, 137-145.

Goerges, S., Mounier, J., Rea, M.C., Gelsomino, R., Heise, V., Beduhn, R., Cogan, T.M., Vancanneyt, M., et Scherer, S. (2008). Commercial ripening starter microorganisms inoculated into cheese milk do not successfully establish themselves in the resident microbial ripening consortia of a South german red smear cheese. Applied and environmental microbiology *74*, 2210-2217.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M.*, et al.* (1996). Life with 6000 genes. Science *274*, 546, 563-547.

Gordon, D., Abajian, C., et Green, P. (1998). Consed: a graphical tool for sequence finishing. Genome research *8*, 195-202.

Goto-Yamamoto, N., Kitano, K., Shiki, K., Yoshida, Y., Suzuki, T., Iwata, T., Yamane, Y., et Hara, S. (1998). SSU1-R, a sulfite resistance gene of wine yeast, is an allele of SSU1 with a different upstream sequence. Journal of Fermentation and Bioengineering *86*, 427-433.

Gouy, M., Guindon, S., et Gascuel, O. (2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Molecular biology and evolution *27*, 221-224.

Groenewald, M., Coutinho, T., Smith, M.T., et van der Walt, J.P. (2012). Species reassignment of *Geotrichum bryndzae*, *Geotrichum phurueaensis*, *Geotrichum silvicola* and *Geotrichum vulgare* based on phylogenetic analyses and mating compatibility. International journal of systematic and evolutionary microbiology.

Grossetete, S., Labedan, B., et Lespinet, O. (2010). FUNGIpath: a tool to assess fungal metabolic pathways predicted by orthology. BMC genomics 11, 81.

Gueguen, M., et Jacquet, J. (1982). Etudes sur les caractères culturaux et la morphologie de Geotrichum candidum Link. Lait 62, 625-644.

Gueguen, M.S., M. (1992). Les levures et *Geotrichum candidum*. CEPIL,, pp. 165-219.

Guessous, Z., Ouhssine, M., Mokhtari, A., Faid, M., et El Yachioui, M. (2000). Isolement et caracterisation de *Geotrichum candidum* pour la production d'une polygalacturonase extracellulaire, Vol 20 (Paris, FRANCE: Lavoisier).

Guindon, S., et Gascuel, O. (2003). A simple, fast, et accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic biology *52*, 696-704.

Gyanchandani, A., Khan, Z.K., Farooqui, N., Goswami, M., et Ranade, S.A. (1998). RAPD analysis of *Candida albicans* strains recovered from different immunocompromised patients (ICP) reveals an apparently non-random infectivity of the strains. Biochem Mol Biol Int *44*, 19-27.

Haasl, R.J., et Payseur, B.A. (2011). Multi-locus inference of population structure: a comparison between single nucleotide polymorphisms and microsatellites. Heredity *106*, 158-171.

Hall, C., Brachat, S., et Dietrich, F.S. (2005). Contribution of horizontal gene transfer to the evolution of *Saccharomyces cerevisiae*. Eukaryotic cell *4*, 1102-1115.

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

Hamasaki-Katagiri, N., Katagiri, Y., Tabor, C.W., et Tabor, H. (1998). Spermine is not essential for growth of *Saccharomyces cerevisiae*: identification of the SPE4 gene (spermine synthase) and characterization of a spe4 deletion mutant. Gene *210*, 195-201.

Hansen, T.K., et Jakobsen, M. (2001). Taxonomical and technological characteristics of *Saccharomyces spp.* associated with blue veined cheese. International journal of food microbiology *69*, 59-68.

Hauser, N.C., Fellenberg, K., Gil, R., Bastuck, S., Hoheisel, J.D., et Perez-Ortin, J.E. (2001). Whole genome analysis of a wine yeast strain. Comparative and functional genomics *2*, 69-79.

Hayaloglu, A.A., et Kirbag, S. (2007). Microbial quality and presence of moulds in Kuflu cheese. International journal of food microbiology *115*, 376-380.

Hermet, A., Meheust, D., Mounier, J., Barbier, G., et Jany, J.L. (2012). Molecular systematics in the genus *Mucor* with special regards to species encountered in cheese. Fungal biology *116*, 692-705.

Hughes, A.L. (1994). The evolution of functionally novel proteins after gene duplication. Proceedings Biological sciences / The Royal Society *256*, 119-124.

Illková, K., Zemková, Z., Flodrová, D., Jäger, J., Benkovská, D., Omelková, J., Vadkertiová, R., Bobáľová, J., et Stratilová, E. (2012). Production of *Geotrichum candidum* polygalacturonases via solid state fermentation on grape pomace. Chem Pap *66*, 852-860.

Jacobsen, M.D., Gow, N.A., Maiden, M.C., Shaw, D.J., et Odds, F.C. (2007). Strain typing and determination of population structure of *Candida krusei* by multilocus sequence typing. Journal of clinical microbiology *45*, 317-323.

Jacques, N., Mallet, S., et Casaregola, S. (2009). Delimitation of the species of the *Debaryomyces hansenii* complex by intron sequence analysis. International journal of systematic and evolutionary microbiology *59*, 1242-1251.

Jacques, N., Sacerdot, C., Derkaoui, M., Dujon, B., Ozier-Kalogeropoulos, O., et Casaregola, S. (2010). Population polymorphism of nuclear mitochondrial DNA insertions reveals widespread diploidy associated with loss of heterozygosity in *Debaryomyces hansenii*. Eukaryotic cell *9*, 449-459.

James, S.A., Carvajal, B.E., Portero Barahona, P., Cross, K., Bond, C.J., et Roberts, I.N. (2012). *Candida ecuadorensis sp. nov.*, a novel ascomycetous yeast species found in two separate regions of Ecuador. International journal of systematic and evolutionary microbiology.

James, T.Y., Kauff, F., Schoch, C.L., Matheny, P.B., Hofstetter, V., Cox, C.J., Celio, G., Gueidan, C., Fraker, E., Miadlikowska, J.*, et al.* (2006). Reconstructing the early evolution of Fungi using a six-gene phylogeny. Nature *443*, 818-822.

Kagkli, D.M., Tache, R., Cogan, T.M., Hill, C., Casaregola, S., et Bonnarme, P. (2006). *Kluyveromyces lactis* and *Saccharomyces cerevisiae*, two potent deacidifying and volatile-sulphur-aroma-producing microorganisms of the cheese ecosystem. Applied microbiology and biotechnology *73*, 434-442.

Kamimura, E.S., Medieta, O., Rodrigues, M.I., et Maugeri, F. (2001). Studies on lipase-affinity adsorption using response-surface analysis. Biotechnology and applied biochemistry *33*, 153-159.

Keeling, P.J., et Palmer, J.D. (2008). Horizontal gene transfer in eukaryotic evolution. Nature reviews Genetics *9*, 605-618.

Kellis, M., Birren, B.W., et Lander, E.S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature *428*, 617-624.

Kerscher, S., Durstewitz, G., Casaregola, S., Gaillardin, C., et Brandt, U. (2001). The complete mitochondrial genome of yarrowia lipolytica. Comparative and functional genomics *2*, 80-90.

Khaldi, N., Collemare, J., Lebrun, M.H., et Wolfe, K.H. (2008). Evidence for horizontal transfer of a secondary metabolite gene cluster between fungi. Genome biology *9*, R18.

Khaldi, N., et Wolfe, K.H. (2008). Elusive origins of the extra genes in *Aspergillus oryzae*. PloS one *3*, e3036.

Khaldi, N., et Wolfe, K.H. (2011). Evolutionary Origins of the Fumonisin Secondary Metabolite Gene Cluster in *Fusarium verticillioides* and *Aspergillus niger*. International journal of evolutionary biology *2011*, 423821.

Kidwell, M.G. (1993). Lateral transfer in natural populations of eukaryotes. Annual review of genetics *27*, 235-256.

Kim, S.J., et Shoda, M. (1999). Batch decolorization of molasses by suspended and immobilized fungus of *Geotrichum candidum* Dec 1. Journal of bioscience and bioengineering *88*, 586-589.

Klaassen, C.H. (2009). MLST versus microsatellites for typing *Aspergillus fumigatus* isolates. Medical mycology : official publication of the International Society for Human and Animal Mycology *47 Suppl 1*, S27-33.

Knop, M. (2006). Evolution of the hemiascomycete yeasts: on life styles and the importance of inbreeding. BioEssays : news and reviews in molecular, cellular and developmental biology *28*, 696-708.

Koszul, R., Caburet, S., Dujon, B., et Fischer, G. (2004). Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. The EMBO journal *23*, 234-243.

Koszul, R., Malpertuy, A., Frangeul, L., Bouchier, C., Wincker, P., Thierry, A., Duthoy, S., Ferris, S., Hennequin, C., et Dujon, B. (2003). The complete mitochondrial genome sequence of the pathogenic yeast *Candida* (*Torulopsis*) *glabrata*. FEBS letters *534*, 39-48.

Kurtzman, C.P. (1992). rRNA sequence comparisons for assessing phylogenetic relationships among yeasts. International journal of systematic bacteriology *42*, 1-6.

Kurtzman, C.P. (2003). Phylogenetic circumscription of *Saccharomyces*, *Kluyveromyces* and other members of the Saccharomycetaceae, et the proposal of the new genera *Lachancea*, *Nakaseomyces*, *Naumovia*, *Vanderwaltozyma* and *Zygotorulaspora*. FEMS yeast research *4*, 233-245.

Kurtzman, C.P. (2011). Phylogeny of the ascomycetous yeasts and the renaming of *Pichia anomala* to Wickerhamomyces anomalus. Antonie van Leeuwenhoek *99*, 13-23.

Kurtzman, C.P., et Fell, J.W. (1998). The Yeasts - A Taxonomic Study (Elsevier Science).

Kurtzman, C.P., Fell, J.W., et Boekhout, T. (2011). The Yeasts: A Taxonomic Study (Elsevier).

Kurtzman, C.P., et Robnett, C.J. (1995). Molecular relationships among hyphal ascomycetous yeasts and yeastlike taxa. Canadian Journal of Botany *73*, 824-830.

Kurtzman, C.P., et Robnett, C.J. (1998). Identification and phylogeny of ascomycetous yeasts from analysis of nuclear large subunit (26S) ribosomal DNA partial sequences. Antonie van Leeuwenhoek *73*, 331-371.

Kurtzman, C.P., et Robnett, C.J. (2003). Phylogenetic relationships among yeasts of the '*Saccharomyces* complex' determined from multigene sequence analyses. FEMS yeast research *3*, 417-432.

Kurtzman, C.P., et Robnett, C.J. (2012). Relationships Among Genera of the Saccharomycotina (Ascomycota) from Multigene Phylogenetic Analysis of Type Species. FEMS yeast research *17*, 1567-1364.

Lachance, M.A., Daniel, H.M., Meyer, W., Prasad, G.S., Gautam, S.P., et Boundy-Mills, K. (2003). The D1/D2 domain of the large-subunit rDNA of the yeast species *Clavispora lusitaniae* is unusually polymorphic. FEMS yeast research *4*, 253-258.

Langkjaer, R.B., Casaregola, S., Ussery, D.W., Gaillardin, C., et Piskur, J. (2003). Sequence analysis of three mitochondrial DNA molecules reveals interesting differences among *Saccharomyces yeasts*. Nucleic Acids Res *31*, 3081-3091.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R.*, et al.* (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947-2948.

Larpin, S., Mondoloni, C., Goerges, S., Vernoux, J.P., Gueguen, M., et Desmasures, N. (2006). *Geotrichum candidum* dominates in yeast population dynamics in Livarot, a French red-smear cheese. FEMS yeast research *6*, 1243-1253.

Leclercq-Perlat, M.N., Oumer, A., Bergere, J.L., Spinnler, H.E., et Corrieu, G. (2000). Behavior of *Brevibacterium linens* and *Debaryomyces hansenii* as ripening flora in controlled production of smear soft cheese from reconstituted milk: growth and substrate consumption dairy foods. J Dairy Sci *83*, 1665-1673.

Lecocq, J., Gueguen, M., et Coiffier, O. (1996). Importance de l'association *Geotrichum candidum* : Brevibacterium linens pour l'affinage de fromages, Vol 16 (Paris, FRANCE: Lavoisier).

Lee, T.H., Aoki, H., Sugano, Y., et Shoda, M. (2000). Effect of molasses on the production and activity of dye-decolorizing peroxidase from *Geotrichum candidum* Dec1. Journal of bioscience and bioengineering *89*, 545-549.

Legras, J.L., et Karst, F. (2003). Optimisation of interdelta analysis for *Saccharomyces cerevisiae* strain characterisation. FEMS microbiology letters *221*, 249-255.

Legras, J.L., Merdinoglu, D., Cornuet, J.M., et Karst, F. (2007). Bread, beer and wine: *Saccharomyces cerevisiae* diversity reflects human history. Molecular ecology *16*, 2091-2102.

Legras, J.L., Ruh, O., Merdinoglu, D., et Karst, F. (2005). Selection of hypervariable microsatellite loci for the characterization of *Saccharomyces cerevisiae* strains. International journal of food microbiology *102*, 73-83.

Lépingle, A., Casaregola, S., Neuvéglise, C., Bon, E., Nguyen, H.V., Artiguenave, F., Wincker, P., et Gaillardin, C. (2000). Genomic Exploration of the Hemiascomycetous Yeasts: 14. *Debaryomyces hansenii var*. *hansenii*. FEBS letters *487*, 82-86.

Li, R., Li, Y., Kristiansen, K., et Wang, J. (2008). SOAP: short oligonucleotide alignment program. Bioinformatics *24*, 713-714.

Liti, G., Barton, D.B., et Louis, E.J. (2006). Sequence diversity, reproductive isolation and species concepts in *Saccharomyces*. Genetics *174*, 839-850.

Liti, G., Carter, D.M., Moses, A.M., Warringer, J., Parts, L., James, S.A., Davey, R.P., Roberts, I.N., Burt, A., Koufopanou, V.*, et al.* (2009). Population genomics of domestic and wild yeasts. Nature *458*, 337-341.

Liu, Z.L., et Moon, J. (2009). A novel NADPH-dependent aldehyde reductase gene from Saccharomyces cerevisiae NRRL Y-12632 involved in the detoxification of aldehyde inhibitors derived from lignocellulosic biomass conversion. Gene *446*, 1-10.

Lopandic, K., Zelger, S., Banszky, L.K., Eliskases-Lechner, F., et Prillinger, H. (2006). Identification of yeasts associated with milk products using traditional and molecular techniques. Food microbiology *23*, 341-350.

Lott, T.J., Frade, J.P., et Lockhart, S.R. (2010). Multilocus sequence type analysis reveals both clonality and recombination in populations of *Candida glabrata* bloodstream isolates from U.S. surveillance studies. Eukaryotic cell *9*, 619-625.

Lowe, T.M., et Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res *25*, 955-964.

Luo, Z., et van Vuuren, H.J. (2009). Functional analyses of PAU genes in *Saccharomyces cerevisiae*. Microbiology *155*, 4036-4049.

Lynch, D.B., Logue, M.E., Butler, G., et Wolfe, K.H. (2010). Chromosomal G + C content evolution in yeasts: systematic interspecies differences, and GC-poor troughs at centromeres. Genome biology and evolution *2*, 572-583.

Lynch, M., et Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. Science 290, 1151-1155.

Maiden, M.C., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., et *al.* (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proceedings of the National Academy of Sciences of the United States of America *95*, 3140-3145.

Mallet, L.V., Becq, J., et Deschavanne, P. (2010). Whole genome evaluation of horizontal transfers in the pathogenic fungus *Aspergillus fumigatus*. BMC genomics *11*, 171.

Mansour, S., Beckerich, J.M., et Bonnarme, P. (2008). Lactate and amino acid catabolism in the cheese-ripening yeast *Yarrowia lipolytica*. Applied and environmental microbiology *74*, 6505-6512.

Marcellino, N., Beuvier, E., Grappin, R., Gueguen, M., et Benson, D.R. (2001). Diversity of *Geotrichum candidum* strains isolated from traditional cheesemaking fabrications in France. Applied and environmental microbiology *67*, 4752-4759.

Mariani, C., Briandet, R., Chamba, J.F., Notz, E., Carnet-Pantiez, A., Eyoug, R.N., et Oulahal, N. (2007). Biofilm Ecology of Wooden Shelves Used in Ripening the French Raw Milk Smear Cheese Reblochon de Savoie. Journal of Dairy Science *90*, 1653-1661.

Mariani, C., Oulahal, N., Chamba, J.F., Dubois-Brissonnet, F., Notz, E., et Briandet, R. (2011). Inhibition of *Listeria monocytogenes* by resident biofilms present on wooden shelves used for cheese ripening. Food Control *22*, 1357-1362.

Marshall, M., Russo, G., Van Etten, J., et Nickerson, K. (1979). Polyamines in dimorphic fungi. Current microbiology *2*, 187-190.

Martin, N., Berger, C., Le Du, C., et Spinnler, H.E. (2001). Aroma compound production in cheese curd by coculturing with selected yeast and bacteria. J Dairy Sci *84*, 2125-2135.

Martin, T., Lu, S.W., van Tilbeurgh, H., Ripoll, D.R., Dixelius, C., Turgeon, B.G., et Debuchy, R. (2010). Tracing the origin of the fungal alpha1 domain places its ancestor in the HMG-box superfamily: implication for fungal mating-type evolution. PloS one *5*.

Mattanovich, D., Graf, A., Stadlmann, J., Dragosits, M., Redl, A., Maurer, M., Kleinheinz, M., Sauer, M., Altmann, F., et Gasser, B. (2009). Genome, secretome and glucose transport highlight unique features of the protein production host *Pichia pastoris*. Microbial cell factories *8*, 29.

McManus, B.A., Coleman, D.C., Moran, G., Pinjon, E., Diogo, D., Bougnoux, M.E., Borecka-Melkusova, S., Bujdakova, H., Murphy, P., d'Enfert, C.*, et al.* (2008). Multilocus sequence typing reveals that the population structure of *Candida dubliniensis* is significantly less divergent than that of Candida albicans. Journal of clinical microbiology *46*, 652-664.

Meyer, W., Aanensen, D.M., Boekhout, T., Cogliati, M., Diaz, M.R., Esposto, M.C., Fisher, M., Gilgado, F., Hagen, F., Kaocharoen, S.*, et al.* (2009). Consensus multi-locus sequence typing scheme for *Cryptococcus neoformans* and *Cryptococcus gattii*. Medical mycology : official publication of the International Society for Human and Animal Mycology *47*, 561-570.

Minguet, E.G., Vera-Sirera, F., Marina, A., Carbonell, J., et Blazquez, M.A. (2008). Evolutionary diversification in polyamine biosynthesis. Molecular biology and evolution *25*, 2119-2128.

Mounier, J., Monnet, C., Jacques, N., Antoinette, A., et Irlinger, F. (2009). Assessment of the microbial diversity at the surface of Livarot cheese using culture-dependent and independent approaches. International journal of food microbiology *133*, 31-37.

Mounier, J., Monnet, C., Vallaeys, T., Arditi, R., Sarthou, A.S., Helias, A., et Irlinger, F. (2008). Microbial interactions within a cheese microbial community. Applied and environmental microbiology *74*, 172-181.

Mourgues, R., Bergère, J.L., et Vassal, L. (1983). Possibilités d'améliorer les qualités organoleptiques des fromages de Camembert grâce à l'utilisation de *Geotrichum candidum*. La Technique Laitière 978, 11-15.

Munoz, R., Gomez, A., Robles, V., Rodriguez, P., Cebollero, E., Tabera, L., Carrascosa, A.V., et Gonzalez, R. (2009). Multilocus sequence typing of oenological *Saccharomyces cerevisiae* strains. Food microbiology *26*, 841-846.

Nakao, Y., Kanamori, T., Itoh, T., Kodama, Y., Rainieri, S., Nakamura, N., Shimonaga, T., Hattori, M., et Ashikari, T. (2009). Genome sequence of the lager brewing yeast, an interspecies hybrid. DNA research : an international journal for rapid publication of reports on genes and genomes *16*, 115-129.

Naumov, G.I., Naumova, E.S., et Korhola, M.P. (1995). Chromosomal polymorphism of MEL genes in some populations of *Saccharomyces cerevisiae*. FEMS microbiology letters *127*, 41-45.

Ness, F., Lavallée, F., Dubourdieu, D., Aigle, M., et Dulau, L. (1993). Identification of yeast strains using the polymerase chain reaction. Journal of the science of food and agriculture *62*, 89-94.

Neuveglise, C., Feldmann, H., Bon, E., Gaillardin, C., et Casaregola, S. (2002). Genomic evolution of the long terminal repeat retrotransposons in hemiascomycetous yeasts. Genome research *12*, 930-943.

Neuveglise, C., Marck, C., et Gaillardin, C. (2011). The intronome of budding yeasts. Comptes rendus biologies *334*, 662-670.

Nickerson, K.W., Dunkle, L.D., et Van Etten, J.L. (1977). Absence of spermine in filamentous fungi. Journal of bacteriology *129*, 173-176.

Nilsson-Tillgren, T., Gjermansen, C., Kielland-Brandt, M.C., Petersen, J.G.L., et Holmberg., S. (1981). Genetic differences between *Saccharomyces carlsbergensis* and *S. cerevisiae*. Analysis of chromosome III by single chromosome transfer. Carlsberg Res Commun *46*, 65-76.

Novo, M., Bigey, F., Beyne, E., Galeote, V., Gavory, F., Mallet, S., Cambon, B., Legras, J.L., Wincker, P., Casaregola, S.*, et al.* (2009). Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast *Saccharomyces cerevisiae* EC1118. Proceedings of the National Academy of Sciences of the United States of America *106*, 16333-16338.

Ohno, S. (1970). Evolution by Gene Duplication. Springer.

Pan, W., Khayhan, K., Hagen, F., Wahyuningsih, R., Chakrabarti, A., Chowdhary, A., Ikeda, R., Taj-Aldeen, S.J., Khan, Z., Imran, D.*, et al.* (2012). Resistance of Asian *Cryptococcus neoformans* serotype A is confined to few microsatellite genotypes. PloS one *7*, e32868.

Park, H., et Bakalinsky, A.T. (2000). SSU1 mediates sulphite efflux in *Saccharomyces cerevisiae*. Yeast *16*, 881-888.

Paulus, T.J., Kiyono, P., et Davis, R.H. (1982). Polyamine-deficient *Neurospora crassa* mutants and synthesis of cadaverine. Journal of bacteriology *152*, 291-297.

Perez-Ortin, J.E., Querol, A., Puig, S., et Barrio, E. (2002). Molecular characterization of a chromosomal rearrangement involved in the adaptive evolution of yeast strains. Genome research *12*, 1533-1539.

Perriere, G., et Gouy, M. (1996). WWW-query: an on-line retrieval system for biological sequence banks. Biochimie *78*, 364-369.

Pfliegler, W.P., Antunovics, Z., et Sipiczki, M. (2012). Double sterility barrier between *Saccharomyces* species and its breakdown in allopolyploid hybrids by chromosome loss. FEMS yeast research *12*, 703-718.

Pottier, I., Gente, S., Vernoux, J.P., et Gueguen, M. (2008). Safety assessment of dairy microorganisms: *Geotrichum candidum*. International journal of food microbiology *126*, 327-332.

Pryde, F.E., Gorham, H.C., et Louis, E.J. (1997). Chromosome ends: all the same under their caps. Current Opinion in Genetics & Development *7*, 822-828.

Rachidi, N., Barre, P., et Blondin, B. (1999). Multiple *Ty*-mediated chromosomal translocations lead to karyotype changes in a wine strain of *Saccharomyces cerevisiae*. Molecular & general genetics : MGG *261*, 841-850.

Rainieri, S., Kodama, Y., Kaneko, Y., Mikata, K., Nakao, Y., et Ashikari, T. (2006). Pure and mixed genetic lines of *Saccharomyces bayanus* and *Saccharomyces pastorianus* and their contribution to the lager brewing strain genome. Applied and environmental microbiology *72*, 3968-3974.

Ramezani-Rad, M., Hollenberg, C.P., Lauber, J., Wedler, H., Griess, E., Wagner, C., Albermann, K., Hani, J., Piontek, M., Dahlems, U.*, et al.* (2003). The *Hansenula polymorpha* (strain CBS4732) genome sequencing and analysis. FEMS yeast research *4*, 207-215.

Redhead, S.A., et Malloch, D.W. (1977). The Endomycetaceae: new concepts, new taxa. Canadian Journal of Botany *55*, 1701-1711.

Reedy, J.L., Floyd, A.M., et Heitman, J. (2009). Mechanistic plasticity of sexual reproduction and meiosis in the *Candida* pathogenic species complex. Curr Biol 19, 891-899.

Rey, A. (1994). Le Robert Dictionnaire historique de la langue française. Dictionnaires Le Robert tome I, 848.

Rodionova, N.A., Dubovaia, N.V., Eneiskaia, E.V., Martinovich, L.I., Gracheva, I.M., et Bezborodov, A.M. (2000). [Purification and characteristic of endo-(1--4)-beta-xylanase from *Geotrichum candidum* 3C]. Prikladnaia biokhimiia i mikrobiologiia *36*, 535-540.

Romano, A., Casaregola, S., Torre, P., et Gaillardin, C. (1996). Use of RAPD and mitochondrial DNA RFLP for typing of *Candida zeylanoides* and *Debaryomyces hansenii* yeast strains isolated from cheese. Systematic and applied microbiology *19*, 255-264.

Romano, P., Ricciardi, A., Salzano, G., et Suzzi, G. (2001). Yeasts from Water Buffalo Mozzarella, a traditional cheese of the Mediterranean area. International journal of food microbiology *69*, 45-51.

Roostita, R., et Fleet, G.H. (1996a). Growth of yeasts in milk and associated changes to milk composition. International journal of food microbiology *31*, 205-219.

Roostita, R., et Fleet, G.H. (1996b). The occurrence and growth of yeasts in Camembert and blue-veined cheeses. International journal of food microbiology *28*, 393-404.

Ropars, J., Cruaud, C., Lacoste, S., et Dupont, J. (2012). A taxonomic and ecological overview of cheese fungi. International journal of food microbiology *155*, 199-210.

Rosewich, U.L., et Kistler, H.C. (2000). Role of Horizontal Gene Transfer in the Evolution of Fungi. Annual review of phytopathology *38*, 325-363.

Sarda, L., et Desnuelle, P. (1958). [Actions of pancreatic lipase on esters in emulsions]. Biochimica et biophysica acta *30*, 513-521.

Schacherer, J., Shapiro, J.A., Ruderfer, D.M., et Kruglyak, L. (2009). Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. Nature *458*, 342-345.

Schacherer, J., Tourrette, Y., Souciet, J.L., Potier, S., et De Montigny, J. (2004). Recovery of a function involving gene duplication by retroposition in *Saccharomyces cerevisiae*. Genome research *14*, 1291-1297.

Schiex, T., Moisan, A., et Rouzé, P. (2001). EuGène: An eucaryotic gene finder that combines several sources of evidence. Computational biology 111-125.

Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A., et Chen, W. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proceedings of the National Academy of Sciences of the United States of America *109*, 6241-6246.

Seiler, H., et Busse, M. (1990). The yeasts of cheese brines. International journal of food microbiology *11*, 289-303.

Serrat, M., Bermudez, R.C., et Villa, T.G. (2002). Production, purification, et characterization of a polygalacturonase from a new strain of *Kluyveromyces marxianus* isolated from coffee wet-processing wastewater. Applied biochemistry and biotechnology *97*, 193-208.

Sherman, D.J., Martin, T., Nikolski, M., Cayla, C., Souciet, J.L., et Durrens, P. (2009). Genolevures: protein families and synteny among complete hemiascomycetous yeast proteomes and genomes. Nucleic Acids Res *37*, 16.

Sidrim, J.J., Costa, A.K., Cordeiro, R.A., Brilhante, R.S., Moura, F.E., Castelo-Branco, D.S., Neto, M.P., et Rocha, M.F. (2010). Molecular methods for the diagnosis and characterization of *Cryptococcus*: a review. Canadian journal of microbiology *56*, 445-458.

Slot, J.C., et Hibbett, D.S. (2007). Horizontal transfer of a nitrate assimilation gene cluster and ecological transitions in fungi: a phylogenetic study. PloS one *2*, e1097.

Slot, J.C., et Rokas, A. (2011). Horizontal Transfer of a Large and Highly Toxic Secondary Metabolic Gene Cluster between Fungi. Current biology : CB *21*, 134-139.

Smith, M.T., de Cock, A.W., Poot, G.A., et Steensma, H.Y. (1995). Genome comparisons in the yeastlike fungal genus *Galactomyces* Redhead et Malloch. International journal of systematic bacteriology *45*, 826-831.

Smith, M.T., Poot, G.A., et de Cock, A.W. (2000). Re-examination of some species of the genus *Geotrichum* Link: Fr. Antonie van Leeuwenhoek *77*, 71-81.

Sohier, D., Dizes, A.S., Thuault, D., Neuveglise, C., Coton, E., et Casaregola, S. (2009). Important genetic diversity revealed by inter-LTR PCR fingerprinting of *Kluyveromyces marxianus* and *Debaryomyces hansenii* strains from French traditional cheeses. Dairy Sci Technol *89*, 569-581.

Souciet, J.L., Dujon, B., Gaillardin, C., Johnston, M., Baret, P.V., Cliften, P., Sherman, D.J., Weissenbach, J., Westhof, E., Wincker, P.*, et al.* (2009). Comparative genomics of protoploid Saccharomycetaceae. Genome research *19*, 1696-1709.

Spencer, J.F., et Spencer, D.M. (1996). Rare-mating and cytoduction in *Saccharomyces cerevisiae*. Methods Mol Biol *53*, 39-44.

Spinnler, H.E., Berger, C., Lapadatescu, C., et Bonnarme, P. (2001). Production of sulfur compounds by several yeasts of technological interest for cheese ripening. International Dairy Journal *11*, 245-252.

Stajich, J.E., Dietrich, F.S., et Roy, S.W. (2007). Comparative genomic analysis of fungal genomes reveals intron-rich ancestors. Genome biology *8*, R223.

Sugano, Y., Matsuo, C., et Shoda, M. (2001). Efficient production of a heterologous peroxidase, DyP from *Geotrichum candidum* Dec 1, on solid-state culture of *Aspergillus oryzae* RD005. Journal of bioscience and bioengineering *92*, 594-597.

Sugano, Y., Nakano, R., Sasaki, K., et Shoda, M. (2000). Efficient heterologous expression in *Aspergillus oryzae* of a unique dye-decolorizing peroxidase, DyP, of *Geotrichum candidum* Dec 1. Applied and environmental microbiology *66*, 1754-1758.

Sugano, Y., Sasaki, K., et Shoda, M. (1999). cDNA cloning and genetic analysis of a novel decolorizing enzyme, peroxidase gene dyp from *Geotrichum candidum* Dec 1. Journal of bioscience and bioengineering *87*, 411-417.

Suh, S.O., et Blackwell, M. (2006). Three new asexual arthroconidial yeasts, Geotrichum carabidarum sp. nov., *Geotrichum histeridarum sp*. *nov*., and *Geotrichum cucujoidarum sp*. *nov*., isolated from the gut of insects. Mycological research *110*, 220-228.

Sulo, P., Laurencik, M., Polakova, S., Minarik, G., et Slavikova, E. (2009). *Geotrichum bryndzae sp*. *nov*., a novel asexual arthroconidial yeast species related to the genus Galactomyces. International journal of systematic and evolutionary microbiology *59*, 2370-2374.

Talla, E., Anthouard, V., Bouchier, C., Frangeul, L., et Dujon, B. (2005). The complete mitochondrial genome of the yeast *Kluyveromyces thermotolerans*. FEBS letters *579*, 30-40.

Tavanti, A., Davidson, A.D., Johnson, E.M., Maiden, M.C., Shaw, D.J., Gow, N.A., et Odds, F.C. (2005). Multilocus sequence typing for differentiation of strains of *Candida tropicalis*. Journal of clinical microbiology *43*, 5593-5600.

Taylor, J.W., et Fisher, M.C. (2003). Fungal multilocus sequence typing--it's not just for bacteria. Current opinion in microbiology *6*, 351-356.

Temporini, E.D., et VanEtten, H.D. (2004). An analysis of the phylogenetic distribution of the pea pathogenicity genes of *Nectria haematococca* MPVI supports the hypothesis of their origin by horizontal transfer and uncovers a potentially new pathogen of garden pea: *Neocosmospora boniensis*. Current genetics *46*, 29-36.

van den Berg, M.A., Albang, R., Albermann, K., Badger, J.H., Daran, J.M., Driessen, A.J., Garcia-Estrada, C., Fedorova, N.D., Harris, D.M., Heijne, W.H.*, et al.* (2008). Genome sequencing and analysis of the filamentous fungus *Penicillium chrysogenum*. Nat Biotechnol *26*, 1161-1168.

Vanhee, L.M., Nelis, H.J., et Coenye, T. (2010). What can be learned from genotyping of fungi? Medical mycology : official publication of the International Society for Human and Animal Mycology *48 Suppl 1*, S60-69.

Vanhee, L.M., Symoens, F., Jacobsen, M.D., Nelis, H.J., et Coenye, T. (2009). Comparison of multiple typing methods for *Aspergillus fumigatus*. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases *15*, 643-650.

Vasei, M., et Imanieh, M.H. (1999). Duodenal colonization by *Geotrichum candidum* in a child with transient low serum levels of IgA and IgM. APMIS : acta pathologica, microbiologica, et immunologica Scandinavica *107*, 681-684.

Vassal, L., Monnet, V., Le Bars, D., Rouc, C., et Gripon, J.-C. (1986). Relation entre le pH, la composition chimique et la texture des fromages de type Camembert. Lait *66*, 341-351.

Viljoen, B.C., et Greyling, T. (1995). Yeasts associated with Cheddar and Gouda making. International journal of food microbiology *28*, 79-88.

Wang, Q.M., Liu, W.Q., Liti, G., Wang, S.A., et Bai, F.Y. (2012). Surprisingly diverged populations of *Saccharomyces cerevisiae* in natural environments remote from human activity. Molecular ecology.

Wang, S.A., et Bai, F.Y. (2008). *Saccharomyces arboricolus sp. nov.*, a yeast species from tree bark. International journal of systematic and evolutionary microbiology *58*, 510-514.

Warmington, J.R., Green, R.P., Newlon, C.S., et Oliver, S.G. (1987). Polymorphisms on the right arm of yeast chromosome III associated with *Ty* transposition and recombination events. Nucleic Acids Res *15*, 8963-8982.

Welsh, J., et McClelland, M. (1990). Fingerprinting genomes using PCR with arbitrary primers. Nucleic Acids Res *18*, 7213-7218.

White, T.J., Bruns , T.D., Lee, S., et Taylor, J.W. (1990). Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. PCR Protocols: a guide to methods and applications, 315–322.

Williams, J.G., Kubelik, A.R., Livak, K.J., Rafalski, J.A., et Tingey, S.V. (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res *18*, 6531-6535.

Wojtatowicz, M., Chrzanowska, J., Juszczyk, P., Skiba, A., et Gdula, A. (2001). Identification and biochemical characteristics of yeast microflora of Rokpol cheese. International journal of food microbiology *69*, 135-140.

Wolfe, K.H., et Shields, D.C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. Nature *387*, 708-713.

Wong, S., Fares, M.A., Zimmermann, W., Butler, G., et Wolfe, K.H. (2003). Evidence from comparative genomics for a complete sexual cycle in the 'asexual' pathogenic yeast *Candida glabrata*. Genome biology *4*, R10.

Woolfit, M., Rozpedowska, E., Piskur, J., et Wolfe, K.H. (2007). Genome survey sequencing of the wine spoilage yeast *Dekkera* (*Brettanomyces*) *bruxellensis*. Eukaryotic cell *6*, 721-733.

Wouters, J.T.M., Ayad, E.H.E., Hugenholtz, J., et Smit, G. (2002). Microbes from raw milk for fermented dairy products. International Dairy Journal *12*, 91-109.

Wright, S., et Finnegan, D. (2001). Genome evolution: sex and the transposable element. Curr Biol *11*, R296-299.

Xufre, A., Albergaria, H., Girio, F., et Spencer-Martins, I. (2011). Use of interdelta polymorphisms of *Saccharomyces cerevisiae* strains to monitor population evolution during wine fermentation. Journal of industrial microbiology & biotechnology *38*, 127-132.

Zivanovic, Y., Wincker, P., Vacherie, B., Bolotin-Fukuhara, M., et Fukuhara, H. (2005). Complete nucleotide sequence of the mitochondrial DNA from *Kluyveromyces lactis*. FEMS yeast research *5*, 315-322.

Zolan, M.E. (1995). Chromosome-length polymorphism in fungi. Microbiological reviews *59*, 686-698.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31, 3406-3415.

ANNEXE 1

C. R. Biologies 334 (2011) 590–598

Contents lists available at ScienceDirect

Comptes Rendus Biologies

www.sciencedirect.com

Taxonomy/Taxinomie

New perspectives in hemiascomycetous yeast taxonomy

Nouvelles perspectives en taxinomie des levures hémiascomycètes

Serge Casaregola *, Stéphanie Weiss, Guillaume Morel

INRA, UMR 1319, Micalis Institute, AgroParisTech, CIRM-Levures, 78850 Thiverval-Grignon, France

ARTICLE INFO

Article history:
Received 8 November 2010
Accepted after revision 1 April 2011
Available online 30 June 2011

Keywords:
Hemiascomycetous yeasts
Phylogeny
Evolution
Ribosomal DNA
Interspecific hybrid
Systematics

ABSTRACT

DNA sequencing has revolutionized yeast taxonomy. Although initially rDNA sequences proved to be universal and convenient for assigning phylogenetic relationships, it was eventually supplanted by multigene analysis, which provided more discriminating and robust results. This led to a new classification of the major yeast clades, which is still used as a reference today. More recently, the availability of a large number of complete genome sequences has given a new perspective on the molecular taxonomy of yeasts by providing a high number of genes to compare. It also highlighted an unexpected aspect of yeast genome evolution: the existence of interspecific hybrids outside of the industrial *Saccharomyces* clade. Together with the loss of heterozygosity in interspecific hybrids and a reduced sexuality leading to clonal propagation, this observation obliges us to reexamine the present concept of species. In parallel, the ongoing challenge is to find a universal molecular marker, to improve fast authentication and, if possible, phylogeny of yeasts. The future of yeast taxonomy will involve the sequencing of more genomes, thorough analysis of populations to obtain a good representation of the biodiversity and integration of these data into dedicated databases.

Mots clés :
Levures
Phylogénie
Évolution
ADN ribosomique
Hybride inter-espèce
Systématique

RÉSUMÉ

Le séquençage de l'ADN a complètement bouleversé notre vision de la taxinomie des espèces. Alors que les séquences d'ADN ribosomiques ont fait preuve de leur universalité et se montrent appropriées à cette tâche, il est rapidement apparu que les analyses multigéniques sont plus discriminantes et fournissent des arbres plus robustes. Cela a conduit à une nouvelle classification des principaux clades de levures, qui est toujours d'actualité aujourd'hui. Plus récemment, la disponibilité d'un nombre croissant de génomes complets a apporté une nouvelle perspective à la taxinomie moléculaire, en fournissant un grand nombre de gènes à comparer. Cela a également mis au jour un aspect inattendu de l'évolution des génomes de levure : l'existence d'hybrides inter-espèces, en dehors du clade du cas bien connu des *Saccharomyces* industrielles. Combinées à la perte d'hétérozygotie et à une sexualité réduite conduisant à une propagation clonale, ces observations nous obligent à réexaminer la définition actuelle des espèces. En parallèle, le challenge actuel est toujours de trouver un marqueur moléculaire universel, afin d'améliorer la rapidité de l'identification, si possible, la phylogénie des levures. Le futur de la taxinomie des levures passera par le séquençage de génomes supplémentaires, en

Abbreviations: KOG, euKaryotic Orthologous Group; LOH, Loss of Heterozygosity; mtDNA, mitochondrial DNA; RFLP, Restriction Fragment Length Polymorphism; WGD, Whole Genome Duplication.

* Corresponding author.
E-mail address: serge.casaregola@grignon.inra.fr (S. Casaregola).

1631-0691/$ – see front matter
doi:10.1016/j.crvi.2011.05.006

incluant une analyse minutieuse des populations afin d'obtenir une bonne représentation de la diversité, et enfin d'intégrer ces données dans des bases de données dédiées.

1. Introduction

The taxonomy of fungal species has been debated for a long time. Among fungi, the taxonomy of hemiascomycetes provides an extra challenge. Although it has been proposed that hemiascomycetous yeasts have evolved for as long as the chordates [1], this taxon is morphologically very homogenous. Therefore, characteristics such as morphology that are still valuable in establishing a fungal taxonomy are less used, if used at all, in yeasts.

The most comprehensive description of yeast species, The Yeasts: a taxonomic study, is now 13 years old and describes over 750 species [2]. A new edition of this series is now published [3]. It lists 1500 yeast species [3]. The major work by Suh and collaborators on the discovery of new taxa in beetles [4] led Boekhout [5] to estimate the number of yeast species to be discovered by 2010 to be close to 3000. The discrepancy between the number of described species and the predicted number of species to be described suggests that a number of ecological niches have not been investigated yet and that cryptic species may have not received enough attention. These figures are somewhat low compared to the 1.5 million predicted extant fungal species [6]. Yeast species are distinguished according to the following characteristics: cellular morphology, type of conidiogenesis, comparative physiology, type of coenzyme Q and G+C content. Since these characteristics are prone to intra-species variability, the DNA/DNA reassociation technique was retained as the method of choice to distinguish species. Although it is still the recognized method used in bacterial taxonomy, in yeast taxonomy, this method may be affected by the large amount of highly conserved and highly repetitive sequences of ribosomal DNA and by the occurrence of hybrids. Furthermore, this method is time-consuming, inapplicable to a large number of strains and it does not provide a consistent phylogeny [7].

Molecular systematics has revolutionized taxonomy and our view of yeast evolution. It is interesting to look back and analyze how the choices of molecular markers have developed. The favorite marker for molecular taxonomy is rDNA, since it is slow evolving and therefore well conserved, thus allowing easy sequence comparison and facilitating of PCR amplification [8]. The first comprehensive analysis based on the entire small subunit 18S rDNA gene was published in 1993 [9]. This approach also proved time-consuming, since this part of the rDNA unit is around 1800 bp long. Nevertheless, the data allowed a clear separation of the hemiascomycetes from the filamentous euascomycetes.

Works by several authors on various fungi belonging to basidiomycetous and hemiascomycetous yeasts as well as euascomycetes led to the choice of the around 600 bp long D1/D2 variable region at the 5′ end of the 26S rDNA [10–13]. Most authors concentrated on this region and a comprehensive database of D1/D2 sequences for over 500 species became available in 1998 [14,15]. It provided the D1/D2 rDNA barcode for identification of hemiascomycetous yeasts to the species level for the ensuing decade, contributing thereby to a more straightforward phylogeny. Since 1998, numerous parts of the rDNA unit have been used for this purpose, including various studies using Restriction Fragment Length Polymorphism (RFLP) of different parts of the rDNA unit for rapid identification and species delineation, mainly the Internal Transcribed Spacers (ITS) [16] and the Non-Transcribed Spacers (NTS) [17]. Considering the properties of D1/D2 (short size, ease of amplification, and ubiquity), the use of other parts of rDNA was abandoned. However, in the meantime, work on fungi started using protein coding genes, which yielded better species delineation and led to evidence for sexuality in some fungi at the end of the 1990s [18,19]. First, the single-copy genes *RPB2* [20] and *ACT1* [21,22] were used in addition to D1/D2, but the taxonomy and the phylogeny of hemiascomycetes really developed concomitantly with the availability of genome sequences [1,23–34].

2. Multigene analysis

The sequence of the complete genome of *Saccharomyces cerevisiae* and the first Genolevures project opened up new horizons for yeast taxonomy [23,24]. The classification of the so-called "*Saccharomyces* complex" clade was brought in a pioneer work [35] based on the concatenation of various sequences: rDNA repeat (18S, 26S), single-copy nuclear genes (translation elongation factor 1, actin, RNA polymerase II) and mitochondrially encoded genes (rDNA small-subunit, *COXII*). It provided a new standard for the delineation of genera, based on the exclusion of polyphyly [36]. In particular, it made the relationship clear between the genus *Saccharomyces*, in the current sense (so-called *sensu stricto*) and now reduced to six species, and its closest neighbors, the *Saccharomyces sensu lato* species that are now classified into various genera such as *Kazachstania, Lachancea* and *Naumovia*. (Fig. 1).

Kurtzman and his collaborators have applied this method to a large number of clades. This led to the circumscription of many genera [36–42]. Most of the transfers of existing species to new genera established by Kurtzman and his collaborators are shown in Fig. 1. A more detailed account of Kurtzman's work is also described in his recent review [43]. It is noteworthy that in this analysis of 83 species, Kurtzman used only three markers, two being rDNA markers, which have been questioned for the bias they can introduce into phylogenetic analysis [44–46]. For instance, rDNA analysis incorrectly groups *Zygosaccharomyces rouxii* with *Nakaseomyces delphensis/Candida glabrata* [1,15], the latter having undergone Whole Genome Duplication (WGD) [1] like *Saccharomyces cerevisiae* [47]. The branching of *Z. rouxii* in rDNA phylogenies

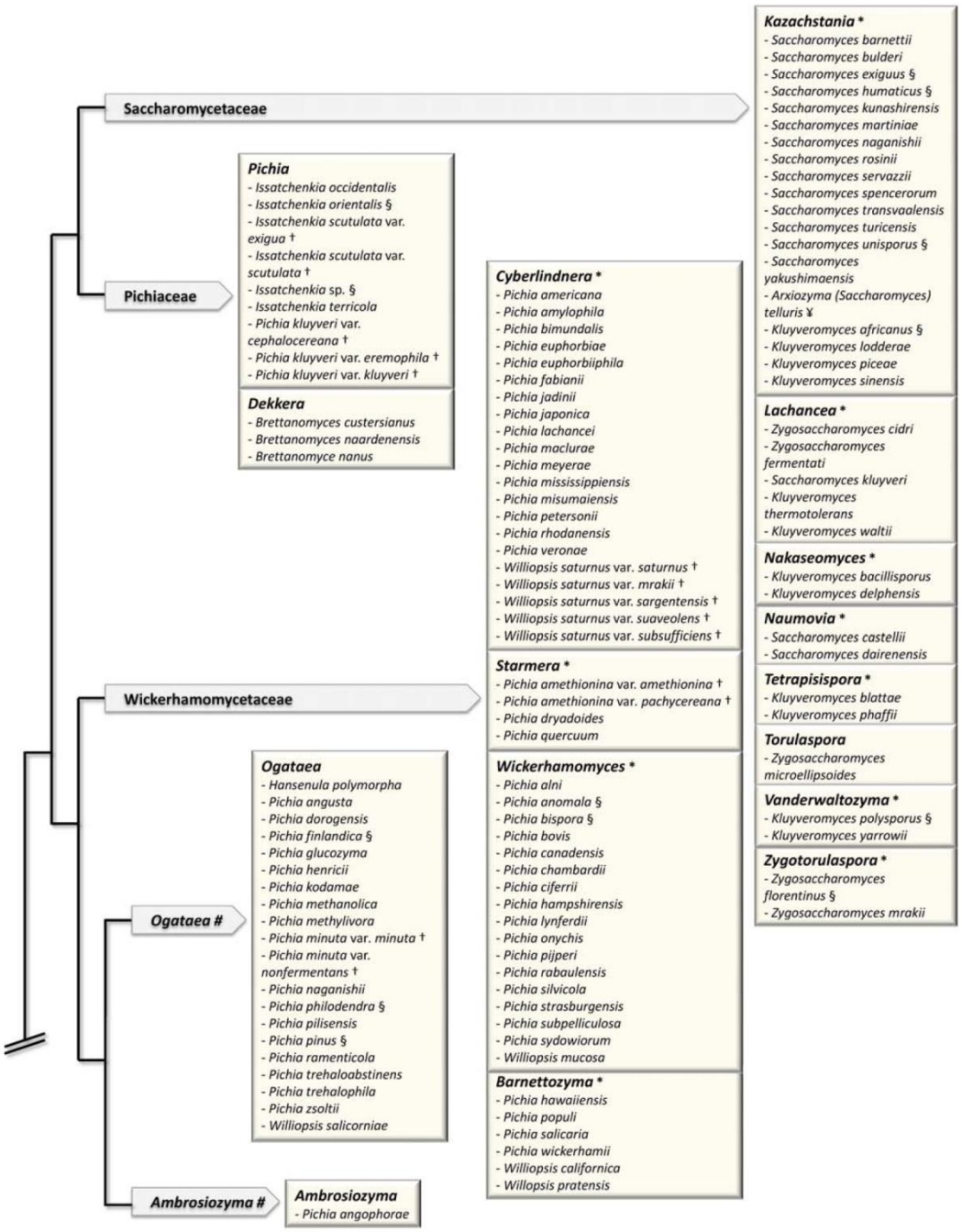


Fig. 1. Major changes in the classification of hemiascomycetes since 1998. The schematic tree of the families is adapted from [43]. Arrows contain names of families. Boxes correspond to genera in bold letters. (#) indicates genera, which are not yet assigned to a family. (*) indicates newly defined genera. Only the species, which have been reassigned to a different genus (that defined by the box) are shown. The previous name of the species, i.e. the genus to which each species belonged previously, is shown. For example, the species *Saccharomyces servazzii* is found in the box defined by the new genus *Kazachstania*. *Saccharomyces servazzii* is now *Kazachstania servazzii*. (§) indicates the species, which have also changed names: *Saccharomyces exiguus* = *Kazachstania exigua*; *Saccharomyces humaticus* = *Kazachstania humatica*; *Saccharomyces unisporus* = *Kazachstania unispora*; *Kluyveromyces africanus* = *Kazachstania africana*; *Kluyveromyces polysporus* = *Vanderwaltozyma polyspora*; *Zygosaccharomyces florentinus* = *Zygotorulaspora florentina*; *Issatchenka orientalis* = *Pichia kudriavzevii*; *Pichia finlandica* = *Ogataea wickerhamii*; *Pichia philodendra* = *Ogataea philodendri*; *Pichia pinus* = *Ogataea pini*; *Pichia anomala* = *Wickerhamomyces anomalus*; *Pichia bispora* = *Wickerhamomyces bisporus*; *Pichia haplophila* = *Priceomyces haplophilus*; *Pichia media* = *Priceomyces medius*; *Debaryomyces castellii* = *Schwanniomyces capriottii*. The two subspecies of *Lipomyces kononenkoae* var. *kononenkoae* and var. *spencermartinsiae* had their name changed to *L. kononenkoae* and *L. spencermartinsiae*, respectively. (†) indicates the species, whose name was previously a variety; for example *Pichia nakazawae* var. *akitaensis* is now *Yamadazyma akitaensis*. (¥) indicates the species, which were subdived into several species. *Saccharomyces* (*Arxiozyma*) *telluris* species is now subdivided into *Kazachstania telluris* and *Kazachstania bovina*. *Zygoascus stearolyticus* and *Zygoascus hellenicus* varieties are now replaced by *Z. hellenicus* and *Z. meyerae*.

Note that the *Candida* species that were shown to have a teleomorph were not included in the figure.

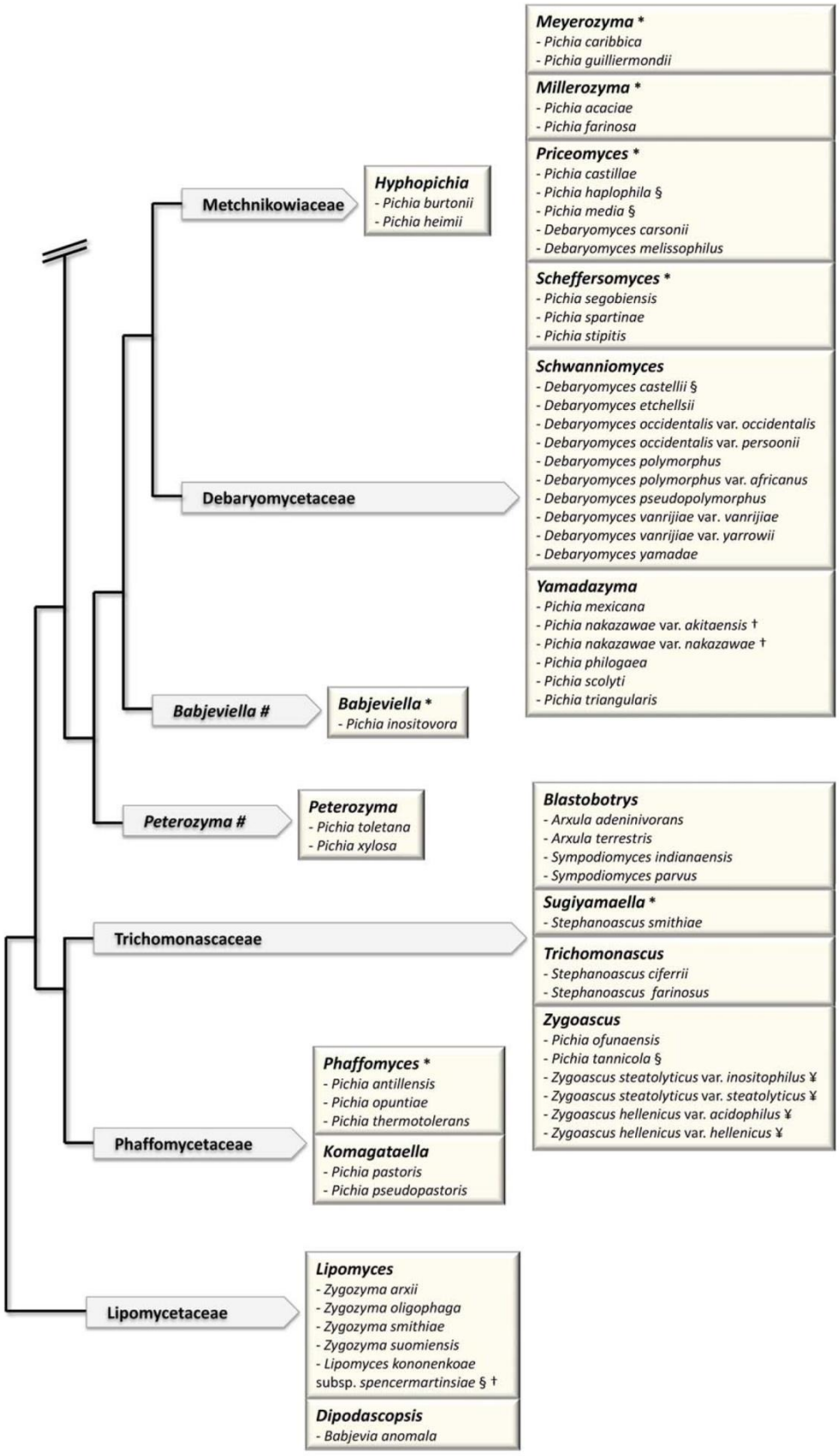


Fig. 1. (*Continued*).

is therefore inexact, since *Z. rouxii* has diverged from the ancestor of the clade before WGD occurred [48]. The need for genomic markers, other than repetitive sequences such as rDNA, that are informative from the phylogenetic point of view is therefore crucial. Genome sequence data have provided the potential for solving these problems.

3. Phylogenomics

With the increased availability of complete genome sequences, multigene analysis can extend to a large number of genes as long as real orthologs can be compared. Duplicated genes cannot be used for phylogeny, since the two copies generally evolve independently. It has been demonstrated that roughly 40% of yeasts genes have paralogues [1], which exclude them for phylogeny. With the availability of large datasets, an old question reappeared: which of the number of genes or taxa available are the most important in phylogeny reconstruction? An early study analyzing 14 yeast genomes and 106 genes [49] proposed that robustness of phylogenies was linked to the number of genes used, whereas the number of species had hardly any effect on the phylogenies. This study may seem audacious, since the species, which were analyzed, were widely diverging. Later on, studies that were more cautious included all available fungal genomes to yield the beginning of a tree of life for this part of the eukarya. A total of 531 genes derived from the euKaryotic Orthologous Groups (KOGs) of 25 species led to a well supported unique tree [45]; however, by simply reducing the number of genes used by 1/3, these authors showed that the phylogeny of some parts of the tree could not be resolved. Fitzpatrick et al. [44] found that, using 153 universally distributed orthologs of 42 genomes, robust relationships could not be established between some species like *C. glabrata* or *Saccharomyces castellii* and the WGD clade. The conflicting data obtained using various methods suggested that more taxa were needed to resolve this node. A similar result was obtained in a study in which no prior selection of genes or sites was performed [50]. Further work by Kuramae et al. [51] on 33 genomes helped to solve this problem. These studies therefore revealed that the number of species analyzed was crucial when inferring phylogenies by these approaches.

One of the most interesting conclusions of these studies was that the number of taxa to introduce in phylogenomic analysis depended on the genetic distance which separates the taxa. Indeed, the aforementioned studies analyzed species that span the whole fungal tree and that evolved over 1 billion years. Considering these large genetic distances, it was not clear whether this approach would be successful with closely related species. The difficulty at inferring robust relationships between closely related species was confirmed in a systematic comparison of all the models used for elaborating phylogenies (superalignments, supertrees, distance and gene content...) from complete genomes [52]. In response to Rokas and Carroll [49], it was even suggested that the inclusion of more genes in phylogenetic reconstructions could decrease accuracy, especially in the case of bias sampling [53]. Conversely, the reduction of those biases by the addition of extra taxa may result in the use of fewer genes for the phylogenetic analysis. A similar conclusion was reached by Aguileta et al. [54]. The direct consequence of this proposition is that a limited number of genes may be sufficient in order to establish robust phylogenies. This may be good news because we have to consider ten to hundred times more species in phylogenetic studies in the near future.

4. Hybrids, hybrids, hybrids...

The most famous example of yeast hybrids are the ones involved in beer making (reviewed by Kielland-Brandt et al. [55]). Since the discovery of the complexity of the brewing yeast genome that contained material DNA from at least two contributors, many hybrids between *Saccharomyces* species were evidenced (this is described in the Fermentative *Saccharomyces* chapter of this volume). The high occurrence of hybrids in this genus may be attributed to the fact that these yeasts have been used for millennia in biotechnology and that their genome was shaped by human activities. However, a few hybrids were found in other clades not belonging to the *Saccharomyces* genus. This is the case in several genera and species, e.g. *Candida*, *Kazachstania, Metschnikowia, Zygosaccharomyces* and *Debaryomyces* [56–59]. Hybrids are not specific to hemiascomycetes; they have been shown to exist in basidiomycetous yeasts [60].

Interspecific hybridization may yield stable haploid strains, which may or may not mate after the hybridization event, or after the resolution of the first hybridization events [61]. Examples can be found in a number of *S. paradoxus* strains in which Liti et al. [62] found an introgression of a large fraction of chromosome III of *S. cerevisiae*. More complex situations were found: for instance, the presence of several sub-telomeric Y' sequences, a family of repeated DNA sequences of *S. cerevisiae*, were detected in some strains of *Saccharomyces bayanus* var. *bayanus* [17]. It is not known whether several Y' sequences were transferred from *S. cerevisiae* or if a single Y' sequence originated from *S. cerevisiae* that was subsequently duplicated in *S. bayanus* var. *bayanus*. The high variability of chromosomal organization in many hybrid strains of *Saccharomyces pastorianus* is also the result of many rearrangements including chromosome duplication, fusions, etc. subsequent to the original hybridization event(s) [17,63]. It was further shown that some *S. pastorianus* hybrids were the result of hybridizations involving a third species in addition to *S. cerevisiae* and *S. bayanus* var. *bayanus*; the third contributor to these hybrids remains to be isolated [64].

Recent work has shown that genetic diversity might be generated differently according to the yeast clade considered. Whereas classical sexuality maybe the rule in the species of the *Saccharomyces* complex, it may be otherwise in the clades like the CTG clade [65], a monophyletic group of yeast species, which share a deviation of the universal genetic code in which the CUG codon is read as Serine instead of Leucine. The much studied *Candida albicans*, which is diploid heterozygote, was recently shown to undergo loss of heterozygosity (LOH), leaving large regions

of chromosomes or even entire chromosomes homozygote (see [66] and references therein). By studying crosses in *Candida lusitaniae*, recombinant and aneuploid progeny was obtained that may expand genetic diversity [65]. By applying a "gene genealogies" approach, which is used to evidence sexuality among cryptic fungal species [67] and by analyzing informative genomic markers, it was shown that *Debaryomyces hansenii*, the biotechnological species of the CTG clade, was in fact a complex made of cryptic species. Some of these species were partly made of diploid heterozygotes, which like *C. albicans*, undergo LOH [59,68]. The presence of cryptic species that form hybrids was observed in another species *Millerozyma (Pichia) farinosa* (Mallet et al., in preparation). One of the species belonging to the *M. farinosa* complex, the well known *Pichia sorbitophila*, was shown to be a diploid heterozygote that also underwent LOH (The Genolévures consortium, personal communication), indicating that this may be common to many, if not all, CTG clade species. The combined existence of hybrids and associated LOH can explain some of the difficulties encountered when reconstructing phylogenies in this part of the yeast tree. In most cases, the ploidy and heterozygote status of the appraised strains was not taking into account in previous phylogenetic studies [69,70], which led to discrepancies between the resulting trees. Indeed, in our experience, diploid strains were shown to contain markers belonging to different species that were redistributed following LOH (Mallet et al., in preparation). As a result, a phylogenetic analysis with multi-species markers led inevitably to erroneous trees.

Overall, the combination of numerous diploid heterozygotes, LOH and clonality will need to be considered in the future phylogenetic studies in specific clades like that of the CTG. The mating process does not seem well conserved for many heterothallic species, thus leading to mating between closely related species. Some of the progeny from these matings survive leading to an abundance of interspecific forms.

5. Bar coding with a unique molecular marker: the Graal in taxonomy

Whatever the type of organism studied, a unique universal marker is desirable. Indeed, although very efficient and reliable, identification and classification of yeast strains through the amplification and sequencing of several markers is burdensome. The attempts to adapt techniques devised for bacteria, Fourier-Transform Infrared Microspectroscopy (see [71] for review) and MALDI-TOF mass spectrometry [72], to yeasts and fungi show promise, and may solve the problem of rapid identification to diagnose infections due to fungi. Nevertheless, these methods cannot cater for (1) new taxa analysis, since by definition, the new species which is about to be described cannot be represented in databases and (2) phylogeny. One of the goals of modern taxonomy is to find a single easily PCR-amplifiable marker that is relatively short, to allow a single run of Sanger sequencing or pyrosequencing, and informative, to provide a clear distinction between all species. The preferred yeast rDNA D1/D2 marker cannot fulfill this role, since its reduced variability does not allow for differentiation of a number of taxa (see above). Attention was given to a similar type of moderately repeated marker in eukaryotes, the mitochondrial *COX1* gene, in order to barcode biodiversity [73]. Like other markers, it proved to be useful [74], although problems associated with (1) the nature of mtDNA itself considering its peculiar inheritance and its mode of evolution, and (2) interspecific hybridizations, were observed (for review, see [75,76]). More practical considerations arose with the use of the *COX1* gene, such as the variable location of introns within the gene of interest [77].

Fungal taxonomists turned towards the Internal Transcribed Sequence 1 and 2 separated by the slow evolving 5.8S gene (ITS), which by its nature is much more discriminating than the D1D2 part of rDNA. This proved to be useful in hemiascomycetes, although some exceptions were observed in basidiomycetous yeasts [78]. Indeed, the lack of strong selection pressure on the two non-coding regions is such that, although sufficiently variable to allow for barcoding, it is subjected to many indels leading to extraneous size variability, making it unsuitable for phylogeny in hemiascomycetes. Attempts have been made to use this marker as a barcode (G. Verkeij, personal communication). Our experience is that one could take advantage of its important size variation to strengthen species delineation (Weiss et al., in preparation). Nevertheless, in *Debaryomyces*, the ITS region is unable to differentiate between the cryptic species related to *D. hansenii* (our unpublished data), whereas spliceosomal introns of various housekeeping genes or coding sequence for actin can do this.

Again, comparative genomics could help in this matter and a bioinformatics search for genes that could perform as well as the large numbers of markers used to construct species phylogenies was undertaken. A first study based on 33 genomes and the comparison of distance matrixes of each KOG and that of the concatenated KOGs led to a number of single gene candidates [51]. A similar study based on the exhaustive comparison of the topologies of the phylogenies between orthologs from 30 genomes and single phylogeny topologies led to the selection of over 200 candidates [54]. Interestingly, these genes were shown to perform well, i.e. the phylogeny of these genes is very similar or identical to the phylogeny of the species established with 246 genes, independently of the set of species to be analyzed. The commonly used markers such as *ACT1*, *RPB2*, etc. did not perform well in this study. The first attempts at using the best of these selected markers, *TSR1* and *MCM7*, are promising [79], but more genomes are needed to ascertain these candidates as "high phylogenic performers". A major drawback of this approach is that the amplification of many of these genes is highly problematic because they are not well conserved. One may imagine that the constant accumulation of sequence data on these genes from many species can permit the design of a number of nucleotide primers, which could be efficient in PCR amplification when used as a mixture.

6. The future

A number of studies have highlighted the two key problems in reconstructing phylogeny: (1) the difficulty to

find common markers that are informative enough for species distinction, (2) the difficulty of assessing whether a marker or a combination of markers can reflect the evolution of hemiascomycetous yeasts. Additional questions like the minimal data set necessary for molecular definition of a species are also relevant, since most of the newly described species have only "passed" the "D1D2 test". It is also clear that the relevance of the use of ribosomal DNA as the source of unique markers is questionable. Finally, future work will attempt at harmonizing the combination of used markers (compare [69,70]).

What could change and/or improve taxonomy of hemiascomycetous yeasts in the future?

1. Population genomics is clearly the most informative approach to determine phylogenetic relationship between species as shown by Liti et al. [80] and Schacherer et al. [81]. Although the price of sequencing will continue to go down and this approach will be without doubt applied to major pathogens like *Candida* and the basidiomycetes *Cryptoccocus*, and to important biotechnological isolates, it is very unlikely that many clades will be analyzed in such a fashion.
2. Already started in bacteria, a large project aiming at sequencing all the existing type strains will certainly be undertaken for yeasts. "Dikaryome", an international effort aimed at sequencing a large number of hemiascomycetous and basidiomycetous yeasts has recently been initiated. Such large-scale genome sequencing will solve most of the problems of taxonomy by providing a wide phylogeny of yeasts and reference genomes.
3. Yeast taxonomy has always been hampered by a certain self-consciousness at defining new species and by the lack of curiosity of exploring entire taxa, in contrast to only comparing type strains. Our recent work [59,68] has shown that at least in the CTG clade, the analysis of a large number of strains within a species could reveal cryptic species that were previously ignored, as well as unexpectedly larger biodiversity due to genetic exchanges between divergent strains and species at high frequency.
4. The need for integrated up-to-date databases is important, since the search through the large generalist sequence databases such as NCBI and EBI may prove disappointing; in these databases individual strains and ecological samples are over represented, largely diluting the type strains or representative strains. The trend toward unified taxonomy has led to many initiatives that better facilitate non-specialist needs such as Mycobank (http://www.mycobank.org/). Staininfo (http://www.staininfo.net) is one of the most innovative tools created recently. Such integrated databases may allow, through an ingenious updating system, the search of 13 international collections. In our view, an integrated database would gather databases on (1) taxonomical nomenclature like Mycobank, (2) genome sequence resources as that provided by Genolevures (http://www.genolevures.org) or the *Candida* database (http://www.broadinstitute.org), and (3) taxonomical marker sequence resources associated to easy-to-use tools; this taxonomy-dedicated marker database remains to be built. This would overcome the need to search for scattered information in independent, not always updated, databases. It must be stressed that more manual annotation (or less automatized annotation) is crucial for its success. Such a database would also associate with Biological Resource Centers constituted in networks like the one that the ongoing European program EMbaRC (http://www.embarc.eu) is currently building.

High-throughput sequencing has changed many aspects of biology, taxonomy not the least. It needs to be applied more systematically to new species as well as to previously discovered species. High-throughput sequencing of genomes and of specific markers may not have provided an immediate solution to taxonomical problems, but it surely has raised more questions, like the minimal data set necessary to characterize a species. No doubt, new generation sequencing will bring a number of surprises regarding the evolution of yeasts, and it can be foreseen that a robust taxonomy will be generated in the near future.

Disclosure of interest

The authors declare that they have no conflicts of interest concerning this article.

Acknowledgements

This work has received funding from the European Community's Seventh Framework Programme (FP7, 2007–2013), Research Infrastructures action, under the grant agreement No. FP7-228310 (EMbaRC project). S.W. is a post-doctoral fellow in the EMbaRC project. G.M. is a PhD student supported by the CNIEL. This work was financially supported by INRA. The authors would like to thank the anonymous referees for their help in improving the manuscript. The authors are grateful to Vidya Rajan for reading the manuscript.

References

[1] B. Dujon, D. Sherman, G. Fischer, P. Durrens, S. Casaregola, I. Lafontaine, J. De Montigny, C. Marck, C. Neuveglise, E. Talla, N. Goffard, L. Frangeul, M. Aigle, V. Anthouard, A. Babour, V. Barbe, S. Barnay, S. Blanchin, J.M. Beckerich, E. Beyne, C. Bleykasten, A. Boisrame, J. Boyer, L. Cattolico, F. Confanioleri, A. De Daruvar, L. Despons, E. Fabre, C. Fairhead, H. Ferry-Dumazet, A. Groppi, F. Hantraye, C. Hennequin, N. Jauniaux, P. Joyet, R. Kachouri, A. Kerrest, R. Koszul, M. Lemaire, I. Lesur, L. Ma, H. Muller, J.M. Nicaud, M. Nikolski, S. Oztas, O. Ozier-Kalogeropoulos, S. Pellenz, S. Potier, G.F. Richard, M.L. Straub, A. Suleau, D. Swennen, F. Tekaia, M. Wesolowski-Louvel, E. Westhof, B. Wirth, M. Zeniou-Meyer, I. Zivanovic, M. Bolotin-Fukuhara, A. Thierry, C. Bouchier, B. Caudron, C. Scarpelli, C. Gaillardin, J. Weissenbach, P. Wincker, J.L. Souciet, Genome evolution in yeasts, Nature 430 (2004) 35–44.
[2] C.P. Kurtzman, J.W. Fell (Eds.), The yeasts, a taxonomic study, fourth edition, Elsevier, Amsterdam, 1998.
[3] C.P. Kurtzman, J.W. Fell, T. Boekhout (Eds.), The yeasts, fourth edition: a taxonomic study, Elsevier, Amsterdam, 2011.
[4] S.O. Suh, J.V. McHugh, D.D. Pollock, M. Blackwell, The beetle gut: a hyperdiverse source of novel yeasts, Mycol. Res. 109 (2005) 261–265.
[5] T. Boekhout, Biodiversity: gut feeling for yeasts, Nature 434 (2005) 449–451.
[6] D.L. Hawksworth, The fungal dimension of biodiversity: magnitude, significance, and conservation, Mycol. Res. 95 (1991) 641–655.

[7] A. Vaughan-Martini, C.P. Kurtzman, Deoxyribonucleic acid relatedness among species of the genus *Saccharomyces sensu stricto*, Int. J. Syst. Bacteriol. 35 (1985) 508–511.

[8] T.J. White, T.D. Bruns, S. Lee, J.W. Taylor, Amplication and direct sequencing of fungal ribosomal RNA genes for phylogenetics, in: M.A. Innis, D.H. Gelfand, J. Sninsky, T.J. White (Eds.), PCR Protocols: a guide to methods and applications, Academic Press, San Diego, 1990, pp. 315–322.

[9] A. Wilmotte, Y. Van de Peer, A. Goris, S. Chapell, R. De Baere, B. Nelissen, J.M. Neefs, H.G. L, R. De Watcher, Evolutionnary relationships among higher fungi inferred from small ribosomal subunit RNA sequence analysis, Syst. Appl. Microbiol. 16 (1993) 436–444.

[10] J. Guadet, J. Julien, J.F. Lafay, Y. Brygoo, Phylogeny of some Fusarium species, as determined by large-subunit rRNA sequence comparison, Mol. Biol. Evol. 6 (1989) 227–242.

[11] C.P. Kurtzman, rRNA sequence comparisons for assessing phylogenetic relationships among yeasts, Int. J. Syst. Bacteriol. 42 (1992) 1–6.

[12] K. O'Donnell, Fusarium and its relatives, in: D.R. Reynolds, J.W. Taylor (Eds.), The fungal holomorph: mitotc, meiotic and pleomorphic speciation in fungal systematics, CAB International, Wallingford, UK, 1993, pp. 225–233.

[13] T. Boekhout, C.P. Kurtzman, K. O'Donnell, M.T. Smith, Phylogeny of the yeast genera *Hanseniaspora* (anamorph *Kloeckera*), *Dekkera* (anamorph *Brettanomyces*), and *Eeniella* as inferred from partial 26S ribosomal DNA nucleotide sequences, Int. J. Syst. Bacteriol. 44 (1994) 781–786.

[14] C.P. Kurtzman, C.J. Robnett, Identification of clinically important ascomycetous yeasts based on nucleotide divergence in the 5' end of the large-subunit (26S) ribosomal DNA gene, J. Clin. Microbiol. 35 (1997) 1216–1223.

[15] C.P. Kurtzman, C.J. Robnett, Identification and phylogeny of ascomycetous yeasts from analysis of nuclear large subunit (26S) ribosomal DNA partial sequences, Antonie Van Leeuwenhoek 73 (1998) 331–371.

[16] B. Esteve-Zarzoso, C. Belloch, F. Uruburu, A. Querol, Identification of yeasts by RFLP analysis of the 5.8S rRNA gene and the two ribosomal internal transcribed spacers, Int. J. Syst. Bacteriol. 49 (Pt 1) (1999) 329–337.

[17] H.V. Nguyen, A. Lepingle, C.A. Gaillardin, Molecular typing demonstrates homogeneity of *Saccharomyces uvarum* strains and reveals the existence of hybrids between *S. uvarum* and *S. cerevisiae*, including the *S. bayanus* type strain CBS 380, Syst. Appl. Microbiol. 23 (2000) 71–85.

[18] D.M. Geiser, J.I. Pitt, J.W. Taylor, Cryptic speciation and recombination in the aflatoxin-producing fungus *Aspergillus flavus*, Proc. Natl. Acad. Sci. U. S. A. 95 (1998) 388–393.

[19] K. O'Donnell, H.C. Kistler, B.K. Tacke, H.H. Casper, Gene genealogies reveal global phylogeographic structure and reproductive isolation among lineages of *Fusarium graminearum*, the fungus causing wheat scab, Proc. Natl. Acad. Sci. U. S. A. 97 (2000) 7905–7910.

[20] Y.J. Liu, S. Whelen, B.D. Hall, Phylogenetic relationships among ascomycetes: evidence from an RNA polymerse II subunit, Mol. Biol. Evol. 16 (1999) 1799–1808.

[21] H.M. Daniel, T.C. Sorrell, W. Meyer, Partial sequence analysis of the actin gene and its potential for studying the phylogeny of *Candida* species and their teleomorphs, Int. J. Syst. Evol. Microbiol. 51 (2001) 1593–1606.

[22] H.M. Daniel, W. Meyer, Evaluation of ribosomal RNA and actin gene sequences for the identification of ascomycetous yeasts, Int. J. Food Microbiol. 86 (2003) 61–78.

[23] A. Goffeau, B.G. Barrell, H. Bussey, R.W. Davis, B. Dujon, H. Feldmann, F. Galibert, J.D. Hoheisel, C. Jacq, M. Johnston, E.J. Louis, H.W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, S.G. Oliver, Life with 6000 genes, Science 274 (1996) 546–563.

[24] J. Souciet, M. Aigle, F. Artiguenave, G. Blandin, M. Bolotin-Fukuhara, E. Bon, P. Brottier, S. Casaregola, J. de Montigny, B. Dujon, P. Durrens, C. Gaillardin, A. Lepingle, B. Llorente, A. Malpertuy, C. Neuveglise, O. Ozier-Kalogeropoulos, S. Potier, W. Saurin, F. Tekaia, C. Toffano-Nioche, M. Wesolowski-Louvel, P. Wincker, J. Weissenbach, Genomic exploration of the hemiascomycetous yeasts: 1. A set of yeast species for molecular evolution studies, FEBS Lett. 487 (2000) 3–12.

[25] M. Kellis, N. Patterson, M. Endrizzi, B. Birren, E.S. Lander, Sequencing and comparison of yeast species to identify genes and regulatory elements, Nature 423 (2003) 241–254.

[26] P. Cliften, P. Sudarsanam, A. Desikan, L. Fulton, B. Fulton, J. Majors, R. Waterston, B.A. Cohen, M. Johnston, Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting, Science 301 (2003) 71–76.

[27] M. Kellis, B.W. Birren, E.S. Lander, Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*, Nature 428 (2004) 617–624.

[28] F.S. Dietrich, S. Voegeli, S. Brachat, A. Lerch, K. Gates, S. Steiner, C. Mohr, R. Pohlmann, P. Luedi, S. Choi, R.A. Wing, A. Flavier, T.D. Gaffney, P. Philippsen, The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome, Science 304 (2004) 304–307.

[29] T.W. Jeffries, I.V. Grigoriev, J. Grimwood, J.M. Laplaza, A. Aerts, A. Salamov, J. Schmutz, E. Lindquist, P. Dehal, H. Shapiro, Y.S. Jin, V. Passoth, P.M. Richardson, Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast *Pichia stipitis*, Nat. Biotechnol. 25 (2007) 319–326.

[30] K. De Schutter, Y.C. Lin, P. Tiels, A. Van Hecke, S. Glinka, J. Weber-Lehmann, P. Rouze, Y. Van de Peer, N. Callewaert, Genome sequence of the recombinant protein production host *Pichia pastoris*, Nat. Biotechnol. 27 (2009) 561–566.

[31] D. Mattanovich, A. Graf, J. Stadlmann, M. Dragosits, A. Redl, M. Maurer, M. Kleinheinz, M. Sauer, F. Altmann, B. Gasser, Genome, secretome and glucose transport highlight unique features of the protein production host *Pichia pastoris*, Microb. Cell Fact. 8 (2009) 29.

[32] M. Woolfit, E. Rozpedowska, J. Piskur, K.H. Wolfe, Genome survey sequencing of the wine spoilage yeast *Dekkera* (*Brettanomyces*) *bruxellensis*, Eukaryot. Cell 6 (2007) 721–733.

[33] A.P. Jackson, J.A. Gamble, T. Yeomans, G.P. Moran, D. Saunders, D. Harris, M. Aslett, J.F. Barrell, G. Butler, F. Citiulo, D.C. Coleman, P.W. de Groot, T.J. Goodwin, M.A. Quail, J. McQuillan, C.A. Munro, A. Pain, R.T. Poulter, M.A. Rajandream, H. Renauld, M.J. Spiering, A. Tivey, N.A. Gow, B. Barrell, D.J. Sullivan, M. Berriman, Comparative genomics of the fungal pathogens *Candida dubliniensis* and *Candida albicans*, Genome Res. 19 (2009) 2231–2244.

[34] G. Butler, M.D. Rasmussen, M.F. Lin, M.A. Santos, S. Sakthikumar, C.A. Munro, E. Rheinbay, M. Grabherr, A. Forche, J.L. Reedy, I. Agrafioti, M.B. Arnaud, S. Bates, A.J. Brown, S. Brunke, M.C. Costanzo, D.A. Fitzpatrick, P.W. de Groot, D. Harris, L.L. Hoyer, B. Hube, F.M. Klis, C. Kodira, N. Lennard, M.E. Logue, R. Martin, A.M. Neiman, E. Nikolaou, M.A. Quail, J. Quinn, M.C. Santos, F.F. Schmitzberger, G. Sherlock, P. Shah, K.A. Silverstein, M.S. Skrzypek, D. Soll, R. Staggs, I. Stansfield, M.P. Stumpf, P.E. Sudbery, T. Srikantha, Q. Zeng, J. Berman, M. Berriman, J. Heitman, N.A. Gow, M.C. Lorenz, B.W. Birren, M. Kellis, C.A. Cuomo, Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes, Nature 459 (2009) 657–662.

[35] C.P. Kurtzman, C.J. Robnett, Phylogenetic relationships among yeasts of the '*Saccharomyces* complex' determined from multigene sequence analyses, FEMS Yeast Res. 3 (2003) 417–432.

[36] C.P. Kurtzman, Phylogenetic circumscription of *Saccharomyces*, *Kluyveromyces* and other members of the *Saccharomycetaceae*, and the proposal of the new genera *Lachancea*, *Nakaseomyces*, *Naumovia*, *Vanderwaltozyma* and *Zygotorulaspora*, FEMS Yeast Res. 4 (2003) 233–245.

[37] C.P. Kurtzman, C.J. Robnett, Systematics of methanol assimilating yeasts and neighboring taxa from multigene sequence analysis and the proposal of *Peterozyma* gen. nov., a new member of the Saccharomycetales, FEMS Yeast Res. 10 (2010) 353–361.

[38] C.P. Kurtzman, C.J. Robnett, E. Basehoar-Powers, Phylogenetic relationships among species of *Pichia*, *Issatchenkia* and *Williopsis* determined from multigene sequence analysis, and the proposal of *Barnettozyma* gen. nov., *Lindnera* gen. nov. and *Wickerhamomyces* gen. nov, FEMS Yeast Res. 8 (2008) 939–954.

[39] C.P. Kurtzman, J. Albertyn, E. Basehoar-Powers, Multigene phylogenetic analysis of the *Lipomycetaceae* and the proposed transfer of *Zygozyma* species to *Lipomyces* and *Babjevia* anomala to *Dipodascopsis*, FEMS Yeast Res. 7 (2007) 1027–1034.

[40] C.P. Kurtzman, C.J. Robnett, Multigene phylogenetic analysis of the *Trichomonascus*, *Wickerhamiella* and *Zygoascus* yeast clades, and the proposal of *Sugiyamaella* gen. nov. and 14 new species combinations, FEMS Yeast Res. 7 (2007) 141–151.

[41] C.P. Kurtzman, New species and new combinations in the yeast genera *Kregervanrija* gen. nov., *Saturnispora* and *Candida*, FEMS Yeast Res. 6 (2006) 288–297.

[42] C.P. Kurtzman, C.J. Robnett, J.M. Ward, C. Brayton, P. Gorelick, T.J. Walsh, Multigene phylogenetic analysis of pathogenic candida species in the *Kazachstania* (*Arxiozyma*) *telluris* complex and description of their ascosporic states as *Kazachstania bovina* sp. nov., *K. heterogenica* sp. nov., *K. pintolopesii* sp. nov., and *K. slooffiae* sp. nov., J. Clin. Microbiol. 43 (2005) 101–111.

[43] C.P. Kurtzman, Phylogeny of the ascomycetous yeasts and the renaming of *Pichia anomala* to *Wickerhamomyces anomalus*, Antonie Van Leeuwenhoek 99 (2010) 13–23.

[44] D.A. Fitzpatrick, M.E. Logue, J.E. Stajich, G. Butler, A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis, BMC Evol. Biol. 6 (2006) 99.

[45] E.E. Kuramae, V. Robert, B. Snel, M. Weiss, T. Boekhout, Phylogenomics reveal a robust fungal tree of life, FEMS Yeast Res. 6 (2006) 1213–1220.

[46] B. Robbertse, J.B. Reeves, C.L. Schoch, J.W. Spatafora, A phylogenomic analysis of the *Ascomycota*, Fungal Genet. Biol. 43 (2006) 715–725.

[47] K.H. Wolfe, D.C. Shields, Molecular evidence for an ancient duplication of the entire yeast genome, Nature 387 (1997) 708–713.
[48] J.L. Souciet, B. Dujon, C. Gaillardin, M. Johnston, P.V. Baret, P. Cliften, D.J. Sherman, J. Weissenbach, E. Westhof, P. Wincker, C. Jubin, J. Poulain, V. Barbe, B. Segurens, F. Artiguenave, V. Anthouard, B. Vacherie, M.E. Val, R.S. Fulton, P. Minx, R. Wilson, P. Durrens, G. Jean, C. Marck, T. Martin, M. Nikolski, T. Rolland, M.L. Seret, S. Casaregola, L. Despons, C. Fairhead, G. Fischer, I. Lafontaine, V. Leh, M. Lemaire, J. de Montigny, C. Neuveglise, A. Thierry, I. Blanc-Lenfle, C. Bleykasten, J. Diffels, E. Fritsch, L. Frangeul, A. Goeffon, N. Jauniaux, R. Kachouri-Lafond, C. Payen, S. Potier, L. Pribylova, C. Ozanne, G.F. Richard, C. Sacerdot, M.L. Straub, E. Talla, Comparative genomics of protoploid *Saccharomycetaceae*, Genome Res. 19 (2009) 1696–1709.
[49] A. Rokas, S.B. Carroll, More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy, Mol. Biol. Evol. 22 (2005) 1337–1344.
[50] H. Wang, Z. Xu, L. Gao, B. Hao, A fungal phylogeny based on 82 complete genomes using the composition vector method, BMC Evol. Biol. 9 (2009) 195.
[51] E.E. Kuramae, V. Robert, C. Echavarri-Erasun, T. Boekhout, Cophenetic correlation analysis as a strategy to select phylogenetically informative proteins: an example from the fungal kingdom, BMC Evol. Biol. 7 (2007) 134.
[52] B.E. Dutilh, V. van Noort, R.T.J.M. van der Heijden, T. Boekhout, B. Snel, M.A. Huynen, Assessment of phylogenomic and orthology approaches for phylogenetic inference, Bioinformatics 23 (2007) 815–824.
[53] S.M. Hedtke, T.M. Townsend, D.M. Hillis, Resolution of phylogenetic conflict in large data sets by increased taxon sampling, Syst. Biol. 55 (2006) 522–529.
[54] G. Aguileta, S. Marthey, H. Chiapello, M.H. Lebrun, F. Rodolphe, E. Fournier, A. Gendrault-Jacquemard, T. Giraud, Assessing the performance of single-copy genes for recovering robust phylogenies, Syst. Biol. 57 (2008) 613–627.
[55] M.C. Kielland-Brandt, T. Nilsson-Tillgren, C. Gjermansen, S. Holmberg, M.B. Pedersen, Genetics of brewing yeasts, in: A.H. Rose, E. Wheals, J.S. Harrison (Eds.), The yeasts, Academic Press, London, 1995, pp. 223–254.
[56] S.A. James, C.J. Bond, M. Stratford, I.N. Roberts, Molecular evidence for the existence of natural hybrids in the genus *Zygosaccharomyces*, FEMS Yeast Res. 5 (2005) 747–755.
[57] L. Solieri, S. Cassanelli, M.A. Croce, P. Giudici, Genome size and ploidy level: new insights for elucidating relationships in *Zygosaccharomyces* species, Fungal Genet. Biol. 45 (2008) 1582–1590.
[58] J.L. Gordon, K.H. Wolfe, Recent allopolyploid origin of *Zygosaccharomyces rouxii* strain ATCC 42981, Yeast 25 (2008) 449–456.
[59] N. Jacques, S. Mallet, S. Casaregola, Delimitation of the species of the *Debaryomyces hansenii* complex by intron sequence analysis, Int. J. Syst. Evol. Microbiol. 59 (2009) 1242–1251.
[60] M. Bovers, F. Hagen, T. Boekhout, Diversity of the *Cryptococcus neoformans–Cryptococcus gattii* species complex, Rev. Iberoam Micol. 25 (2008) S4–12.
[61] M. Sipiczki, Interspecies hybridization and recombination in *Saccharomyces* wine yeasts, FEMS Yeast Res. 8 (2008) 996–1007.
[62] G. Liti, D.B. Barton, E.J. Louis, Sequence diversity, reproductive isolation and species concepts in *Saccharomyces*, Genetics 174 (2006) 839–850.
[63] S. Casaregola, H.V. Nguyen, G. Lapathitis, A. Kotyk, C. Gaillardin, Analysis of the constitution of the beer yeast genome by PCR, sequencing and subtelomeric sequence hybridization, Int. J. Syst. Evol. Microbiol. 51 (2001) 1607–1618.
[64] S. Rainieri, Y. Kodama, Y. Kaneko, K. Mikata, Y. Nakao, T. Ashikari, Pure and mixed genetic lines of *Saccharomyces bayanus* and *Saccharomyces pastorianus* and their contribution to the lager brewing strain genome, Appl. Environ. Microbiol. 72 (2006) 3968–3974.
[65] J.L. Reedy, A.M. Floyd, J. Heitman, Mechanistic plasticity of sexual reproduction and meiosis in the *Candida* pathogenic species complex, Curr. Biol. 19 (2009) 891–899.
[66] D. Diogo, C. Bouchier, C. d'Enfert, M.E. Bougnoux, Loss of heterozygosity in commensal isolates of the asexual diploid yeast *Candida albicans*, Fungal Genet. Biol. 46 (2009) 159–168.
[67] J.W. Taylor, D.J. Jacobson, S. Kroken, T. Kasuga, D.M. Geiser, D.S. Hibbett, M.C. Fisher, Phylogenetic species recognition and species concepts in fungi, Fungal Genet. Biol. 31 (2000) 21–32.
[68] N. Jacques, C. Sacerdot, M. Derkaoui, B. Dujon, O. Ozier-Kalogeropoulos, S. Casaregola, Population polymorphism of nuclear mitochondrial DNA insertions reveals widespread diploidy associated with loss of heterozygosity in *Debaryomyces hansenii*, Eukaryot. Cell 9 (2010) 449–459.
[69] C.K. Tsui, H.M. Daniel, V. Robert, W. Meyer, Re-examining the phylogeny of clinically relevant *Candida* species and allied genera based on multigene analyses, FEMS Yeast Res. 8 (2008) 651–659.
[70] C.P. Kurtzman, M. Suzuki, Phylogenetic analysis of ascomycete yeasts that form coenzyme Q-9 and the proposal of the new genera *Babjeviella*, *Meyerozyma*, *Millerozyma*, *Priceomyces* and *Scheffersomyces*, Mycoscience 51 (2010).
[71] C. Santos, M.E. Fraga, Z. Kozakiewicz, N. Lima, Fourier transform infrared as a powerful technique for the identification and characterization of filamentous fungi and yeasts, Res. Microbiol. 161 (2010) 168–175.
[72] F. Seyfarth, M. Ziemer, H.G. Sayer, A. Burmester, M. Erhard, M. Welker, S. Schliemann, E. Straube, U.C. Hipler, The use of ITS DNA sequence analysis and MALDI-TOF mass spectrometry in diagnosing an infection with *Fusarium proliferatum*, Exp. Dermatol. 17 (2008) 965–971.
[73] P.D. Hebert, A. Cywinska, S.L. Ball, J.R. deWaard, Biological identifications through DNA barcodes, Proc. Biol. Sci. 270 (2003) 313–321.
[74] P.D. Hebert, E.H. Penton, J.M. Burns, D.H. Janzen, W. Hallwachs, Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 14812–14817.
[75] K.L. Shaw, Conflict between nuclear and mitochondrial DNA phylogenies of a recent species radiation: what mtDNA reveals and conceals about modes of speciation in Hawaiian crickets, Proc. Natl. Acad. Sci. U. S. A. 99 (2002) 16122–16127.
[76] G.D. Hurst, F.M. Jiggins, Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts, Proc. Biol. Sci. 272 (2005) 1525–1534.
[77] K.A. Seifert, Barcoding fungi: progress towards DNA barcoding of fungi, Mol. Ecol. Resour. 9 (2009) 83–89.
[78] G. Scorzetti, J.W. Fell, A. Fonseca, A. Statzell-Tallman, Systematics of basidiomycetous yeasts: a comparison of large subunit D1/D2 and internal transcribed spacer rDNA regions, FEMS Yeast Res. 2 (2002) 495–517.
[79] I. Schmitt, A. Crespo, P.K. Divakar, J.D. Fankhauser, E. Herman-Sackett, K. Kalb, M.P. Nelsen, N.A. Nelson, E. Rivas-Plata, A.D. Shimp, T. Widhelm, H.T. Lumbsch, New primers for promising single-copy genes in fungal phylogenetics and systematics, Persoonia 23 (2009) 35–40.
[80] G. Liti, D.M. Carter, A.M. Moses, J. Warringer, L. Parts, S.A. James, R.P. Davey, I.N. Roberts, A. Burt, V. Koufopanou, I.J. Tsai, C.M. Bergman, D. Bensasson, M.J. O'Kelly, A. van Oudenaarden, D.B. Barton, E. Bailes, A.N. Nguyen, M. Jones, M.A. Quail, I. Goodhead, S. Sims, F. Smith, A. Blomberg, R. Durbin, E.J. Louis, Population genomics of domestic and wild yeasts, Nature 458 (2009) 337–341.
[81] J. Schacherer, J.A. Shapiro, D.M. Ruderfer, L. Kruglyak, Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*, Nature 458 (2009) 342–345.

ANNEXE 2

INVESTIGATION

Pichia sorbitophila, an Interspecies Yeast Hybrid, Reveals Early Steps of Genome Resolution After Polyploidization

Véronique Leh Louis,*,[1] Laurence Despons,* Anne Friedrich,* Tiphaine Martin,[†] Pascal Durrens,[†] Serge Casarégola,[‡] Cécile Neuvéglise,[‡] Cécile Fairhead,[§] Christian Marck,[**] José A. Cruz,[††] Marie-Laure Straub,* Valérie Kugler,* Christine Sacerdot,[‡‡] Zlatyo Uzunov,[§§] Agnes Thierry,[‡‡] Stéphanie Weiss,[‡] Claudine Bleykasten,* Jacky De Montigny,* Noemie Jacques,[‡] Paul Jung,* Marc Lemaire,[***] Sandrine Mallet,[‡] Guillaume Morel,[‡] Guy-Franck Richard,[‡‡] Anasua Sarkar,[†] Guilhem Savel,[†††] Joseph Schacherer,* Marie-Line Seret,[‡‡‡] Emmanuel Talla,[§§§] Gaelle Samson,[****] Claire Jubin,[****] Julie Poulain,[****] Benoît Vacherie,[****] Valérie Barbe,[****] Eric Pelletier,[****] David J. Sherman,[†] Eric Westhof,[††] Jean Weissenbach,[****] Philippe V. Baret,[‡‡‡] Patrick Wincker,[****] Claude Gaillardin,[‡] Bernard Dujon,[‡‡] and Jean-Luc Souciet*,[1]

*Université de Strasbourg, CNRS UMR7156, F-67000 Strasbourg, France; [†]Université de Bordeaux 1, LaBRI INRIA Bordeaux Sud-Ouest (MAGNOME), F-33405 Talence, France; [‡]INRA UMR 1319 Micalis, AgroParisTech, Bat. CBAI, F-78850 Thiverval-Grignon, France; [§]Institut de Génétique et Microbiologie, Université Paris-Sud, UMR CNRS 8621, F-91405 Orsay CEDEX, France; **Institut de Biologie et de Technologies de Saclay (iBiTec-S), CEA, F-91191 Gif-sur-Yvette CEDEX, France; [††]Université de Strasbourg, Architecture et Réactivité de l'ARN, Institut de Biologie Moléculaire et Cellulaire du CNRS, F-67084 Strasbourg, France; [‡‡]Institut Pasteur, CNRS URA2171, Université Pierre et Maris Curie, Paris 6 UFR927, F-75724, Paris-CEDEX 15, France; [§§]Sofia University St. Kliment Ohridski, Faculty of Biology, Department of General and Applied Microbiology, 1164, Sofia, Bulgaria; ***Université de Lyon, F-69622, Lyon, France; Université Lyon 1, Villeurbanne; CNRS, UMR5240 Microbiologie, Adaptation et Pathogénie; INSA de Lyon, F-69621, Villeurbanne, France; [†††]Université de Bordeaux 1, CNRS UMR5800, F-33405 Talence, France; [‡‡‡]Earth and Life Institute, Université Catholique de Louvain, B-1348, Louvain-la-Neuve, Belgium; [§§§]Université de la Méditerranée, Laboratoire de Chimie Bactérienne, CNRS-UPR9043, 31 chemin Joseph Aiguier, F-13402 Marseille CEDEX 20, France, and ****CEA, DSV, IG, Génoscope; CNRS UMR 8030; Université d'Evry Val d' Essonne, 2 rue Gaston Crémieux, F-91057 Evry, France.

ABSTRACT Polyploidization is an important process in the evolution of eukaryotic genomes, but ensuing molecular mechanisms remain to be clarified. Autopolyploidization or whole-genome duplication events frequently are resolved in resulting lineages by the loss of single genes from most duplicated pairs, causing transient gene dosage imbalance and accelerating speciation through meiotic infertility. Allopolyploidization or formation of interspecies hybrids raises the problem of genetic incompatibility (Bateson-Dobzhansky-Muller effect) and may be resolved by the accumulation of mutational changes in resulting lineages. In this article, we show that an osmotolerant yeast species, *Pichia sorbitophila*, recently isolated in a concentrated sorbitol solution in industry, illustrates this last situation. Its genome is a mosaic of homologous and homeologous chromosomes, or parts thereof, that corresponds to a recently formed hybrid in the process of evolution. The respective parental contributions to this genome were characterized using existing variations in GC content. The genomic changes that occurred during the short period since hybrid formation were identified (*e.g.*, loss of heterozygosity, unilateral loss of rDNA, reciprocal exchange) and distinguished from those undergone by the two parental genomes after separation from their common ancestor (*i.e.*, NUMT (NUclear sequences of MiTochondrial origin) insertions, gene acquisitions, gene location movements, reciprocal translocation). We found that the physiological characteristics of this new yeast species are determined by specific but unequal contributions of its two parents, one of which could be identified as very closely related to an extant *Pichia farinosa* strain.

KEYWORDS
osmotolerant yeast *P. sorbitophila*
allopolyploidy
hybridization
genome evolution
loss of heterozygosity

doi: 10.1534/g3.111.000745
Manuscript received July 21, 2011; accepted for publication December 16, 2011

Supporting information is available online at http://www.g3journal.org/lookup/suppl/doi:10.1534/g3.111.000745/-/DC1
P. sorbitophila genome sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. FO082046–FO082059.
[1]Corresponding author: Université de Strasbourg, Génétique Moléculaire, Génomique, Microbiologie CNRS UMR7156, Institut de Botanique, 28 rue Goethe, F-67000 Strasbourg, France. E-mail: jlsouciet@unistra.fr; vleh@unistra.fr

A new species could arise via a process of interspecific hybridization, that is, the union of different organisms across a species barrier. At the early stage, hybridization produces allodiploid or allopolyploid hybrids, depending on the ploidy of parents. This situation gives rise to a period of genome resolution where diverse genetic events act simultaneously and successively to form a stable chimerical genome. Recombination, loss of chromosomes and loss of heterozygosity (LOH) are frequent mechanisms involved in this genome shuffling that stabilizes hybrids (Belloch *et al.* 2009; Forche *et al.* 2005; Gerstein *et al.* 2006; Scannell *et al.* 2006; Sipiczki 2008). Because of the combination and modification of at least two distinct gene pools, hybrids

could display better adaptive properties than their parental species, called heterosis (Arnold and Martin 2010; Belloch *et al.* 2008). The limitation on the amount of such interspecies hybrids is often related to genetic incompatibilities, leading to fitness and survival decrease and/or fertility reduction (Johnson 2008). Hybridization occurs approximately in 25% of plant species and 10% of animal species (Mallet 2005).

Recent genomic data and experimental analyses suggest that interspecies hybrids are frequent in yeast populations. Several interspecific hybrids have been described among species that belong to the genus *Saccharomyces* and play key roles in industrial fermentations (Bond *et al.* 2009; Dunn and Sherlock 2008; Nakao *et al.* 2009; Querol and Bond 2009; Rainieri *et al.* 2006; Sipiczki 2008). In contrast, only few cases of interspecific hybrids have been reported among other yeasts or fungi: the fungal pathogens *Candida albicans* and *Candida dubliniensis* can mate to produce tetraploid hybrids (Pujol *et al.* 2004), a random genomic sequencing of a *Zygosaccharomyces rouxii* wild isolate showed that it contains two different sets of genes (Gordon and Wolfe 2008), and finally, among *Basidiomycota*, anomalous *Cryptococcus neoformans* strains isolated from patients appeared to be hybrids between *C. neoformans* and *C. gattii* (Bovers *et al.* 2006). If stress conditions in cultures have been invoked to stimulate the formation of yeast hybrids (Belloch *et al.* 2008; Querol and Bond 2009), experiments show that yeast species in general tend to have no or limited prezygotic barriers (Chou *et al.* 2010; Greig *et al.* 2002; Liti *et al.* 2006; Marinoni and Lachance 2004; Murphy *et al.* 2006). In addition, only rare cases of heterosis or Bateson-Dobzhansky-Muller incompatibility are reported in yeasts (Belloch *et al.* 2008; Chou *et al.* 2010; Kao *et al.* 2010; Lee *et al.* 2008). Therefore, the role of interspecies crosses in yeast populations need to be better clarified, as well as the events involved in the hybrid genome formation and stabilization.

The yeast *P. sorbitophila* (de Miranda *et al.* 1980), a member of the "CTG" group of *Saccharomycotina* (De Montigny *et al.* 2000; Kurtzman and Suzuki 2010), has been largely studied for its resistance to osmotic and salt stress (Banuelos *et al.* 2002; Benito *et al.* 2004; Lages *et al.* 1999; Lages and Lucas 1995; Maresova and Sychrova 2003; Neves *et al.* 2004; Oliveira *et al.* 1996; Prista *et al.* 2005). This osmotolerant yeast was isolated as a contaminant of a 70% sorbitol solution. *P. sorbitophila* resists to very high NaCl concentration (4M NaCl), whereas *Debaryomyces hansenii* and *Pichia farinosa*, two closely related species, are inhibited. de Miranda *et al.* (1980) and Oliveira *et al.* (1996) reported that *P. sorbitophila* is able to form asci with ascospores. The ability of crossing clones from ascospores to each other led to propose that *P. sorbitophila* may be homothallic (Oliveira *et al.* 1996).

By the complete sequencing and the detailed analysis of its genome, we demonstrate here that *P. sorbitophila* is, in fact, a hybrid yeast. From its genome, we extracted the two subgenomes, inherited from both parents. We identified genes for specific metabolism pathways, acquired either from both parents or from a sole parent, giving now the opportunity to study possible cases of heterosis. We finally demonstrate that *P. sorbitophila* is a very recent hybrid in the process of its resolution. The genomic rearrangements characterizing the early steps of genome stabilization were distinguished from those appeared in the parental genomes before the hybridization event. The hybrid state of the *P. sorbitophila* genome raises now the question about the fertility of this yeast.

METHODS

Full methods are available in the online version with the supporting information. Here are described the methods used for the genome sequencing and assembly, and the determination of the global GC content variations between Pγ and Pε subgenomes, depicted in Figure 1.

Sequencing and assembly

The sequencing of the nuclear genome of *Pichia sorbitophila* CBS 7064 (de Miranda *et al.* 1980) was performed using a whole-genome shotgun strategy, with two plasmid libraries (4-kb inserts and 40-kb inserts) and ABI/Sanger technology to 10X-15X depth. In a first read assembly performed using ARACHNE assembler_version 03 (Batzoglou *et al.* 2002), we obtained 17 supercontigs, from 0.144 Mb to 2.114 Mb in size, with a read coverage of 7.5X on average for 11 contigs and 14X for 6 contigs (supporting information, Figure S1). The weakly covered contigs were joined into five distinct pairs showing approximately 85% sequence identity between contigs forming a pair, reflecting the partial heterozygosity of the genome (Figure S3). The homozygous genomic regions were represented by the six contigs 14X covered. Some of the weakly and highly covered contigs could also be joined on the basis of the sequence identity shared at the end of the contigs, corresponding to partly heterozygous and partly homozygous chromosomes (Figure S1). We confirmed this assembly by reiterating all the process using ARACHNE assembler_version 04 and with a manual finishing step to join contigs and resolve low-quality sequences. Homozygous sequences were doubled, and some of them concatenated with heterozygous sequences if necessary, to be representative of the genome state. The presence of small heterozygous regions inside homozygous areas, potentially masked during the assembly process, was checked by analyzing single-nucleotide polymorphism (SNP) distribution (see Table S2 for method). No area with a complex pattern of heterozygosity/homozygosity was highlighted inside the homozygous regions. At the end of that process, 14 contigs were obtained and named in alphabetical order according to their size. Only two sequence gaps remained in chromosomes G and H telomeric repeats. This assembly was validated by comparing with the number and the size of chromosomes estimated by pulsed-field gel electrophoresis (Figure S2), one contig corresponding to one chromosome.

GC content calculation

The GC content along chromosomes was computed using all protein coding genes present in two syntenic copies between chromosomes forming a pair and having identical sizes to minimize the effect of insertions/deletions on the GC values (2499 gene pairs considered). The two GC% obtained for a gene pair were compared with each other by calculating their ratio (dGC) from the mean GC% value: $dGC_1 = GC_1 / ((GC_1 + GC_2)/2)$ and $dGC_2 = GC_2/((GC_1 + GC_2)/2)$, with GC_1 and GC_2 corresponding to the GC% values obtained for copies 1 and 2, respectively. This normalization does not allow us to take into account the GC content variation between genes. Therefore, when both copies of a gene are identical at the nucleotide level (a situation observed in homozygous regions), the following identity relation is attributed to these two alleles: $dGC_1 = dGC_2 = 1$. Conversely, for two nonidentical copies, the gene copy that contains the highest GC content has its $dGC > 1$. The two dGC values obtained for each gene pair were plotted along chromosome pairs using a sliding window of 11 genes with a step of 1 for representing a moving average of 11 adjacent genes along each chromosome. Curve superposition was performed using a single chromosomal coordinate. This coordinate corresponds to the start codon of one gene located on an arbitrarily chosen chromosome of each pair. The GC content variations in all protein coding genes were also calculated using tRNA that pair with two codons (Crick 1966). The codon usage was determined for each heterozygous region, independently from its parental origin, using genes that are present in two syntenic copies (5516 gene pairs

considered). According to the GC trend curves, data obtained for all heterozygous regions belonging to the same subgenome were pooled to determine the average variation observed between both subgenomes.

RESULTS AND DISCUSSION

Structure of the *Pichia sorbitophila* hybrid genome

Size of the genome, number of chromosome pairs, and global heterozygosity: The nuclear genome of *Pichia sorbitophila* CBS 7064 (also referenced as *Pichia* or *Millerozyma farinosa*) (Kurtzman and Suzuki 2010) was completely sequenced using a whole-genome shotgun strategy and initially assembled into 17 supercontigs (see *Materials and Methods* and Figure S1). Sequence alignments revealed nearly perfect synteny conservation and approximately 85% identity at nucleotide level between some pairs of supercontigs (11 supercontigs), suggesting a partial heterozygosity of the genome. The other supercontigs (6 supercontigs) remained unique. Interestingly, sequencing coverage was double for the latter compared with the former (Figure S1), suggesting that they represent homozygous regions with two identical or almost identical copies of sequence and that the read assembly led to the production of a single consensus sequence for each homozygous part. A manually finishing process confirmed that the *P. sorbitophila* genome is a mosaic of homologous and homeologous chromosome pairs (see *Materials and Methods*). We obtained 14 contigs (named chr. A to N), from 1.05 Mb to 2.12 Mb, representing the 14 chromosomes of the *P. sorbitophila* nuclear genome and giving a global genome size of 21.5 Mb (Table S1). Contig sizes largely match chromosome sizes estimated by pulsed-field gel electrophoresis (Figure S2). Telomeric repeats were found at the ends of almost all contigs (Table S1). As suspected, the 14 chromosomes could be gathered into seven pairs of homologous or homeologous chromosomes (Figure 1) on the basis of the sequence identity along the chromosomes (Figure S3). The chromosome pairs A/B and C/D were found to be partly heterozygous and partly homozygous, G/H and K/L homozygous, and M/N and E/F/I/J heterozygous. These last four chromosomes could not be separated into two clearly distinct pairs since a switch of identity was observed from E/I to E/F and from F/J to I/J (Figure 1 and Figure S3), suggesting that a translocation event has occurred between two of these chromosomes.

The global nucleotide sequence identity between heterozygous regions range from 70.6 to 88.6% (see Table S1). Some local loss of identity (hemizygous regions) are detected at the end of some chromosomes and in dispersed intrachromosomal areas (see Figure S3). If we do not consider these hemizygous regions, the remaining syntenic parts of the genome (11.85 Mb) shows no more than 89.16% nucleotide identity, revealing that *P. sorbitophila* is an hybrid genome derived from two distinct progenitors that have a high level of nucleotide polymorphism (10.84% of divergence) but a very well conserved synteny.

Homozygous regions and junctions between homozygous and heterozygous regions: Forty percent of the genome is homozygous (two identical or almost identical copies of sequence) and result from LOH events (Figure 1 and Table S1). LOHs concern chromosome pairs either in their totality (chr. G/H and K/L) or in part (chr. A/B and C/D). We quantified the number of SNPs in homozygous regions by realigning initial reads against the consensus sequences given by the supercontigs (see Table S2 for data and method). As a whole, the polymorphism is very low (0.46 SNP per 10 kb) and reveals a LOH process recently started. According to the chromosome pairs considered, different polymorphism levels are observed, from 1 SNP per 51,474 bp within C/D to 1 SNP per 18,000 bp within G/H. The first LOH event involved probably the G/H pair, the most polymorphic region, and may have been followed by the additional events in successive order: K/L, A/B, and C/D.

We also examined the junctions between the heterozygous and homozygous regions in the A/B and C/D chromosome pairs. In both cases, they are located inside protein coding genes and the two allelic copies of these genes are highly conserved at the nucleotide level (95.7% for A/B and 99.3% for C/D). It contrasts with the average identity observed for the others heterozygous genes (90.5%) and for the surrounding heterozygous regions (91%). No repetitive element or low complexity sequence is present at these junctions.

Position of centromeres: In *Debarymoyces hansenii* (Dujon *et al.* 2004), the most closely related yeast to *P. sorbitophila* (Figure 3), each chromosome has a unique island of highly repeated and degenerated sequences of retrotransposons Tdh5 with a poor GC content, which probably corresponds to the centromere (Butler *et al.* 2009; Dujon 2010; Lynch *et al.* 2010). In *P. sorbitophila*, we did not find any repetitive elements such as retrotransposons, but we identified for each chromosome a unique island with a poor GC content (10.8% less than the global GC content, see Figure S4), corresponding likely to the centromere position. These GC-poor regions range from 2.2 to 3.1 kb in size, are devoid of protein-coding genes or other features and are at equivalent positions in chromosomes forming a pair (Figure 1 and Figure S4). For homeologous pairs of chromosomes, all the proposed centromeres are located in heterozygous regions and share a very weak sequence identity or none, although surrounding allelic regions are well conserved (see Figure S3).

Extraction of Pγ and Pε parental genomes

In the case of *P. sorbitophila* and contrary to the analyses conducted so far on other yeast hybrids (Bovers *et al.* 2006; Dunn and Sherlock 2008; Gordon and Wolfe 2008; Nakao *et al.* 2009; Rainieri *et al.* 2006; Sipiczki 2008), the extraction method of the two parental subgenomes cannot be based on knowledge of the genomic sequence of at least one species closely related to one parent. Therefore, we used two complementary strategies to extract the parental subgenomes. A first strategy, determined by the GC content, led us to propose a subdivision of the genome. The proposed assignation of genomic regions to each subgenome was then tested and completed by a comparative analysis performed with the species *Pichia farinosa* CBS 2001.

GC variation and bias in codon usage: The global GC content calculated along chromosomes systematically revealed two different GC tendencies in heterozygous regions: each homeologous pair shows one chromosome with a greater GC content (average dGC = 1.006; $\sigma = 0.012$) and one chromosome with a lower GC content (average dGC = 0.994; $\sigma = 0.012$), as shown in Figure 1 using protein coding genes present in two syntenic copies between chromosomes forming a pair (Method described in the *Materials and Methods*). We hypothesized that it may reflect the subtle differences in nucleotide composition between the two parental genomes at the origin of the *P. sorbitophila* hybrid genome. We arbitrarily named "Pγ" the parental subgenome with the highest GC content and "Pε" the parental subgenome with the lowest one. We tested that hypothesis by analyzing the GC content variations in all protein coding genes (independently of their size) using tRNAs that pair with two different codons (Crick 1966). As shown in Table 1, genes in heterozygous regions defined as Pγ contain on average 1.57%

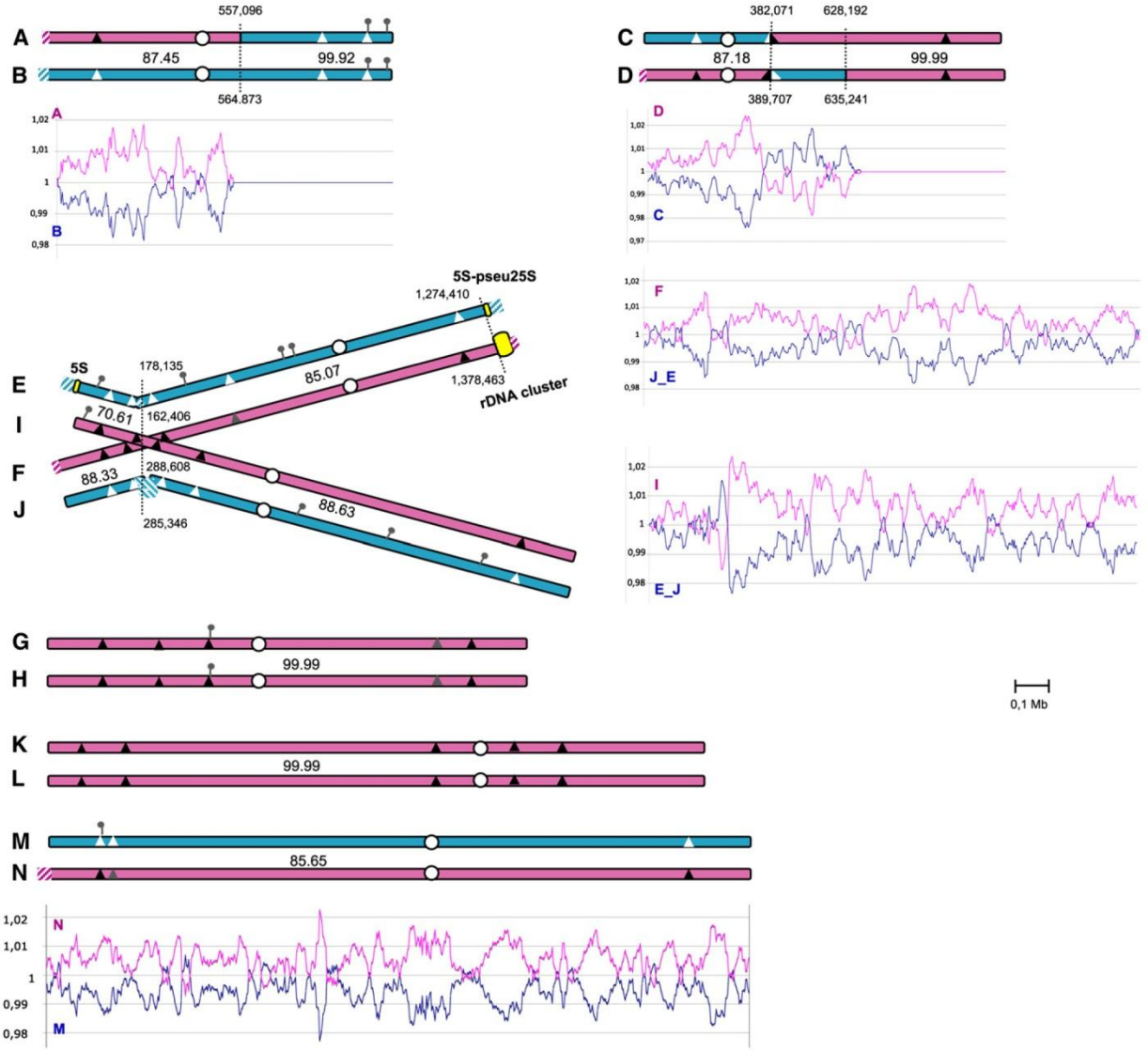


Figure 1 The hybrid nuclear genome of *P. sorbitophila*. The 14 chromosomes are represented by pairs according to their synteny. Chromosomes A/B and C/D are partly heterozygous, partly homozygous. Chromosomes E/F/I/J and M/N are heterozygous, and G/H and K/L are homozygous. The translocation breakpoint observed between chromosomes E, F, I, and J is represented by crossed chromosomes. Red and blue colors correspond to the proposed Pγ and Pε parental subgenomes, respectively, and hatched boxes to synteny losses between the homeologous chromosomes. Percentages of nucleotide identity between heterozygous regions are also indicated. Position of predicted centromeres (Figure S4) and of the 15 clusters of NUMT loci comprising a total of 24 NUMTs (Table S3) are shown by white dots and pins, respectively. Triangles indicate the positions of selected sequences used in *P. farinosa* CBS2001 strain (Table S4). The nucleotide identity level observed between *P. sorbitophila* and *P. farinosa* for the selected sequences is represented as followed: black triangles for 100% identity, gray for 99% and white for less than 96%. Below each chromosome pair (or at right for the crossed chromosomes) is represented the GC content variation (as shown in *Materials and Methods* and Figure S5) calculated for genes present at two allelic copies in chromosome pairs. The colors for curves correspond to subgenomes as for chromosomes.

more codons with a C at the third position than genes of Pε regions (mean = 1.56 [1.08-1.95]), allowing to propose an assignation of heterozygous regions to each subgenome (Figure 1, see also Figure S5). Applied to homozygous regions, the GC-based analyses suggested that both G/H and K/L pairs belong to Pγ subgenome (Figure S5). As for A/B and C/D homozygous parts, the discrimination was not obvious.

We finally observed a large exchange between GC% trend curves in the C/D chromosomal pair (Figure 1), with a bias of 2.92% (mean = 2.44 [1.39-4.36]) in the codon usage between upstream and downstream regions (Figure 2). We checked that it was not relevant to a sequence assembly error by confirming both chromosomal sequences using polymerase chain reaction amplifications and resequencing. Thus, the GC trend curve profile suggests that

a reciprocal exchange of sequences took place between both chromosomes. The left border of this chromosomal exchange (Figure 2) is located into a putative isoleucine tRNA synthetase coding sequence, with 100% sequence identity between both alleles. Other GC% trend curve exchanges were not taken into account in the following analyses, because they were of limited size and some of them likely corresponded to noise.

Pγ subgenome allelic sequences in *P. farinosa* CBS 2001: A taxonomical study of the *Pichia (Millerozyma) farinosa* group of species (S. Mallet *et al.*, unpublished results) revealed that several sequences from the *P. sorbitophila* Pγ subgenome were nearly 100% identical to the haploid yeast *Pichia farinosa* var. *farinosa* CBS 2001. To confirm the origin of each part of the *P. sorbitophila* genome, we completed this study by sequencing 17 other chromosomal sites in *P. farinosa* CBS 2001 (Table S4). Special attention was paid to cover the E/F/I/J translocation breakpoint as well as the C/D chromosomal exchange and NUMTs (NUclear sequences of MiTochondrial origin) positions (see section *Unequal acquisition of mitochondrial DNA sequences*, below). We observed that all markers in regions defined as belonging to Pγ subgenome shared 99% to 100% identity with *P. farinosa* CBS 2001, whereas all markers in regions attributed to Pε subgenome shared only 90% to 96% identity (Figure 1 and Table S4). Sequencing data for chr. A/B and C/D showed also that their homozygous regions belonged to Pε and Pγ, respectively. As a whole, these results confirm that: (1) the *P. sorbitophila* genome is an admixture of two parental subgenomes (Pγ and Pε); (2) the A/B, C/D, G/H, and K/L pairs were subjected to LOH processes; and (3) a C/D chromosomal exchange took place in the *P. sorbitophila* hybrid genome after its formation. Indeed, SNPs identified between both *P. sorbitophila* subgenomes and the genome of *P. farinosa* CBS 2001 (Figure 2C) are strongly correlated with the GC% trend curves exchange observed for chr. C/D (Figure 2A).

Phylogenetic position of Pγ and Pε: The sequence identity shared by *P. farinosa* CBS 2001 and Pγ for several positions on the genome suggests that one of the *P. sorbitophila* progenitors belongs to the *Millerozyma* group of species and is closely related to *P. farinosa* CBS 2001. We determined the phylogenetic position of Pγ and Pε in the CTG group of yeasts (Butler *et al.* 2009; Dujon *et al.* 2004; Jackson *et al.* 2009; Jeffries *et al.* 2007; Jones *et al.* 2004) (Table S5), using protein coding genes simultaneously present in both subgenomes (Figure S6). As shown in Figure 3, most of the branches of the phylogenetic tree obtained agree with previously published trees (Butler *et al.* 2009; Kurtzman and Suzuki 2010). Pγ and Pε are positioned in distinct strongly supported branches (bootstrap value = 100), with a phylogenetic distance corresponding to the separation of the species and with *D. hansenii* as the nearest known yeast of the CTG group with a fully sequenced genome. According to recent phylogenetic studies (Kurtzman and Suzuki 2010; S. Mallet *et al.*, unpublished results) Pγ and Pε correspond to two distinct species of the *Millerozyma* (*Pichia) farinosa* group of species.

Comparison between Pγ and Pε subgenomes

Divergence of protein-coding gene alleles: The genome of *P. sorbitophila* proves to be one of the smallest identified in the CTG group according to its size (10.75 Mb) and the number of annotated genes

Table 1 Bias in codon usage between Pγ and Pε subgenomes

Amino Acid	Codon	No. Codons		Usage, %		Pγ-Pε
		Pγ	Pε	Pγ	Pε	
Phe	TTT	35,134	37,411	50.56	53.55	2.99
	TTC	34,354	32,448	49.44	46.45	
Val	GTT	29,941	30,325	59.98	61.56	1.58
	GTC	19,977	18,937	40.02	38.44	
Ser	TCT	34,498	36,153	64.45	66.26	1.80
	TCC	19,030	18,412	35.55	33.74	
Pro	CCT	24,501	25,360	66.37	67.88	1.51
	CCC	12,414	11,999	33.63	32.12	
Thr	ACT	25,044	26,245	58.24	60.78	2.55
	ACC	17,961	16,932	41.76	39.22	
Ala	GCT	30,768	31,479	61.14	63.54	2.4
	GCC	19,557	18,060	38.86	36.46	
His	CAT	20,343	20,787	58.51	60.07	1.56
	CAC	14,426	13,818	41.49	39.93	
Asn	AAT	48,995	51,063	53.57	55.28	1.71
	AAC	42,467	41,316	46.43	44.72	
Asp	GAT	55,556	56,476	57.36	58.56	1.21
	GAC	41,307	39,959	42.64	41.44	
Cys	TGT	10,196	10,161	55.71	55.19	0.52
	TGC	8105	8250	44.29	44.81	
Ser	AGT	21,219	21,365	50.48	50.81	0.33
	AGC	20,817	20,682	49.52	49.19	
Gly	GGT	30,801	29,869	64.52	63.82	0.70
	GGC	16,937	16,931	35.48	36.18	
Avg.						1.57
Mean						1.56
Q1-Q3						[1.08-1.95]

For tRNA species that pair with two codons, the usage % of each codon was determined as follows: (number of one codon/number of both codon ×100). The values were calculated for all chromosomal regions defined as belonging to Pγ or Pε.

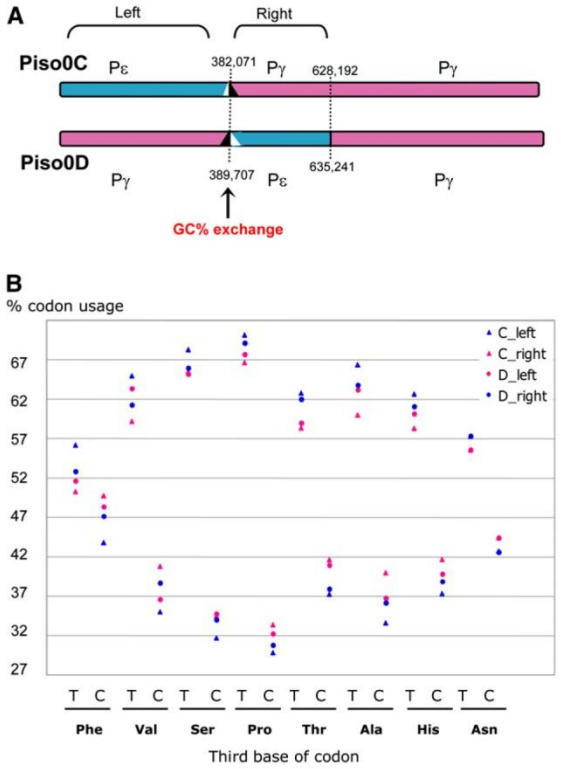


Figure 2 Analysis of the GC trend curve exchange between chr. C and D. (A) Position of the GC exchange determined by the global GC content analysis along chromosomes using a sliding window of 10 kb and a step of 1 kb. (B) Distribution of the codon usage percentages calculated from tRNA species that pair with two codons and showing more than 1.5 variation between both codons (extracted from Table 1, lane Pγ-Pε). Values for each tRNA were calculated for the left and the right regions of the C/D chromosomal exchange, respectively. (C) Multiple alignments of *P. sorbitophila* and *P. farinosa* CBS 2001 sequences around the GC trend curve exchange. This exchange area is characterized by a 100% identical region (1934-nt long) between chr. C and D. Positions of SNPs between *P. sorbitophila* chr. C, chr. D and *P. farinosa* CBS 2001 are indicated in red. As shown, *P. farinosa* chromosomal sequence is first identical to *P. sorbitophila* chr. D sequence and then to chr. C (after the 1934-nt long sequence), confirming that the GC exchange ensues from a reciprocal translocation event.

(5626 protein-coding genes), if we consider the equivalent haploid genome (Table 2 and Table S7). The gene redundancy is limited for both subgenomes compared to *D. hansenii* with a total of 33.2% protein-coding genes belonging to multigene families against 51.5% and three times less tandemly duplicated gene arrays (Table S8). In each chromosomal pair, a protein coding gene is present in most cases in two allelic copies coming either from both parents (in heterozygous regions) or from a sole parent (in homozygous regions), giving a total of 11,252 loci (Table 2 and supporting information). When genes with introns are located in heterozygous regions, both alleles of the genes contain the same number of introns, almost identical in size (Table S6).

Among the 3425 genes in heterozygous regions, 3205 (93.6%) are represented by two nonidentical coding-alleles (Table 2) demonstrating on average 92.1% of identity at the protein level (Id_{prot}) and a mean ratio of nonsynonymous substitutions per synonymous substitutions (dN/dS) of 0.121 (Figure S9). Three subsets of heterozygous genes are particularly interesting. First, genes showing highly conserved alleles (44 genes, dN/dS $<$ 0.0046, $Id_{prot}>$ 99.8%) likely encode for essential functions (Table S9). Second, some genes with highly divergent alleles (15 among 90 genes) correspond to "*Millerozyma*-specific genes" found only in Pγ and Pε subgenomes but not in the other yeasts of the CTG group (Table S10). They probably appeared in the Pγ and Pε common ancestor after its separation with the ancestor of *D. hansenii* and are under a relaxed selection pressure. The third subset is composed of 38 CDS-pseudogene allele pairs (CDS or CoDing Sequence is defined here as the region of nucleotides that corresponds to the sequence of amino acids in the predicted protein and that is not interrupted by internal frameshift or stop codon). The allele distribution for these genes is not homogeneous since almost all pseudogene alleles (81.6%) are located in the Pε subgenome whereas the corresponding coding alleles are in the Pγ. Finally, the two groups of genes with highly divergent alleles (90 CDS-CDS pairs and 38 CDS-pseudogene pairs) may constitute an interesting pool of genes probably required for specific adaptations to environmental conditions, since they show bias in Gene Ontology frequencies in favor of transporters, phosphatases, oxidoreductases, and cell wall proteins (Table S11 and Table S12).

Only a limited number of genes (220 genes) in one subgenome has no allelic counterpart in the other subgenome. Among them, 102 correspond probably to "dubious open reading frames" (Figure S8), reducing the number of single allele genes to 118. They are located in three different genomic areas: 83 genes (70.3%) at the end of contigs, 12 genes (10.2%) in the E/F/I/J translocation breakpoint and 23 genes (19.5%) in the heterozygous regions. The distribution of single allele genes between Pγ and Pε is slightly in favor of Pε with 78 genes (66.7%) against 40 for Pγ (33.3%). Despite their small number (2% of the total number of genes), they are at the origin of some metabolic pathways in the *P. sorbitophila* hybrid as described in the section *Unilateral acquisition of genes for sugar degradation.*

Uniparental conservation of ribosomal RNA genes: Despite the fact that the *P. sorbitophila* nuclear genome has three different ribosomal DNA (rDNA) loci, the rDNA is at a hemizygous state with

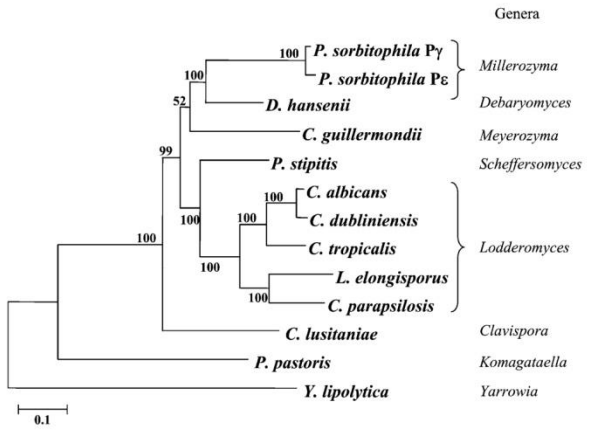


Figure 3 Phylogenetic positions of the two subgenome sequences identified in *P. sorbitophila* hybrid. *Yarrowia lipolytica* is used as outgroup to root the CTG tree. The tree was built from the alignment of 233 protein families (87,181 amino acids per species) having a single member in each analyzed species. Amino acid sequences for each family were aligned with MAFFT (Katoh *et al.* 2009) and cleaned with Gblocks (Talavera and Castresana 2007). The tree was built from the resulting alignment with the maximum likelihood method using PHYML with a JTT substitution model corrected for heterogeneity among sites by a Γ-law distribution using four different categories of evolution rates (Guindon and Gascuel 2003). The proportion of invariable sites and the α-parameter of the Γ-law distribution were optimized according to the data. Bootstraps were calculated from 100 replicates. They are indicated before each node and the scale for branch length at the bottom of the figure.

a sole cluster of repeats (Figure 4). The latter is located on the right arm of chr. F (Pγ) and contains the 5S and 35S transcript units (precursor of 18S, 5.8S, and 25S RNAs), repeated around 73 times (Figure 4B). At the allelic position on chr. E (Pε, right arm), a whole 5S unit and a deleted 25S unit are present, but without repetition. Finally, another relic of rDNA cluster (a single 5S rDNA locus) is at the opposite left arm of chr. E (Pε). It seems therefore that, after the hybrid formation, the Pε-like parental chromosome (chr. E) underwent an entire loss of the rDNA repeats. The two rDNA relics on chr. E and issued from Pε share 90.3% sequence identity (Figure S10) suggesting that they have already been present in the Pε parental genome before the hybrid formation. These two loci are located in highly polymorphic subtelomeric regions (Figure S11), that could have contributed to the loss of the rDNA repeats. The hemizygous state of the ribosomal DNA contrasts with the distribution of other noncoding RNA genes (ncRNA and tRNA), which are all represented by two well-conserved alleles (see supporting information).

Unequal acquisition of mitochondrial DNA sequences: Most eukaryotic nuclear genomes contain pieces of mitochondrial sequences (Lenglez *et al.* 2010; Sacerdot *et al.* 2008), designated as NUMTs for NUclear sequences of MiTochondrial origin. They result from the transfer of fragments of mitochondrial DNA into chromosomes. The number and size of the NUMTs vary significantly between yeasts within the monophyletic group of hemiascomycetes (Sacerdot *et al.* 2008). For the *P. sorbitophila* nuclear genome, we observed a highly unequal distribution of NUMTs (Figure 1 and Table S3). First, of the 24 NUMTs identified from its extant mitochondrial genome (Jung *et al.* 2009), 20 are exclusively located in the subgenome deriving from the Pε parent. Therefore, most of the NUMTs (14 NUMTs) are at a hemizygous state with the equivalent allelic sequence being devoid of NUMT. Three arguments lead us to propose that the mitochondrial sequences have been inserted into the nuclear genomes of both parents before the hybridization event: (1) all NUMTs in heterozygous regions are at a hemizygous state; (2) the present *P. sorbitophila* hybrid is separated from its ancestral hybrid by only few thousand successive generations (estimated from the polymorphism level in homozygous regions and based on the method described by Rolland and Dujon 2011), whereas the number of NUMTs insertions is equivalent to what observed for other yeast genomes; and (3) the NUMT located on the G/H homozygous pair in two identical alleles is also present in *P. farinosa* CBS 2001 (Table S4), closely related to the Pγ progenitor. The nonuniform distribution of NUMTs reveals also that Pε has undergone four times more mitochondrial DNA insertions than Pγ. The sequence divergence between the *P. sorbitophila* and *P. farinosa* CBS 2001 mitochondrial *COX2* genes suggests that the

Table 2 *P. sorbitophila* genomic features in Pγ and Pε subgenomes

Parental Contribution	Chromosomal Region	Total No. ProteinCoding Genes				Total No. Noncoding RNA					Total No. Other Elements NUMTs Loci
		CDS Without Intron	CDS With Introns	Pseudo-Gene	Total Gene	tRNA	snoRNA	snRNA	Pol III ncRNA	Ribosomal DNA[a]	
Pγ	Solo	95	2	7	104	0	0	0	0	73	1
	Heterozygous[b]	2973	219	13	3205	88	21	3	5	0	0
	Homozygous	3834	246	10	4090	92	26	2	0	0	2
	Total	6902	467	30	7399	180	47	5	5	73	3
Pε	Solo	106	4	6	116	0	0	0	0	0	8
	Heterozygous[b]	2950	218	37	3205	88	21	3	5	0	0
	Homozygous	482	46	4	532	20	6	0	0	0	4
	Total	3538	268	47	3853	108	27	3	5	0	12
Total genome		10,440	735	77	11,252	288	74	8	10	73	15
Haplotype equivalent		5,220	367.5	38.5	5626	144	37	4	5	36.5	7.5

CDS, CoDing Sequence; Pol III, polymerase III; NUMTs, NUclear sequences of MiTochondrial origin.
[a] Tandemly repeated units.
[b] Loci (3205 pairs) containing both parental genes (a total of 6410 genes). They correspond to CDS/CDS pairs for 3161 pairs, to CDS/pseudogene pairs for 38 pairs, and to pseudogene/pseudogene pairs for 6 pairs (supporting information).

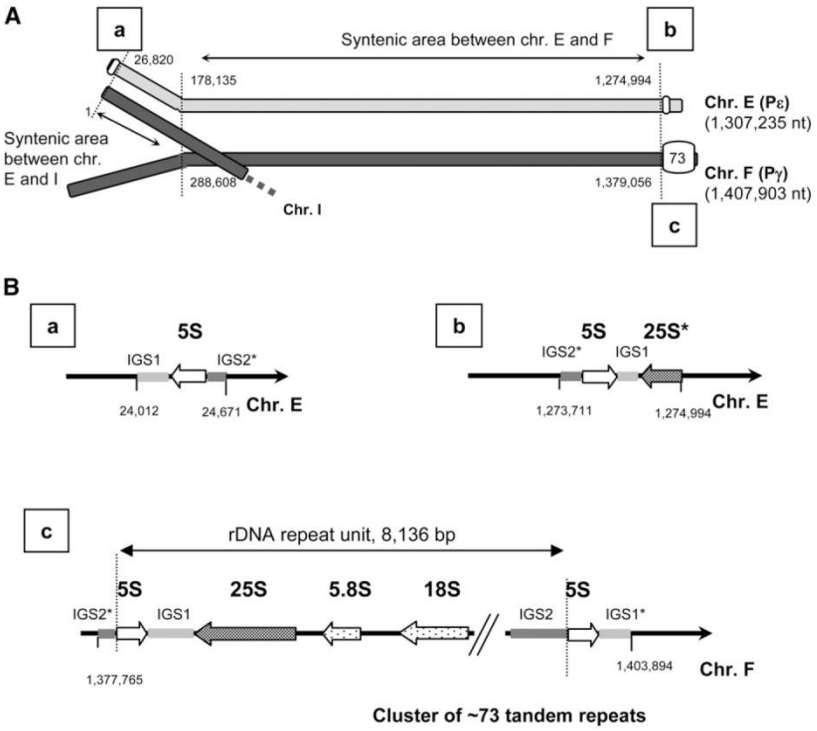


Figure 4 Location and organization of the ribosomal DNA clusters. rDNA sequences were identified by comparison to other yeast genomes (Table S5). Three loci were identified, one at the left border of chr. E ("a" in A and B), a second at the right border of chr. E ("b" in A and B) and a third at the left border of chr. F ("c" in A and B), the latest one containing approximately 73 tandem repeats. (A) Indicates the position and coordinates of each locus on chromosomes. (B) Describes the loci organizations, incomplete elements are indicated by stars. These three organizations were also checked by polymerase chain reaction amplification, end sequencing and PFGE hybridization (Figure S2).

mitochondrial genome (mtDNA) of *P. sorbitophila* was inherited from the Pε progenitor (see supporting information).

Chromosomal rearrangements in parental genomes: Three distinct chromosomal areas of the *P. sorbitophila* genome were analyzed to identify molecular events at the origin of synteny breaks: (1) E/F/I/J chromosomes, (2) subtelomeric regions, and (3) internal chromosomal positions where single allele genes are located.

Gene orders at the E/F/I/J translocation breakpoint were compared with the orthologous regions identified in *D. hansenii* and *Candida guillermondii* (Figure S12). The fact that the gene orders on chr. I and F (Pγ) are identical or almost identical to those of *D. hansenii* and *C. guillermondii* suggests that the chromosomal arm exchange likely took place between the two other chromosomes, chr. E and J (Pε). The short evolutionary period of *P. sorbitophila* since its formation suggests that this event appeared in Pε before the hybridization event, although we cannot exclude that it was part of the early chromosomal rearrangements occurred in the hybrid genome. The translocation breakpoint is particularly enriched in tandemly duplicated genes. Among the 12 single-allele genes located in chr. E and J, eight are tandemly duplicated and form four distinct tandem gene arrays (Figure S12). Insertion and amplification of those duplicated genes probably either induced the chromosomal exchange or were initiated by the exchange.

The plasticity of subtelomeric regions and their capacity to harbor large gene families make their comparison between subgenomes extremely difficult. However, we found that some of the 83 single allele genes located in subtelomeric regions were the result of gene location movements in parental genomes. An example of such cases is given in Figure S13.

The remaining 23 single allele genes in heterozygous regions are mainly concentrated in 10 internal chromosomal areas (Figure S3). Compared with the corresponding gene order with *D. hansenii*, *C. guillermondii*, *Pichia stipitis*, and *C. albicans*, the single allele genes are missing in all other species at these specific locations whereas the surrounding regions are highly syntenic (Figure S14). Consequently, most of these genes were probably inserted into the corresponding genomic areas in one parental genome after the separation of both parents from their common ancestor. These insertions may result from diverse molecular events: gene location movement as observed in subtelomeric regions, gene acquisition (four single-allele genes are species-specific), gene duplication in cases of multigene families (eight single-allele genes), and tandem duplication as observed for chr. F (Figure S14) or more complex events.

Acquisition of metabolic pathways

Unilateral acquisition of genes for sugar degradation: Analysis of single-allele genes at the level of their putative encoded functions shows that some key genes for *P. sorbitophila* metabolism have been acquired from the Pε parent. They are still conserved in *P. sorbitophila* hybrid genome, although the Pε subgenome constitutes no more than 32% of the totality. This is noteworthy for maltose degradation (Table 3): *P. sorbitophila* is able to hydrolyze maltose (de Miranda *et al.* 1980) in contrast to *P. farinosa* CBS 2001 (strain closely related to Pγ). The *MAL* genes (Alves *et al.* 2008; Chow *et al.* 1989) are exclusively single allele genes inherited from Pε and exist in several copies at four different areas of synteny break in heterozygous regions: for example, the E/F/I/J translocation breakpoint contains two tandemly duplicated *MALX3* genes on chr. E and two tandemly duplicated *MALX2* genes on chr. J (Figure S12). At a synteny breakpoint between chr. M and N, four of the five additional genes on chr. M are *MAL* genes, with the corresponding gene order *MALX2-MALX1-MALX3* (pseudogene)–*MALX3* (Figure S14). As a whole, the Pε subgenome encodes three

Table 3 Distribution of single-allele genes between Pγ and Pε for sugar degradation and other transports

Putative Function	Gene Name	Locus in Pγ	Locus in Pε
Sugar metabolism			
Maltose permease	MALX1		PISO0M16930g
			PISO0J21547g
			PISO0M00166g
Maltase	MALX2		PISO0M00188g
			PISO0M16886g
			PISO0J03551g
			PISO0J03441g
MAL activator	MALX3		PISO0E02028g
			PISO0E02050g
			PISO0M16974g
			PISO0J35007g[a]
Invertase	SUC2		PISO0J03639g
			PISO0J03573g
Sorbitol dehydrogenase	SOR1	PISO0N22123g	PISO0M21880g
			PISO0E00180g
		PISO0K00604g/ PISO0L00605[b]	
Gluthatione metabolism			
5-oxoprolinase	OXP1	PISO0I02648g	PISO0J04431g
		PISO0C10078g/PISO0D10145g[b]	
			PISO0E04690g
			PISO0J03595g
Allantoate transport			
Allantoate permease	DAL5		PISO0A12958g/ PISO0B13025g[b]
			PISO0M24916g
			PISO0M17480g
		PISO0K23022g/ PISO0L23023g[b]	
			PISO0J21503g
		PISO0I08192g	PISO0J10019g
			PISO0E04646g
Nicotinic acid transport	NTA1	PISO0I02626g	PISO0J04409g
			PISO0J03617g
			PISO0J03485g
		PISO0K00318g/ PISO0L00319g[b]	
		PISO0K00406g/ PISO0L11407g[b]	
		PISO0N15105g	PISO0M14708g

[a] Pseudogene.
[b] Identical alleles located in homozygous regions.

MALX1 permeases, four *MALX2* maltases and three *MALX3* activators whereas Pγ has none (Table 3).

We also observed a bias in favor of the Pε subgenome for the invertase *SUC2* gene (Carlson *et al.* 1983; Taussig and Carlson 1983) with one tandem array of two *SUC2* genes at the translocation breakpoint (Figure S12) and no copies in Pγ subgenome. The Pε subgenome also contributes for seven of the ten allantoate permease genes and the Pγ for three of five sorbitol dehydrogenase genes (Table 3). In conclusion, a limited number of genes have been deleted or acquired in both parents since their separation from their common ancestor. These minor differences may however have significant consequences on *P. sorbitophila* metabolism.

Biparental acquisition of genes for stress resistance, mating type, and meiosis: *P. sorbitophila* was isolated as a contaminant of a 70% sorbitol solution (de Miranda *et al.* 1980). Compared with *S. cerevisiae* and *D. hansenii*, *P. sorbitophila* is more tolerant to NaCl (4M), LiCl (0.8 M), and KCl (2.5 M) (Maresova and Sychrova 2003). Previous studies on Na^+ intracellular limitation pathways in *P. sorbitophila* allowed the identification of three key transporters: the *NHA1* and *NHA2* cation/H^+ antiporters (Banuelos *et al.* 2002), which correspond to two alleles of the same gene; the P-type ATPase (Benito *et al.* 2004); and a H^+/glycerol symport activity (Lages and Lucas 1995). To complete those data, we searched for the presence of more than 50 known genes involved in ion, glycerol, and water transport or in osmotolerance-related pathways (Table S18), identified in *S. cerevisiae* (obtained from Saccharomyces Genome Database, http://www.yeastgenome.org/) and also well documented for *D. hansenii* (Prista *et al.* 2005). We found that, in contrast to the metabolism of sugar degradation, osmotolerance genes were not specific to one subgenome but rather identified with their two related coding alleles in heterozygous regions.

To achieve Na^+ efflux, several *ENA* genes ("Exitus NAtru" genes) encoding Na^+-ATPases are found in *D. hansenii* and *S. cerevisiae* (Benito *et al.* 2004), whereas only one *ENA* gene has been identified in *P. sorbitophila*. The latter also lacks the *HAL1* gene (Rios *et al.* 1997), which decreases intracellular Na^+ via *ENA1*. Potassium is the most abundant intracellular cation in living cells and plays important roles in biological processes. The *NHA1-2* Na^+/H^+ antiporter is not specific for Na^+ but also mediates K^+ efflux in addition to Na^+, allowing a possible intracellular K^+ depletion (Benito *et al.* 2004). In *P. sorbitophila*, this depletion may be intensified by the presence of *TOK1* (Table S18) a permeable channel for potassium efflux (Ketchum *et al.* 1995; Rios *et al.* 1997) missing in *D. hansenii*. To counterbalance the K^+ efflux at high NaCl concentrations, it has been

proposed that an efficient K^+ uptake system must exist in *P. sorbitophila* (Banuelos *et al.* 2002). We identified the two potassium transporters *HAK1* (high-affinity K transporter) and *TRK1* (TRansport of K) and the P-type ATPase *ACU1* (Benito *et al.* 2004). *P. sorbitophila* lacks the *PHO89* Na^+/Pi cotransporter (Martinez and Persson 1998) in contrast to *D. hansenii*, which has *PHO89* but not *ACU1*. Osmoregulation is also achieved by the production of glycerol or other osmolytes (*e.g.*, arabitol, erythritol) and the capacity to maintain them into the cells (Kayingo *et al.* 2001; Lages and Lucas 1995; Lages *et al.* 1999). Glycerol leaks through the plasma membrane and its retention therefore needs an active transport system. In *P. sorbitophila*, H^+/glycerol symport allows the intracellular accumulation of glycerol (Lages and Lucas 1995) and we also show that the aquaglyceroporin *FPS1*, the glycerol permease responsible for glycerol leakage (Tamás *et al.* 1999), is missing.

As for stress resistance, *P. sorbitophila* genes involved for mating, meiosis, and spore formation are in two allelic versions, suggesting a conservation of these genes from the common ancestor of Pγ and Pε progenitors until the present hybrid. In *P. sorbitophila*, we found two similar mating-type loci at the heterozygous allelic positions in chr. M and N (Figure S15). Both contain *MATa2* and *MATalpha1* genes but lack *MATa1* and *MATalpha2*. Thus, as *P. stipitis* and *D. hansenii* (Butler 2010; Fabre *et al.* 2005; Jeffries *et al.* 2007; Lee *et al.* 2010), *P. sorbitophila* has its mating-type loci fused with no *MATalpha2*. *MATa1* is missing only in *P. sorbitophila*, a gene loss that likely took place in the common ancestral species of both parents. We also searched for the presence of 227 orthologs of *S. cerevisiae* genes involved for mating, meiosis and spore formation since key genes for the sexual development and meiosis in *S. cerevisiae* are missing in *Candida* species (Butler *et al.* 2009), and we also revisited the gene annotations of the 2nd version of the *D. hansenii* genome (Souciet *et al.* 2009). All genes identified in *D. hansenii* (187 genes) were detected in two copies in *P. sorbitophila* (Table S19). Three additional genes are present in *P. sorbitophila*: *DIT1* and *DIT2* involved in dityrosine synthesis (Briza *et al.* 1994) and *YEL023C* which may participate in this process (Butler *et al.* 2009). Dityrosine is the major component of the outer layer of spore wall. It forms an insoluble scaffold on the surface of the spore (Briza *et al.* 1988, 1990) and contributes to its resistance. There are no data available for spore resistance concerning *P. sorbitophila*. However, we can speculate that the presence of these three genes may contribute to the resistance of spores in specific environmental conditions.

CONCLUSION

Allopolyploid hybrids, which are the result of cell fusions between different species, undergo generally rapid and extensive genomic modifications after their formation (Rainieri *et al.* 2006; Sipiczki 2008). With the aim of deciphering hybrid genome evolution, these posthybridization rearrangements have to be distinguished from those which took place in the parental genomes during the long period that preceded the hybrid formation. The complete sequencing of the genome of the osmotolerant yeast *P. sorbitophila* strain CBS 7064 (de Miranda *et al.* 1980) revealed that it is a hybrid genome. Despite the fact that no complete genome sequence closely related to that of one or both parents is available so far, a detailed analysis of its genome led us to retrace its evolution. The *P. sorbitophila* nuclear genome (21.5 Mb) is actually composed of seven pairs of chromosomes issued from two progenitors named Pγ and Pε, with Pγ being closely related to *P. farinosa* CBS 2001. In contrast, the origin of the Pε parent remains unknown because of the lack of close sequence similarity among existing yeasts. The sequence divergence observed between both subgenomes (10.84% at the nucleotide level between syntenic regions) is equivalent to the one described between the genomes of *S. cerevisiae* and *S. paradoxus*, two distinct species of the genus *Saccharomyces* (Cliften *et al.* 2001).

In addition, it reflects the long evolutionary period that separates the Pγ and Pε parents from their common ancestor. During this long evolutionary period, the genomes of both parents underwent very few genomic reshaping events: some gene acquisitions and gene location movements, rare gene duplications, differential levels of NUMTs insertions and probably one chromosomal arm translocation, allowing conservation of large syntenic gene blocs. However, these events might have been decisive for the *P. sorbitophila* hybrid evolution and adaptability, as shown by the uniparental acquisition of *MAL* and *SUC* genes. At the opposite, genes involved in osmotic stress resistance or spore resistance were already acquired by the Pγ and Pε common ancestor. For these classes of genes, heterosis might still result from interaction between distinct alleles. *P. sorbitophila* offers therefore a unique case to study acquisitions of novel functional properties originating from the admixture of the parental genetic contributions.

The *P. sorbitophila* genome also provides an interesting snapshot of the genomic evolutionary events after an interspecific hybridization in eukaryotes (as summarized in Figure 5). On one hand, we observed a diploid level for this genome, in contrast to aneuploid situations reported for hybrids of the genus *Saccharomyces* (Querol and Bond 2009), probably attributable to the hybridization event that generated directly a strict allodiploid hybrid (Gerstein *et al.* 2006). On the other hand, LOH appears to play a prominent role. In total, 40.3% of this genome has only one parental origin (35.5% from Pγ and 4.8% from Pε). For all concerned chromosomes, LOH extends up to telomeres, a phenomenon also observed in *Candida* species (Butler *et al.* 2009; Diogo *et al.* 2009; Forche *et al.* 2005). In the partly homozygotized chromosomal pairs, LOH origin ignores gene borders. The limited but existing sequence polymorphism observed between pairs of homozygotized regions indicates that at least four successive LOH events took place in this hybrid. Assuming the same mutational rate as in *S. cerevisiae* and considering that the few mutational changes were essentially neutral in homozygotized regions, we can estimate that LOH process started only few thousand generations ago (estimated at nearly 185,000 generations from method described in Rolland and Dujon 2011), and that the formation of the hybrid was just anterior to the beginning of this process. According to Fay and Benavides (2005) and to Rolland and Dujon (2011), we can speculate that the hybrid formation occurred in the last centuries. This estimated time is consistent with the fact that *P. sorbitophila* was isolated from a manufacturing product (a highly concentrated sorbitol solution).

In contrast to the extent of LOH, uniparental gene loss played a limited role during the evolution of the *P. sorbitophila* genome, with the notable exceptions of rDNA (inherited solely from Pγ) and mtDNA (inherited from Pε). The unilateral loss of rDNA was also observed in other hybrid eukaryotic genomes. In the allotetraploid grass *Zingeria trichopoda*, the *Z. biebersteiniana* like parental chromosomes would have undergone a massive loss of 45S rDNA (Kotseruba *et al.* 2003). In the lager brewing yeast *S. pastorianus*, the rDNA from the *S. cerevisiae*-type subgenome is approximately 20 times more represented than its *S. bayanus*-type counterpart (Nakao *et al.* 2009), whereas both of the parental rDNA types were retained in the *Zygosaccharomyces* allopolyploid (Gordon and Wolfe 2008). The consequence of this unilateral loss, that is, the transcription of rDNA genes inherited from a sole parent, is comparable with the nucleolar dominance observed in numerous plant interspecific hybrids. In this latter case, rDNA loci inherited from both parents are conserved but

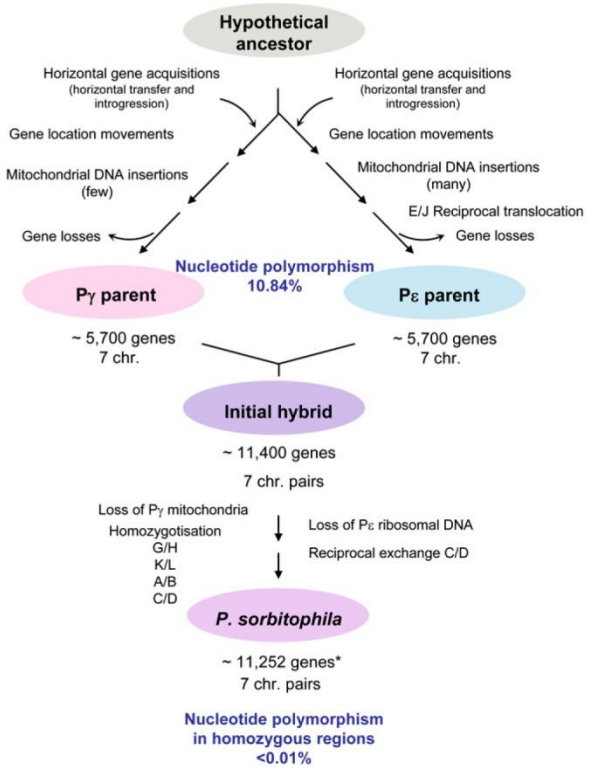


Figure 5 Genomic reshapings identified since the separation of Pγ and Pε parents from their common ancestor to the current *P. sorbitophila* genome. Position of rearrangements in each part of the flowchart (before or after hybridization) is not relevant of their chronology. (*) The 11,252 genes in *P. sorbitophila* are in fact 3205 gene pairs coming from both parents, 2045 pairs coming from only Pγ parent, 266 pairs from Pε parent, and 116 and 104 are single-copy genes derived from either Pγ or Pε, respectively.

rDNA genes derived from one progenitor are silent (Lewis *et al.* 2007; Pikaard 2000). These observations raise the problems of the origin of the rDNA instability and the viability of meiotic products. Our preliminary results of sporulation experiments show that *P. sorbitophila* is able to produce asci containing one to four ascopores, as observed previously by de Miranda *et al.* (1980) and Oliveira *et al.* (1996). Results also suggest that at most two spores per ascus are viable (data not shown). These preliminary results could be consistent with the hemizygous state of the rDNA if we assume that *P. sorbitophila* is able to undergo meiosis, a hypothesis that requires additional data to be confirmed.

The reconstruction of the complete parental subgenomes allowed us to decipher the recent *P. sorbitophila* genome history and, therefore, to depict, gene by gene, how two divergent genomes put together into a viable hybrid are rearranged during the process of genome stabilization. Our results also show that interspecies hybrids, because of poor prezygotic barrier, are widespread in the *Saccharomycotina* group of species. Human activities in industrial contexts provide unusual substrates (70% sorbitol, for example) that may act as bottlenecks for the selection of particularly resistant species of yeasts and fungi or hybrids. This was probably the case for *P. sorbitophila*. Its genome analysis gave us therefore the opportunity to highlight, in the present times, the early steps of genome evolution after the formation of a hybrid.

ACKNOWLEDGMENTS

We thank Henri Grosjean and Fredj Tekaia for useful discussions. An interactive website can be found at http://www.genolevures.org/. 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). The computer framework was supported by the funding of the University of Bordeaux 1, the Aquitaine Région in the program "Génotypage et Génomique Comparée," and the ACI IMPBIO "Génolevures En Ligne." We thank the System and Network Administration team in LaBRI for excellent help and advice. J.A.C. is supported by the PhD Program in Computational Biology of the Instituto Gulbenkian de Ciência, Portugal (sponsored by Fundação Calouste Gulbenkian, Siemens SA, and Fundação para a Ciência e Tecnologia; SFRH/BD/33528/2008). B.D. is a member of Institut Universitaire de France.

LITERATURE CITED

Alves, S. L. J., R. A. Herberts, C. Hollatz, D. Trichez, L. C. Miletti *et al.*, 2008 Molecular analysis of maltotriose active transport and fermentation by *Saccharomyces cerevisiae* reveals a determinant role for the AGT1 permease. Appl. Environ. Microbiol. 74: 1494–1501.

Arnold, M. L., and N. H. Martin, 2010 Hybrid fitness across time and habitats. Trends Ecol. Evol. 25: 530–536.

Banuelos, M. A., J. Ramos, F. Calero, V. Braun, and S. Potier, 2002 Cation/H+ antiporters mediate potassium and sodium fluxes in *Pichia sorbitophila*. Cloning of the PsNHA1 and PsNHA2 genes and expression in *Saccharomyces cerevisiae*. Yeast 19: 1365–1372.

Batzoglou, S., D. B. Jaffe, K. Stanley, J. Butler, S. Gnerre *et al.*, 2002 ARACHNE: a whole-genome shotgun assembler. Genome Res. 12: 177–189.

Belloch, C., S. Orlic, E. Barrio, and A. Querol, 2008 Fermentative stress adaptation of hybrids within the Saccharomyces sensu stricto complex. Int. J. Food Microbiol. 122: 188–195.

Belloch, C., R. Perez-Torrado, S. S. Gonzalez, J. E. Perez-Ortin, J. Garcia-Martinez *et al.*, 2009 Chimeric genomes of natural hybrids of *Saccharomyces cerevisiae* and *Saccharomyces kudriavzevii*. Appl. Environ. Microbiol. 75: 2534–2544.

Benito, B., B. Garciadeblas, P. Schreier, and A. Rodriguez-Navarro, 2004 Novel p-type ATPases mediate high-affinity potassium or sodium uptake in fungi. Eukaryot. Cell 3: 359–368.

Bond, U., I. L. Allen, S. Sima, and M. G. Geoffrey, 2009 The genomes of lager yeasts, pp. 159–182 in *Advances in Applied Microbiology*, Vol. 69, edited by A. I. Laskin, S. Sariaslani, and G. M. Gadd. Academic Press, San Diego.

Bovers, M., F. Hagen, E. E. Kuramae, M. R. Diaz, L. Spanjaard *et al.*, 2006 Unique hybrids between the fungal pathogens *Cryptococcus neoformans* and *Cryptococcus gattii*. FEM. Yeast Res. 6: 599–607.

Briza, P., A. Ellinger, G. Winkler, and M. Breitenbach, 1988 Chemical composition of the yeast ascospore wall. The second outer layer consists of chitosan. J. Biol. Chem. 263: 11569–11574.

Briza, P., A. Ellinger, G. Winkler, and M. Breitenbach, 1990 Characterization of a DL-dityrosine-containing macromolecule from yeast ascospore walls. J. Biol. Chem. 265: 15118–15123.

Briza, P., M. Eckerstorfer, and M. Breitenbach, 1994 The sporulation-specific enzymes encoded by the DIT1 and DIT2 genes catalyze a two-step reaction leading to a soluble LL-dityrosine-containing precursor of the yeast spore wall. Proc. Natl. Acad. Sci. USA91: 4524–4528.

Butler, G., 2010 Fungal sex and pathogenesis. Clin. Microbiol. Rev. 23: 140–159.

Butler, G., M. D. Rasmussen, M. F. Lin, M. A. Santos, S. Sakthikumar *et al.*, 2009 Evolution of pathogenicity and sexual reproduction in eight Candida genomes. Nature 459: 657–662.

Carlson, M., R. Taussig, S. Kustu, and D. Botstein, 1983 The secreted form of invertase in *Saccharomyces cerevisiae* is synthesized from mRNA encoding a signal sequence. Mol. Cell. Biol. 3: 439–447.

Chou, J. Y., Y. S. Hung, K. H. Lin, H. Y. Lee, and J. Y. Leu, 2010 Multiple molecular mechanisms cause reproductive isolation between three yeast species. PLoS Biol. 8: e1000432.
Chow, T. H., P. Sollitti, and J. Marmur, 1989 Structure of the multigene family of MAL loci in *Saccharomyces*. Mol. Gen. Genet. 217: 60–69.
Cliften, P. F., L. W. Hillier, L. Fulton, T. Graves, T. Miner *et al.*, 2001 Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. Genome Res. 11: 1175–1186.
Crick, F. H., 1966 Codon–anticodon pairing: the wobble hypothesis. J. Mol. Biol. 19: 548–555.
de Miranda, R. L., K. R. Appel, and H. Seyfarth, 1980 *Pichia sorbitophila* sp nov. Antonie van Leeuwenhoek 46: 157–159.
De Montigny, J., C. Spehner, J.-L. Souciet, F. Tekaia, B. Dujon *et al.*, 2000 Genomic Exploration of the Hemiascomycetous Yeasts: 15. *Pichia sorbitophila*. FEBS Lett. 487: 87–90.
Diogo, D., C. Bouchier, C. d'Enfert, and M. E. Bougnoux, 2009 Loss of heterozygosity in commensal isolates of the asexual diploid yeast *Candida albicans*. Fungal Genet. Biol. 46: 159–168.
Dujon, B., 2010 Yeast evolutionary genomics. Nat. Rev. Genet. 11: 512–524.
Dujon, B., D. Sherman, G. Fischer, P. Durrens, S. Casaregola *et al.*, 2004 Genome evolution in yeasts. Nature 430: 35–44.
Dunn, B., and G. Sherlock, 2008 Reconstruction of the genome origins and evolution of the hybrid lager yeast *Saccharomyces pastorianus*. Genome Res. 18: 1610–1623.
Fabre, E., H. Muller, P. Therizols, I. Lafontaine, B. Dujon *et al.*, 2005 Comparative genomics in hemiascomycete yeasts: evolution of sex, silencing, and subtelomeres. Mol. Biol. Evol. 22: 856–873.
Fay, J. C., and J. A. Benavides, 2005 Evidence for domesticated and wild populations of *Saccharomyces cerevisiae*. PLoS Genet. 1: e5.
Forche, A., G. May, and P. T. Magee, 2005 Demonstration of loss of heterozygosity by single-nucleotide polymorphism microarray analysis and alterations in strain morphology in *Candida albicans* strains during infection. Eukaryot. Cell 4: 156–165.
Gerstein, A. C., H. J. Chun, A. Grant, and S. P. Otto, 2006 Genomic convergence toward diploidy in *Saccharomyces cerevisiae*. PLoS Genet. 2: e145.
Gordon, J. L., and K. H. Wolfe, 2008 Recent allopolyploid origin of *Zygosaccharomyces rouxii* strain ATCC 42981. Yeast 25: 449–456.
Greig, D., E. J. Louis, R. H. Borts, and M. Travisano, 2002 Hybrid speciation in experimental populations of yeast. Science 298: 1773–1775.
Guindon, S., and O. Gascuel, 2003 A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52: 696–704.
Jackson, A. P., J. A. Gamble, T. Yeomans, G. P. Moran, D. Saunders *et al.*, 2009 Comparative genomics of the fungal pathogens *Candida dubliniensis* and *Candida albicans*. Genome Res. 19: 2231–2244.
Jeffries, T. W., I. V. Grigoriev, J. Grimwood, J. M. Laplaza, A. Aerts *et al.*, 2007 Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast *Pichia stipitis*. Nat. Biotechnol. 25: 319–326.
Johnson, N., 2008 Hybrid incompatibility and speciation. Nat. Ed. Available at: http://www.nature.com/scitable/topicpage/hybrid-incompatibility-and-speciation-820.
Jones, T., N. A. Federspiel, H. Chibana, J. Dungan, S. Kalman *et al.*, 2004 The diploid genome sequence of *Candida albicans*. Proc. Natl. Acad. Sci. USA101: 7329–7334.
Jung, P. P., J. Schacherer, J. L. Souciet, S. Potier, P. Wincker *et al.*, 2009 The complete mitochondrial genome of the yeast *Pichia sorbitophila*. FEM. Yeast Res. 9: 903–910.
Kao, K. C., K. Schwartz, and G. Sherlock, 2010 A genome-wide analysis reveals no nuclear Dobzhansky-Muller pairs of determinants of speciation between *S. cerevisiae* and *S. paradoxus*, but suggests more complex incompatibilities. PLoS Genet. 6: e1001038.
Katoh, K., G. Asimenos, and H. Toh, 2009 Multiple alignment of DNA sequences with MAFFT. Methods Mol. Biol. 537: 39–64.
Kayingo, G., S. G. Kilian, and B. A. Prior, 2001 Conservation and release of osmolytes by yeasts during hypo-osmotic stress. Arch. Microbiol. 177: 29–35.
Ketchum, K. A., W. J. Joiner, A. J. Sellers, L. K. Kaczmarek, and S. A. Goldstein, 1995 A new family of outwardly rectifying potassium channel proteins with two pore domains in tandem. Nature 376: 690–695.
Kotseruba, V., D. Gernand, A. Meister, and A. Houben, 2003 Uniparental loss of ribosomal DNA in the allotetraploid grass *Zingeria trichopoda* (2n = 8). Genome 46: 156–163.
Kurtzman, C. P., and M. Suzuki, 2010 Phylogenetic analysis of ascomycete yeasts that form coenzyme Q-9 and the proposal of the new genera *Babjeviella*, *Meyerozyma*, *Millerozyma*, *Priceomyces*, and *Scheffersomyces*. Mycoscience 21: 2–14.
Lages, F., and C. Lucas, 1995 Characterization of a glycerol/H+ symport in the halotolerant yeast *Pichia sorbitophila*. Yeast 11: 111–119.
Lages, F., M. Silva-Graca, and C. Lucas, 1999 Active glycerol uptake is a mechanism underlying halotolerance in yeasts: a study of 42 species. Microbiology 145(Pt 9): 2577–2585.
Lee, H.-Y., J.-Y. Chou, L. Cheong, N.-H. Chang, S.-Y. Yang *et al.*, 2008 Incompatibility of nuclear and mitochondrial genomes causes hybrid sterility between two yeast species. Cell 135: 1065–1073.
Lee, S. C., M. Ni, W. Li, C. Shertz, and J. Heitman, 2010 The evolution of sex: a perspective from the fungal kingdom. Microbiol. Mol. Biol. Rev. 74: 298–340.
Lenglez, S., D. Hermand, and A. Decottignies, 2010 Genome-wide mapping of nuclear mitochondrial DNA sequences links DNA replication origins to chromosomal double-strand break formation in Schizosaccharomyces pombe. Genome Res. 20: 1250–1261.
Lewis, M., D. Pikaard, M. Nasrallah, J. Doelling, and C. Pikaard, 2007 Locus-specific ribosomal RNA gene silencing in nucleolar dominance. PLoS ONE 29: e815.
Liti, G., D. B. H. Barton, and E. J. Louis, 2006 Sequence diversity, reproductive isolation and species concepts in *Saccharomyces*. Genetics 174: 839–850.
Lynch, D. B., M. E. Logue, G. Butler, and K. H. Wolfe, 2010 Chromosomal G + C content evolution in yeasts: systematic interspecies differences, and GC-poor troughs at centromeres. Genome Biol. Evol. 2: 572–583.
Mallet, J., 2005 Hybridization as an invasion of the genome. Trends Ecol. Evol. 20: 229–237.
Maresova, L., and H. Sychrova, 2003 Physiological characterization of osmotolerant yeast *Pichia sorbitophila* and comparison with a putative synonym *Pichia farinosa*. Folia Microbiol. (Praha) 48: 211–217.
Marinoni, G., and M. A. Lachance, 2004 Speciation in the large-spored *Metschnikowia* clade and establishment of a new species, *Metschnikowia borealis* comb. nov. FEM. Yeast Res. 4: 587–596.
Martinez, P., and B. L. Persson, 1998 Identification, cloning and characterization of a derepressible Na+-coupled phosphate transporter in *Saccharomyces cerevisiae*. Mol. Gen. Genet. 1998: 628–638.
Murphy, H. A., H. A. Kuehne, C. A. Francis, and P. D. Sniegowski, 2006 Mate choice assays and mating propensity differences in natural yeast populations. Biol. Lett. 2: 553–556.
Nakao, Y., T. Kanamori, T. Itoh, Y. Kodama, S. Rainieri *et al.*, 2009 Genome sequence of the lager brewing yeast, an interspecies hybrid. DNA Res. 16: 115–129.
Neves, L., R. Oliveira, and C. Lucas, 2004 Yeast orthologues associated with glycerol transport and metabolism. FEM. Yeast Res. 5: 51–62.
Oliveira, R. P., F. Lages, and C. Lucas, 1996 Isolation and characterisation of mutants from the halotolerant yeast *Pichia sorbitophila* defective in H +/glycerol symport activity. FEMS Microbiol. Lett. 142: 147–153.
Pikaard, C., 2000 Nucleolar dominance: uniparental gene silencing on a multi-megabase scale in genetic hybrids. Plant Mol. Biol. 43: 163–177.
Prista, C., M. C. Loureiro-Dias, V. Montiel, R. Garcia, and J. Ramos, 2005 Mechanisms underlying the halotolerant way of Debaryomyces hansenii. FEM. Yeast Res. 5: 693–701.
Pujol, C., K. J. Daniels, S. R. Lockhart, T. Srikantha, J. B. Radke *et al.*, 2004 The closely related species *Candida albicans* and *Candida dubliniensis* can mate. Eukaryot. Cell 3: 1015–1027.
Querol, A., and U. Bond, 2009 The complex and dynamic genomes of industrial yeasts. FEMS Microbiol. Lett. 293: 1–10.

Rainieri, S., Y. Kodama, Y. Kaneko, K. Mikata, Y. Nakao *et al.*, 2006 Pure and mixed genetic lines of *Saccharomyces bayanus* and *Saccharomyces pastorianus* and their contribution to the lager brewing strain genome. Appl. Environ. Microbiol. 72: 3968–3974.

Rios, G., A. Ferrando, and R. Serrano, 1997 Mechanisms of salt tolerance conferred by overexpression of the HAL1 gene in *Saccharomyces cerevisiae*. Yeast 13: 515–528.

Rolland, T., and B. Dujon, 2011 Yeasty clocks: dating genomic changes in yeasts. C. R. Biol. 334: 620–628.

Sacerdot, C., S. Casaregola, I. Lafontaine, F. Tekaia, B. Dujon *et al.*, 2008 Promiscuous DNA in the nuclear genomes of hemiascomycetous yeasts. FEM. Yeast Res. 8: 846–857.

Scannell, D. R., K. P. Byrne, J. L. Gordon, S. Wong, and K. H. Wolfe, 2006 Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature 440: 341–345.

Sipiczki, M., 2008 Interspecies hybridization and recombination in *Saccharomyces* wine yeasts. FEM. Yeast Res. 8: 996–1007.

Souciet, J. L., B. Dujon, C. Gaillardin, M. Johnston, P. V. Baret *et al.*, 2009 Comparative genomics of protoploid *Saccharomycetaceae*. Genome Res. 19: 1696–1709.

Talavera, G., and J. Castresana, 2007 Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 56: 564–577.

Tamás, M. J., K. Luyten, F. C. Sutherland, A. Hernandez, J. Albertyn *et al.*, 1999 Fps1p controls the accumulation and release of the compatible solute glycerol in yeast osmoregulation. Mol. Microbiol. 31: 1087–1104.

Taussig, R., and M. Carlson, 1983 Nucleotide sequence of the yeast *SUC2* gene for invertase. Nucleic Acids Res. 11: 1943–1954.

Communicating editor: T. R. Hughes

La levure *Geotrichum candidum* : taxonomie, biodiversité et génome

RESUME

Geotrichum candidum est une levure hémiascomycète ubiquitaire longtemps considérée comme un champignon filamenteux. C'est l'une des levures les plus fréquemment trouvées dans les fromages dans les quelles elle contribue à l'affinage. Dans le cadre du projet ANR ALIA Food Microbiomes en partenariat avec des industriels fromagers et producteur de levain, nous avons caractérisé l'espèce *G. candidum* par une étude phylogénétique et placé de manière non ambigüe *G. candidum* parmi les levures hémiascomètes. Une analyse MLST a permis de séparer les souches étudiées en deux groupes. Le premier contient essentiellement des souches environnementales tandis que le second ne contient que des souches isolé du fromage. Cela suggère une certaine sélection ou spécialisation d'un groupe de souche dans la fabrication du fromage. Une méthode de typage inter LTR plus discriminante a permis de typer l'ensemble des souches et peut fournir aux industriels un outil robuste pour le suivi d'une souche en production. Le génome de *G. candidum* CLIB 918 = ATCC 204307 a été séquencé. Les premières analyses ont mis en évidence des discontinuités évolutives parmi les gènes qui le composent. Parmi les 6802 gènes identifiés, 315 gènes présentent des orthologues chez les champignons filamenteux et non chez les levures. Cela suggère que durant l'évolution, *G. candidum* a conservé un grand nombre de gènes qui a été perdu chez les autres levures ou en a reçu certain par transfert horizontal de gènes. L'existence de ce même type de gènes chez d'autres levures ayant une position basale dans l'arbre des hémiascomycètes, suggère que *G. candidum* et ces levures ont une position intermédiaire lors de la transition évolutive champignon vers levure. Il est à noter que certains d'entre eux sont impliqués dans le métabolisme et pourraient jouer un rôle dans l'adaptation de cette levure à la fabrication du fromage.

MOTS CLES

Geotrichum candidum, taxonomie, biodiversité, évolution, génome, MLST, HGT

The yeast *Geotrichum candidum*: taxonomy, biodiversity and genome

ABSTRACT

Geotrichum candidum is a hemiascomycetous yeast frequently found in the environment and foodstuffs. It is one of the main yeasts in cheese and it is widely used as adjunct culture in the maturation of cheese. Within ANR project ALIA Food Microbiomes in partnership with industry, we characterized the species the species *G. candidum* by a multigene phylogenetic study. MLST analysis allowed us to separate the studied strains into two groups. The first contains mainly environmental strains while the second contains only strains isolated from cheese. This suggests a specialization or a selection of a group of strains within industry. We developed a typing method by inter LTR profiles, which can provide a robust tool for an industrial monitoring of strains. The genome of *G. candidum* CLIB 918 = ATCC 204307 was sequenced. Preliminary analyses revealed evolutionary discontinuities among genes. 6802 genes where identified in which 315 genes have orthologs in filamentous fungi and not in yeast. This suggests that during evolution, *G. candidum has* retained a large number of genes which have been lost in other yeasts or has received some by horizontal gene transfer. The existence of this other yeasts also having a basal position in hemiascomycetous tree suggests that *G. candidum* and these other yeasts have an intermediate position during the evolutionary transition fungus to yeast. It is noteworthy that some of them are involved in the metabolism and may play a role in the adaptation of the yeast to the cheese environment.

KEYWORDS

Geotrichum candidum, taxonomy, biodiversity, evolution, genome, MLST, HGT